



# **Multimodal Systems For Scientific and Educational Applications**

A THESIS

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS  
FOR THE DEGREE OF

**DOCTOR OF PHILOSOPHY**

BY

**AVINASH ANAND**

(PhD21003)

Department of Computer Science and Engineering (Joint with the  
Department of Human-centered Design)

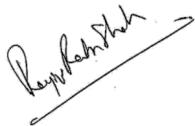
INDRAPRASTHA INSTITUTE OF INFORMATION TECHNOLOGY  
NEW DELHI- 110020

**November, 2024**

## Certificate

The work contained in this thesis entitled, **Multimodal Systems For Scientific and Educational Applications**, has also been submitted to Indraprastha Institute of Information Technology (IIIT), Delhi PhD program. To the best of my knowledge and belief, the thesis contains no material previously published or written by another person except where due reference is made.

**Advisors' Name**

A handwritten signature in black ink, reading "Rajiv Ratn Shah", is written over a horizontal line that has a small arrowhead pointing to the right at its end.

Dr. Rajiv Ratn Shah, Associate and Chair Professor, Depts.  
of HCD & CSE, IIIT Delhi  
Place: New Delhi  
Date: November 2024

## Acknowledgment


I am profoundly thankful to my advisor, Dr. Rajiv Ratn Shah, whose insightful feedback on my analysis and framing greatly improved the quality of this thesis. His encouragement and guidance have been a continual source of motivation throughout my PhD journey.

I am grateful to the Indraprastha Institute of Information Technology and GS1 India for their financial support throughout the development of this thesis. I would also like to extend my appreciation to the dedicated support staff at IIIT-D, especially Mr. Bhawani, Ms. Binu, Ms. Anshu Dureja, and Mr. Raju Biswas, for their exceptional assistance in promptly addressing all technical and administrative matters. A special thanks goes to Mr. Rajendra from the IIIT-D Library, whose invaluable help in accessing essential materials greatly contributed to the success of my thesis.

I am deeply grateful to Professor Shin'ichi Satoh for the invaluable internship opportunity at NII Tokyo, which has greatly enriched my PhD journey.

I would like to extend my heartfelt thanks to my colleagues—Raj Jaisawal, Naman, Kritarth, Mohit, Janak, Pijush, Arnav, Medha, Yaman, Avni, Ritika, Hemant Yadav, and Ritwik Mishra, Mann for their support throughout my PhD journey. I am especially grateful to Rishi, Chayan, Sarthak and Vinayak for their invaluable assistance in the completion of my thesis.

Above all, I wish to express my heartfelt gratitude and love to my wife and colleague, Dr. Astha Verma, whose unwavering support made this work possible. I am also deeply thankful to my parents, Brothers, Rushali Ma'am, Rama, and extended family for their constant encouragement and unconditional support.

  
(Avinash Anand)

## Abstract

Large Language Models (LLMs) have transformative capabilities but limited application in specialized educational and research contexts requiring multimodal reasoning, context-aware processing, and domain-specific understanding. Education and research need tools handling nuanced textual and visual interplay and context-rich language. This thesis advances LLMs in high school physics reasoning, multimodal problem-solving, mathematical reasoning with bilingual understanding, student engagement analysis, grammar correction, and citation generation. The first contribution enhances multimodal reasoning in physics education, where problems combine text and diagrams. Introducing the MM-PhyQA dataset and using retrieval-augmented methods with Multi-Image Chain-of-Thought (MI-CoT), the study achieves 71.60% accuracy on complex physics tasks, improving LLM support for physics education. Next, mathematical problem-solving, especially geometry, is addressed. The GeoVQA and GPSM4K datasets enable training of LLaVA-v1.5 and G-LLaVA models, which outperform Larger LLMs in geometric reasoning benchmarks, showing the benefit of tailored LLMs for visually and linguistically challenging math tasks. The thesis also tackles student engagement prediction in online learning, lacking in-person cues. Using ECLIPSE dataset to capture virtual attention dynamics, fine-tuning CG-ViT and NeuralGaze models yields a 21.45% improvement in engagement accuracy, supporting adaptive, personalized remote education. For grammatical error correction (GEC), traditional neural machine translation methods struggle with long context. The Dynamic Context Learner (DCL) enables LLMs to integrate relevant context dynamically, improving accuracy on CoNLL-2014 and BEA-Dev datasets with F1 score gains, enhancing grammar correction for academic writing. In academic writing, accurate citation generation is vital. Existing models lack depth to capture complex citation relationships. The multi-source citation text generation (M-CTG) framework combines knowledge graphs and keyphrase embeddings with fine-tuned Vicuna and Alpaca models, achieving a 36.98% ROUGE-1 improvement, facilitating better citation and source attribution. Collectively, this thesis demonstrates the potential of multimodal LLMs fine-tuned for domain-specific edu-

cational and scientific tasks. By introducing new datasets, refining architectures, and applying innovative methods, it bridges AI application gaps across fields. In physics education, bilingual mathematical reasoning, and engagement analysis, tailored multimodal LLMs enhance reasoning and context processing. These advances show how domain-specific multimodal AI tools benefit both education and science, paving the way for precise, context-aware, impactful LLM applications across complex, cross-domain challenges.

# TABLE OF CONTENTS

<b>Abstract</b>	<b>iii</b>
<b>List of Tables</b>	<b>xii</b>
<b>List of Figures</b>	<b>xv</b>
<b>List of Papers</b>	<b>xviii</b>
<b>List of Abbreviations</b>	<b>xx</b>
<b>1 Introduction</b>	<b>1</b>
1.1 Motivation . . . . .	2
1.2 Research Aims . . . . .	3
1.2.1 Contributions of this Research . . . . .	5
1.2.2 Dissertation Organization . . . . .	6
<b>2 Related Works</b>	<b>9</b>
2.1 Introduction . . . . .	9
2.2 Educational Domain . . . . .	10
2.2.1 Physics Question Answering . . . . .	10
2.2.2 Online Learning . . . . .	12
2.2.3 Geometric Question Answering . . . . .	14
2.3 Scientific Domain . . . . .	15
2.3.1 Grammatical Error Correction . . . . .	16
2.3.2 Citation Generation . . . . .	17
2.3.3 Scientific Controllable Text Generation . . . . .	19

<b>3</b>	<b>Physics Question Answering</b>	<b>22</b>
3.1	Motivation	22
3.2	Introduction	22
3.3	Experiments	24
3.3.1	Novel Physics Education Dataset	24
3.3.2	SciPhyRAG: Retrieval Augmentation to Improve LLMs on Physics	
	Q&A	32
3.3.3	MM-PhyQA: Multimodal Physics Question-Answering With	
	Multi-Image CoT Prompting	41
3.3.4	Model Performance	46
3.4	Conclusion	49
<b>4</b>	<b>Online Learning</b>	<b>52</b>
4.1	Motivation	52
4.2	Introduction	52
4.3	Experiments	54
4.3.1	Unveiling Learner Dynamics: The ECLIPSE Dataset and Neu-	
	ral Gaze Framework for Prolonged Engagement Assessment in	
	Online Learning	54
4.4	Conclusion & Future Scope	69
<b>5</b>	<b>Mathematical Question Answering</b>	<b>71</b>
5.1	Motivation	71
5.2	Introduction	71
5.3	Experiments	74
5.3.1	GeoVQA: A Comprehensive Multimodal Geometry Dataset for	
	Secondary Education	74
5.3.2	Improving Multimodal LLM's Ability In Geometric Problem	
	Solving, Reasoning, And Multi-Step Scoring	83

5.3.3	Mathify: Evaluating LLMs on Mathematical Problem Solving	
	Tasks	99
5.3.4	Bridging the Language Gap in Mathematical Reasoning of LLMs	106
5.4	Conclusion and Future Scope	115
<b>6</b>	<b>Grammar Correction</b>	<b>118</b>
6.1	Motivation	118
6.2	Introduction	119
6.3	Experiments	120
6.3.1	Datasets	120
6.3.2	Methodology	124
6.3.3	GECToR-RoBERTa & Model Training	126
6.3.4	Training	127
6.3.5	Results	128
6.3.6	Error Analysis	129
6.4	Conclusion	129
<b>7</b>	<b>Citation and Text Generation</b>	<b>131</b>
7.1	Motivation	131
7.2	Introduction	131
7.3	Experiment	133
7.3.1	KG-CTG: Citation Generation with Knowledge Graph-Enhanced	
	LLMs	133
7.3.2	Context-Enhanced Language Models for Generating Multi-paper	
	Citations	140
7.3.3	Advances in CTG: Leveraging Multi-Source Seq2Seq Models	
	and LLMs	147
7.4	Conclusion	158
<b>8</b>	<b>Summary and Future Directions</b>	<b>160</b>

<b>8.1 Introduction</b> . . . . .	160
<b>8.2 Summary</b> . . . . .	160
<b>8.3 Future Work</b> . . . . .	162
<b>Bibliography</b>	<b>182</b>

## LIST OF TABLES

3.1	Distribution of Questions by Chapter . . . . .	25
3.2	Comparison of Baseline and CoT Results . . . . .	30
3.3	Table showing the BERT, ROUGE and METEOR scores of the 2 Fine- Tuned Vicuna-LoRA models . . . . .	40
3.4	Final Answer Accuracy before and after Fine-Tuning with PhyQA . . . . .	41
3.5	Topics and subtopics in the MM-PhyQA dataset . . . . .	44
3.6	Performance of text-only and multimodal (MM) models. Model training spec- ifications such as LoRA Rank and whether MI-CoT Prompting was used have been mentioned. All models were fine-tuned except for GPT-4, for which the answers were extracted using zero-shot prompting . . . . .	46
3.7	Performance of text-only LLMs using zero-shot prompting and fine- tuning . . . . .	47
4.1	Weighted Cohen’s kappa for the three annotators . . . . .	56
4.2	ECLIPSE Dataset: Affective State Label Composition . . . . .	57
4.3	Engagement level Classification Results for DAiSEE, EngageNet, and our dataset ECLIPSE . . . . .	65
4.4	Confusion, Frustration, and Boredom results for DAiSEE and ECLIPSE dataset for various models . . . . .	66
4.5	Engagement level prediction comparison with single-frame and video- based benchmarks. Our model gives competitive performance to video- based models and surpasses existing single-frame benchmarks. We set new benchmarks with GLAMOR-Net+FA+FL and NeuralGaze . . . . .	67
4.6	Engagement level Classification Results for EngageNet and EngageNet & ExCEDA combined respectively . . . . .	68

4.7 Results for binary classification of Content-guided ViT model on the ECLIPSE dataset vs ViT model without video features . . . . .	69
5.1 GeoVQA Dataset statistics . . . . .	75
5.2 Comparison of the Accuracy (%) on Different Models with Zero-Shot Inference on GeoVQA (Ours) Dataset and Benchmark Datasets: PGPS9K and Geometry3K . . . . .	78
5.3 Comparison of Different Models' Accuracy (%) with Few-Shot Inference on GeoVQA (Ours) Dataset . . . . .	80
5.4 Comparison on different model Accuracy (%) with Few-shot inference on GeoVQA (ours) dataset and Benchmark Dataset: PGPS9K and Geometry3K. The highest accuracy among each dataset is highlighted with bold and green, red, and yellow colour highlights across each model and dataset to represent the highest score on open source models. . . . .	80
5.5 GPSM4K Finetuning Dataset Statistics . . . . .	84
5.6 GPSM4K Inference Dataset Statistics . . . . .	84
5.7 Comparison of the Accuracy (%) on different models with Zero-shot inference on GPSM4K (Ours) dataset and Benchmark Dataset: PGPS9K and Geometry3K . . . . .	90
5.8 Performance of LVLMS on GPSM4K test set . . . . .	91
5.9 Image Captioning Result . . . . .	93
5.10 Model performance on GPSM4K test after adding Image captions . . . . .	94
5.11 Model performance on GPSM4K test with RAG . . . . .	95
5.12 Distribution of Question Types in Augmented Math-401 Dataset . . . . .	100
5.13 Exact Match Accuracy on 100 Samples from 5 Datasets and MathQuest (Before Fine-tuning on Math-401). (*) Denotes the Math-401 Subset Used for Augmentation. . . . .	103
5.14 Exact Match Accuracy on 100 Samples from 5 Datasets and MathQuest (After Fine-tuning on Math-401). (*) Denotes the Math-401 Subset Used for Augmentation. . . . .	104

5.15 Performance of LLMs in Hindi Math questions using decomposition strategy. Bold values indicate improvements from the previous step. Underlined values show the highest performance of SLLMs for each operation.	108
5.16 Performance Comparison of Open-Source and Closed-Source Models on English and Hindi Mathematical Benchmarks	113
5.17 Performance of Bilingual Models on IndiMathQA Using SFT_easy Training	116
5.18 Performance of Bilingual Models on IndiMathQA Using SFT_easy+medium Training	116
6.1 Statistics of the datasets that were used in the research	122
6.2 Types of Error and Their Distribution in the Error Script	124
6.3 Our DCL-GEC model is evaluated against two (SOTA) models on sentence variations in the FCE-Test, BEA-Dev, and CoNLL-2014 datasets. Unlike the comparison models, which rely on a fixed number of contextual sentences, our model dynamically adjusts context, leading to superior performance across these benchmarks.	126
6.4 GECToR is compared with the DCL-GEC model. We compare results for sentences (sent) and paragraphs (para) variation of FCE-Test, BEA-Dev, and CoNLL-2014, where we achieved better results on the paragraph variant and were at par for the sentence variant.	126
6.5 Comparison of Random/fixed split vs DCL on CoNLL-2014 having paragraphs.	127
6.6 Gector vs DCL-GEC on XSUM Dataset	127
6.7 GECToR vs DCL-GEC w/ SciBERT as Transformer on S2ORC Dataset	128
6.8 GECToR vs DCL-GEC on CNN Dataset	128
6.9 This table presents a comparison of error type-specific performance between GECToR and DCL-GEC on the FCE-Test dataset, evaluated using the $F_{0.5}$ score. The last column highlights the performance difference between our proposed model and GECToR, demonstrating a significant improvement of 24% with the DCL-GEC model.	128

7.1 Statistics of the dataset extracted from the S2ORC corpus. . . . .	136
7.2 Results of the CTG Task Using LLMs (Without Knowledge Graphs)	139
7.3 Results of CTG Task using LLMs (With Knowledge Graphs) . . . .	139
7.4 Statistics of the dataset created from the S2ORC corpus. . . . .	142
7.5 Results of Fine-Tuned LLM . . . . .	147
7.6 Results of Fine-Tuned Model + Knowledge Graph as Prompt . . . .	147
7.7 Metadata used with abstracts of the papers . . . . .	149
7.8 Comparison of model performance with and without Summary in Knowl- edge Graphs (KG) . . . . .	156
7.9 Comparison of model performance with various methods . . . . .	156
7.10 Model Comparison . . . . .	157

## LIST OF FIGURES

1.1 AI Applications in Education and Scientific Writing.	2
1.2 Overview of AI applications in the education domain.	3
1.3 Overview of AI applications in the scientific domain.	4
3.1 Question length distribution	26
3.3 Insight into the number of questions at chapter-wise granularity in PhyQA	34
3.4 <b>Left:</b> A snippet from the Physics textbook of a section about "Motion of Centre of Mass"; <b>Right:</b> A snippet of the annotation file corresponding to the textbook snippet	36
3.5 MMPhy-QA Dataset questions	43
3.6 Schematic Pipeline of Multimodal Question Answering	45
3.7 Multi-Image Chain of thought (MI-CoT) Prompted text provided as in- put to LMMs during training. The main question to be answered is preceded by two exemplars, with the three questions separated by a de- limiter. The image is a sequence of three comma-separated file names and the label is the ground truth	45
3.8 Comparison of the accuracy and rouge scores of different LLaVA vari- ants when trained using (MI-CoT) Prompting vs their non-CoT prompted supervised fine-tuned (SFT) counterparts	48
3.9 Types of errors encountered by LLaVA-1.5 13b	50
4.1 Sample of images in ECLIPSE for different affects at different levels	55
4.2 Engagement affect’s class distribution for comparison between Enga- geNet, DAiSEE & DEEPER	57
4.3 Model architecture with NCA-based feature embedding module com- bined with Multi-head self-attention and facial landmarks	58

4.4	NCA-based feature extraction module consists of 2 NCA models. For NCA1, the original image is concatenated with 16 additional hidden channels. The output of NCA1 is downsampled using MaxPooling and given as input to NCA2 after the addition of 32 additional channels .	59
5.1	Data Extraction and topic-wise distribution . . . . .	75
5.3	Fine-tuned models performance on solution steps . . . . .	81
5.4	Base models performance on solution steps . . . . .	81
5.5	Top@1 accuracy of base and fine-tuned models . . . . .	82
5.6	Data Extraction and topic-wise distribution . . . . .	84
5.7	Prompt used for Solution Regeneration . . . . .	86
5.8	Image captioning example . . . . .	92
5.9	Multimodal RAG tackles this challenge by combining the strengths of images and descriptions. It analyzes diagrams to understand the problem and uses text descriptions to find the solution. In 1-shot multimodal RAG, a single problem-solution example is used to find similar problems from a vast dataset. These retrieved examples then help solve new inference questions, showcasing the model’s ability to generalize its knowledge . . . . .	95
5.10	Two Prompts Used in Final Solution Extraction and Comparison Respectively . . . . .	96
5.11	Prompt Used for Step-by-Step Evaluation . . . . .	97
5.12	The binary ‘Correct’ or ‘Incorrect’ evaluative approach of existing benchmarks is inadequate to examine the depth and precision of the multi-step reasoning process. To this end, We propose a CoT evaluation strategy to thoroughly assess their mathematical CoT skills in visual contexts, involving Multimodal LLMs. . . . .	98
5.13	Our Dataset <b>MathQuest</b> Sample . . . . .	101
5.14	Distribution of Count of Samples of each Concept . . . . .	102
5.15	This figure shows the fine-tuning flow, the LLMs we use for fine-tuning, and the datasets we use for inference. . . . .	104

5.16 Strategic Decomposition: A Method for Simplifying Complex Math Problems into Smaller, Solvable Units to Improve SLLM Capabilities	107
5.17 Curriculum Learning with Structured Solutions: A Comprehensive Framework to Gradually Guide Models Through Complex Mathematical Challenges.	110
6.1 The DCL-GEC Architecture is composed of two main components: (1) a <i>Sentence Tokenizer</i> , which divides the input paragraph into individual sentences; and (2) the <i>DCL Model</i> , which reassembles these sentences into contextually coherent paragraphs. Once restructured, the <i>GECToR</i> model applies sequence tagging to identify and implement corrections, producing an edited and grammatically accurate paragraph.	123
7.1 Workflow diagram illustrating the experimental setup, including two workflows: <b>1.</b> Prompt creation without knowledge graph relations, and <b>2.</b> Relation extraction from abstracts followed by prompt creation.	134
7.2 Structures of Prompts used for the Large Language Models.	135
7.3 Example of Generated Citation text from <b>Vicuna</b> without knowledge graph relations	137
7.4 Example of Generated Citation text from <b>Alpaca</b> with knowledge graph relations	138
7.5 Multi-Sentence CTG	140
7.6 LLaMA, Vicuna & Alpaca	142
7.7 Prompt structures for LLMs. (a) without KG relationships; (b) with KG relationships.	143
7.8 Example of Generated Citation text from the best Model ( <b>Vicuna</b> ) without knowledge graph relations	145
7.9 Example of Generated Citation text from the best Model ( <b>Vicuna</b> ) with knowledge graph relations	146
7.10 Our proposed <b>M-CTG</b> . BART encoder takes source and cited (target) abstracts for text encoding. Graph and style embedding passed in MLP. BART Encoder output combined with MLP output in the Fusion Layer.	147



## Conferences



Figure 1: ECAI-24  
*Accepted*



Figure 2: AAAI-24  
*Accepted*



Figure 3: 2 papers  
ACM-24,25  
*Accepted*



Figure 4: 5 Papers at  
BDA-24  
*Accepted*



Figure 5: CIKM-24  
*Accepted*



Figure 6: 3 Papers at  
MIPR-24  
*Accepted*



Figure 7: PAKDD-24  
*Accepted*



Figure 8: GAIED-24  
*Accepted*

## List of Papers

### Recently Accepted/Submitted Papers, 2024

1. Avinash Anand, Avni Mittal, Rajiv Ratn Shah, Roger Zimmermann, and Shin'ichi Satoh, "Unveiling Learner Dynamics: The ECLIPSE Dataset and NeuralGaze Framework for Prolonged Engagement Assessment in Online Learning", *Accepted at ECAI-24*. Core-A [4](#)
2. Avinash Anand, Naman Lal, Rajiv Ratn Shah, "Advances in Citation Text Generation: Leveraging Multi-Source Seq2Seq Models and Large Language Models", *Accepted at CIKM-24*. Core-A [7](#)
3. Avinash Anand, Raj Jaiswal, Rajiv Ratn Shah, "Advancing Multimodal LLMs: A Focus on Geometry Problem Solving, Reasoning, and Sequential Scoring", *ACM Multimedia Asia 2024* [5](#)
4. Avinash Anand, Raj Jaiswal, Rajiv Ratn Shah, Roger Zimmermann, "Enhancing Geometric Problem-Solving using GeoVQA: A Multimodal Geometry Dataset for Advanced Vision Language Model", *Accepted at MIPR-24* [5](#)
5. Avinash Anand, Ashwin R Nair, Kritarth Prasad, Rajiv Ratn Shah, "A Bilingual Math Problem Solver: Solving NCERT-Level Science Problems in English and Hindi", *AAAI (A\*)* [5](#)
6. Avinash Anand, Janak Kapuriya, Rajiv Ratn Shah, "Enhancing Low-Resource Physics Q&A in Large Multimodal Models with Caption based Cues", .
7. Avinash Anand, Janak Kapuriya, Rajiv Ratn Shah, "Optimizing Vision Language Models for Scientific VQA through Caption-Aware Supervised Training", .

# Research Publications

## Conference Proceedings

1. A. Anand, Janak, Apoorv, A. Verma, and R. R. Shah, “Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting,” in *PAKDD-24, Taiwan: PAKDD, 2024* [3](#)  
doi:[https://doi.org/10.1007/978-981-97-2262-4\\_5](https://doi.org/10.1007/978-981-97-2262-4_5) Core-A.
2. A. Anand, K. Addala, K. Baghel, et al., “Revolutionizing high school physics education: A novel dataset,” in *2023 Conference on Big Data and Artificial Intelligence, New Delhi, India: IEEE Computer Society, 2023* [3](#)  
doi: [10.1007/978-3-031-49601-1\\_5](https://doi.org/10.1007/978-3-031-49601-1_5)
3. A. Anand, A. Goel, M. Hira, et al., “Sciphyrag - retrieval augmentation to improve LLMs on physics Q&A,” in *2023 Conference on Big Data and Artificial Intelligence, New Delhi, India: IEEE Computer Society, 2023* [3](#)  
doi: [10.1007/978-3-031-49601-1\\_4](https://doi.org/10.1007/978-3-031-49601-1_4)
4. A. Anand, M. Gupta, K. Prasad, et al., “KG-CTG: Citation generation through knowledge graph-guided large language models,” in *2023 Conference on Big Data and Artificial Intelligence, New Delhi, India: IEEE Computer Society, 2023* [7](#)  
doi: [10.1007/978-3-031-49601-1\\_3](https://doi.org/10.1007/978-3-031-49601-1_3)
5. A. Anand, M. Gupta, K. Prasad, et al., “Mathify: Evaluating large language models on mathematical problem-solving tasks,” in *GAIED, 2023, New Orleans: Neural Information Processing Systems, 2023* [5](#)
6. A. Anand, A. Jairath, N. Lal, et al., “GEC-DCL: Grammatical error correction model with dynamic context learning for paragraphs & scholarly papers,” in *2023 Conference on Big Data and Artificial Intelligence, New Delhi, India: IEEE Computer Society, 2023* [6](#)  
doi: [10.1007/978-3-031-49601-1\\_7](https://doi.org/10.1007/978-3-031-49601-1_7)
7. A. Anand, K. Prasad, U. Goel, et al., “Context-enhanced language models for generating multi-paper citations,” in *2023 Conference on Big Data and Artificial Intelligence, New Delhi, India: IEEE Computer Society, 2023* [7](#)  
doi: [10.1007/978-3-031-49601-1\\_6](https://doi.org/10.1007/978-3-031-49601-1_6)
8. A. Anand, A. Goel, M. Hira, S. Bangar, and D. R. R. Shah, “Advancements in scientific controllable text generation methods,” 2023. *arXiv: 2307.05538 [cs.CL]* [7](#)  
*ACM Survey Submitted.*

## List of Abbreviations

<b>Abbreviation</b>	<b>Description</b>
AI	Artificial Intelligence
ML	Machine Learning
LLM	Large Language Model
NLP	Natural Language Processing
CoT	Chain of Thought
ViT	Vision Transformer
GPT	Generative Pretrained Transformer
RAG	Retrieval Augmented Generation
MLP	Multi-Layer Perceptron
DCL	Dynamic Context Learner
CTG	Citation Text Generation
GEC	Grammatical Error Correction
RLHF	Reinforcement learning from human feedback
SOTA	State of the Art
NCA	Neural cellular Automata
VAE	Variational auto-encoder
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Unit
LoRA	Low-Rank Adapters
BERT	Bidirectional Encoder Representations from Transformers
KL	Kullback-Leibler
API	Application programming interface
NCERT	National Council of Educational Research and Training

# CHAPTER 1

## Introduction

Large Language Models (LLMs) mark a clear step forward in Artificial Intelligence, yet their full value in education and research remains partly untapped. Real-world tasks in these settings rarely arrive as plain text; instead they combine prose, diagrams, equations, and long, context-rich conversations, and they call for step-by-step reasoning that mirrors human thought. Current LLMs struggle to keep track of meaning across such mixed inputs and extended interactions, limiting their usefulness where precision and clarity matter most. This thesis tackles those limitations by pairing tailored datasets with lightweight architectural changes and task-specific prompting, aiming to narrow the gap between general LLM ability and the exacting needs of domain experts.

To ground the investigation, we target five high-impact applications that together form a comprehensive test bed for multimodal, context-aware reasoning, as depicted in Figure [1.1](#). First, we develop an intelligent tutoring pipeline for **high-school physics**, which enables a model to read free-body diagrams, translate equations into everyday language, and guide students through conceptually dense problems. Second, in **geometry problem-solving**, we train bilingual models that can weave textual hints and visual cues into theorem-rich, stepwise proofs, making automated assistance accessible in both English and Hindi classrooms. Third, for **virtual learning environments**, we introduce an engagement-detection system that fuses webcam signals and click-stream patterns to gauge attention in real time and adjust the pacing of lesson content. Fourth, our **document-level grammatical-error-correction** module delivers context-sensitive edits that preserve technical terminology while improving overall readability and coherence. Finally, a **contextual citation-generation** framework retrieves, ranks, and inserts references that align with the surrounding discourse, raising the quality and relevance of scientific writing.

By treating these diverse challenges as a unified benchmark, we demonstrate how carefully chosen data, prompts, and model tweaks can transform general-purpose LLMs into dependable, specialized assistants. The chapters that follow detail the methods, report empirical gains, and discuss the broader implications for AI-enhanced education

and research.

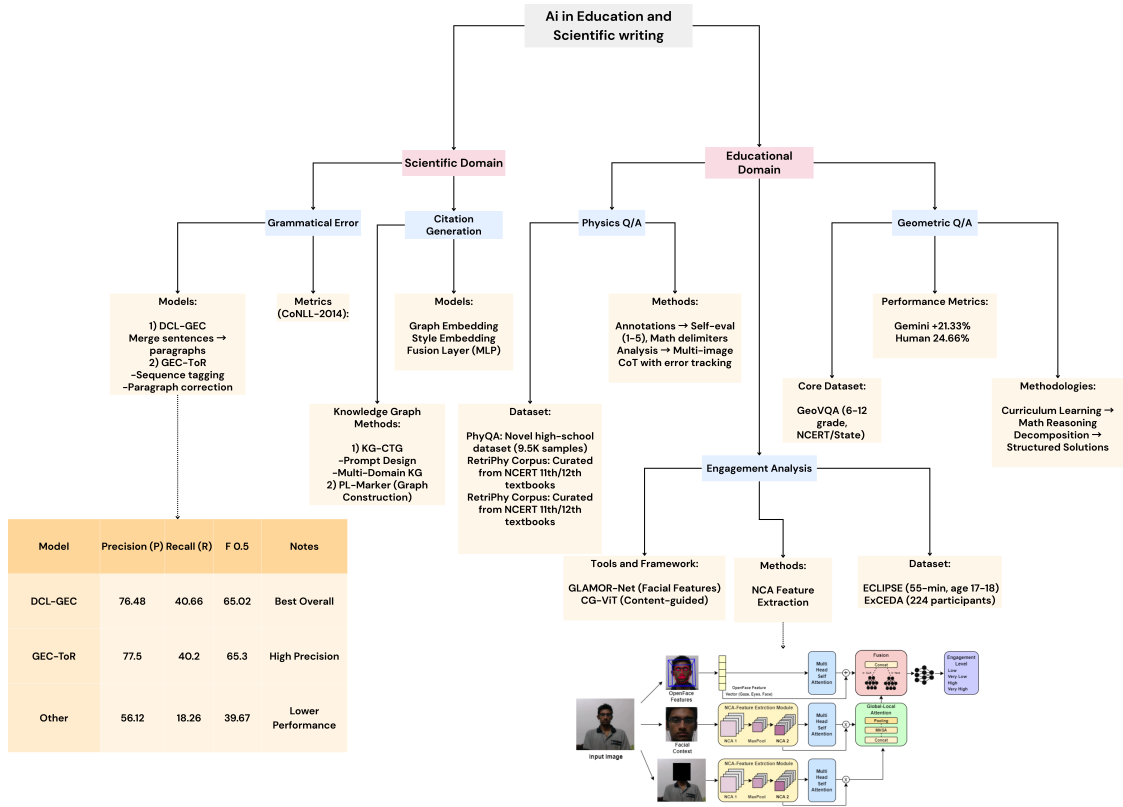


Figure 1.1: AI Applications in Education and Scientific Writing.

## 1.1 Motivation

The motivation for this work stems from the need to extend LLM capabilities beyond generic tasks, focusing instead on high school physics education, geometry problem-solving, student engagement analysis in online education, Grammatical Error Correction (GEC), Citation Text Generation (CTG), and controlled text synthesis. Each of these domains presents unique challenges that require customized approaches. For instance, physics education demands the integration of both textual and visual information to solve knowledge-intensive problems, while geometry and mathematical problem-solving require a model capable of multimodal numerical reasoning. Engagement analysis in online learning depends on real-time emotion and attention recognition, which necessitates the development of robust, adaptable models. Similarly, GEC requires a contextually aware approach to correct language at the paragraph or document level, and CTG involves generating accurate, context-relevant citations essential for scholarly writing. These diverse AI applications in the education domain are thoroughly outlined

in Figure 1.2.

The contributions made in this thesis are rooted in the development of state-of-the-art (SOTA) techniques and resources, which collectively enhance AI’s capacity to support domain-specific tasks, demonstrating the potential for LLMs and multimodal models in real-world educational and research applications.

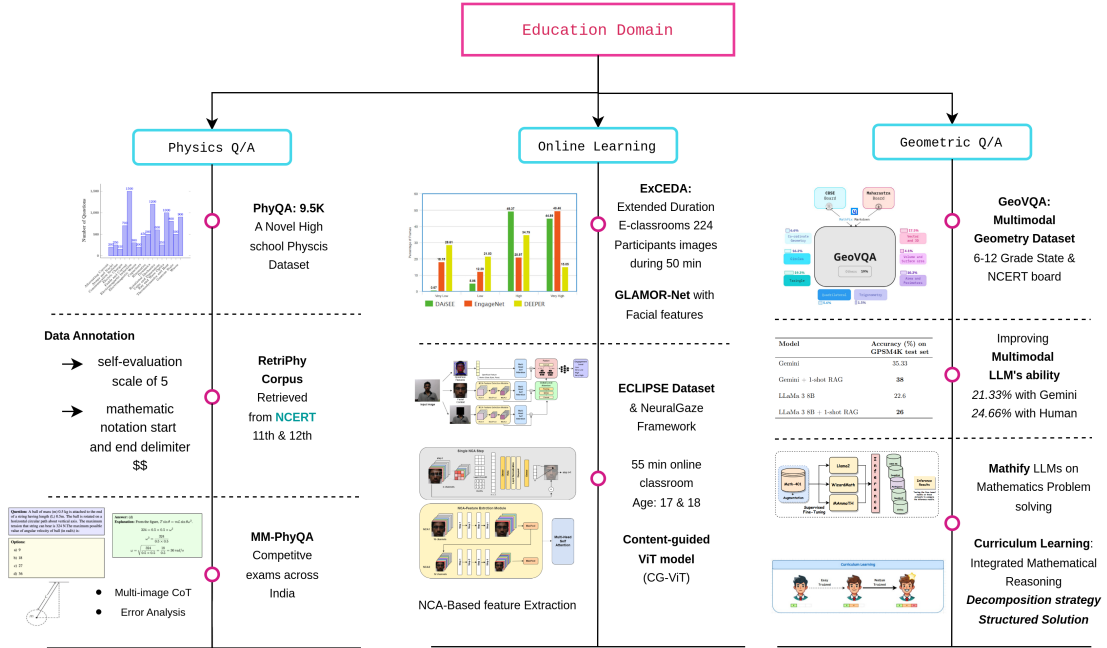


Figure 1.2: Overview of AI applications in the education domain.

## 1.2 Research Aims

The primary aim of this thesis is to advance the application of LLMs and multimodal models in domain-specific tasks by addressing gaps in their reasoning and contextual capabilities. This research is centered around several key areas:

**1. Physics Education and Problem-Solving:** Physics education presents complex challenges as students must navigate a blend of textual descriptions, equations, and diagrams to fully grasp concepts. Traditional LLMs often struggle with cross-modal reasoning required for physics questions, particularly in translating conceptual understanding into accurate solutions. This thesis aims to address the need for LLMs that can integrate and process information across these diverse formats to provide reliable support in physics education.

**2. Engagement Analysis in Online Learning:** As online learning environments

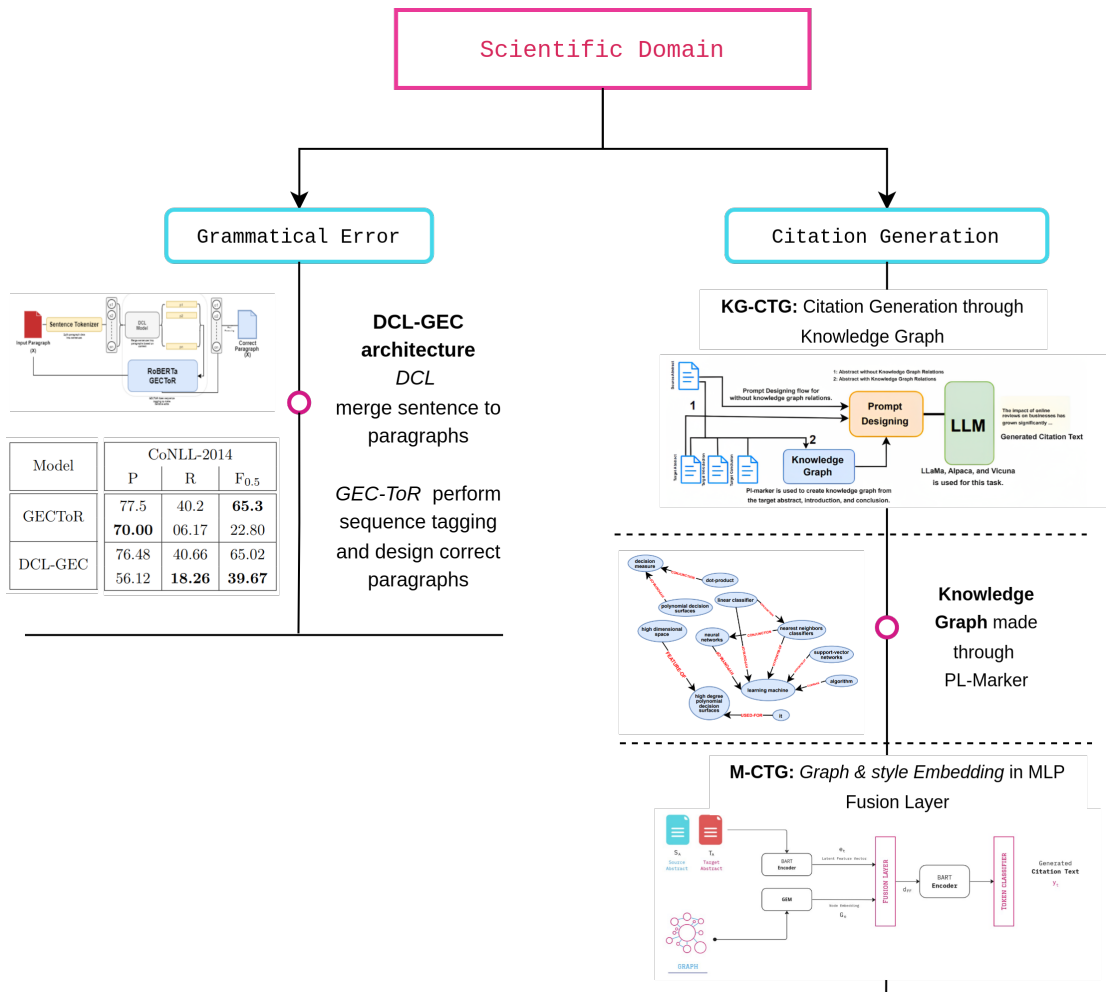


Figure 1.3: Overview of AI applications in the scientific domain.

proliferate, accurately monitoring and understanding student engagement is critical for creating effective, adaptive learning experiences. Engagement detection goes beyond simple face-expression classification—while emotion models map isolated expressions to labels, engagement models must interpret subtle, sustained cues (e.g., micro-expressions, gaze drift, posture shifts, hand movements) over time and across individuals. Virtual classrooms lack the natural, real-time feedback loops of in-person settings, so AI systems must fuse facial, gaze, pose, and contextual signals and operate under strict latency and resource constraints. In particular, detecting low-engagement states in live video streams demands lightweight architectures capable of sub-100 ms inference per frame, ensuring that personalized interventions can be triggered seamlessly and without disrupting the learning flow.

**3. Geometry and Multimodal Numerical Reasoning:** Geometry and other mathematical domains often require simultaneous processing of textual information and visual elements such as shapes and diagrams. Existing AI models and datasets are limited

in supporting reasoning that combines both visual and textual data, particularly for secondary education. Additionally, there is a significant need for bilingual resources in mathematical education, as Hindi-speaking students face barriers to accessing effective AI tools for complex subjects. This thesis aims to address these challenges by enhancing AI's capacity to tackle multimodal problems across languages, providing comprehensive resources that support geometry and mathematical reasoning in bilingual educational settings.

4. **Grammatical Error Correction (GEC):** In educational and scientific writing, effective grammar correction is crucial for clarity and coherence. Traditional GEC models typically operate at the sentence level, which can lead to a loss of context in more extended passages or complex documents. There is a need for approaches that dynamically incorporate context over multiple sentences to provide more accurate and comprehensive grammar correction, especially for academic and technical writing.

5. **Citation Text Generation (CTG):** Accurate citation generation is a critical component of scientific writing, ensuring credibility and proper attribution. Existing citation models often lack the depth to capture nuanced relationships between cited works and the citing context, leading to inaccuracies in citation relevance and coherence. There is a demand for citation systems that can produce contextually rich, precise citations that support researchers and students in maintaining academic integrity in their writing. Figure 1.3 provides an overview of these AI applications in the scientific domain, highlighting our work in Grammatical Error Correction (GEC) and Citation Text Generation (CTG).

### 1.2.1 Contributions of this Research

This thesis makes several significant contributions across a range of AI applications, each enhancing the performance and adaptability of LLMs in specific domains:

- In physics education, this work introduces the MM-PhyQA dataset and advances LLM accuracy in physics question answering through retrieval-augmented techniques and multimodal approaches. Techniques like Multi-Image Chain-of-Thought (MI-CoT) prompting are applied to improve reasoning accuracy, with tailored datasets enabling LLMs to handle complex, knowledge-intensive tasks, thus expanding their role in supporting STEM education.
- In the domain of online learning, this thesis develops novel datasets and models that assess student engagement in real time, focusing on detecting low-engagement

states in virtual learning settings. By integrating content-guided approaches and participant expressions, this research enhances models like GLAMOR-Net and Content-Guided Vision Transformer (CG-ViT) for engagement prediction, supporting adaptive, personalized learning experiences.

- In mathematical problem-solving, this research investigates curriculum learning and bilingual fine-tuning on complex mathematical datasets, particularly in geometry. The work proposes specialized datasets like GPSM4K and Math-Quest to support LLMs in handling structured solution formats and bilingual contexts, comparing models like WizardMath-7B for English and Hindi tasks. These contributions advance AI's role in multilingual STEM education by addressing language-specific needs in mathematical reasoning.
- For multimodal numerical reasoning, especially in secondary-level geometry, the thesis provides new datasets and enhanced model architectures. By incorporating both textual and visual information, this research improves AI systems' ability to interpret and solve complex geometry problems, expanding their utility in mathematics education and bridging gaps in existing geometry-focused resources.
- In Grammatical Error Correction (GEC), this work introduces the Dynamic Context Learner (DCL), which dynamically incorporates contextual cues to improve the accuracy of grammar correction over extended texts. This approach enhances traditional GEC models by capturing a broader context, significantly improving performance in academic and technical writing, especially in scientific and multi-sentence contexts.
- Finally, in Citation Text Generation (CTG), this research presents a multi-source architecture (M-CTG) that integrates knowledge graphs and keyphrase embeddings to improve contextual relevance and accuracy in citation generation. By fine-tuning models like Vicuna with specialized datasets such as CTG-S2ORC and employing advanced prompting techniques, this contribution streamlines citation practices in scholarly writing, offering a robust solution for accurate, context-aware citation generation.

Through these contributions, this thesis demonstrates the potential of LLMs and multimodal models to address complex challenges in education and research. By developing targeted datasets, refining model architectures, and applying innovative fine-tuning strategies, this work lays the foundation for more effective, context-aware applications of AI in specialized fields.

## 1.2.2 Dissertation Organization

The subsequent chapters of this thesis are organized as follows.

**Chapter 2** discusses the literature review that explores advancements in educational and scientific domains, highlighting AI applications in question answering, engagement prediction, geometry problem-solving, grammatical error correction, and citation

generation, with a focus on addressing gaps in dataset quality, contextual coherence, and model adaptability for educational and research needs.

**Chapter 3** presents the MM-PhyQA dataset and explores AI techniques for high school physics question answering, particularly through multimodal approaches, demonstrating the effectiveness of multi-image Chain-of-Thought prompting and fine-tuning techniques like MI-CoT to enhance reasoning accuracy, with future research aimed at expanding these methodologies across other STEM disciplines.

**Chapter 4** introduces engagement level classification across various models, including EfficientNet, ViT, GLAMOR-Net, and the proposed NeuralGaze, using benchmarks like DAiSEE, EngageNet, and ECLIPSE, highlighting that GLAMOR-Net with OpenFace features and focal loss achieves the best accuracy, with significant improvements observed in the Content-Guided Vision Transformer (CG-ViT) for engagement prediction by incorporating content information alongside participant expressions.

**Chapter 5** investigates curriculum learning for bilingual fine-tuning on complex mathematical datasets, comparing performance across lightweight open-source models and closed-source models for tasks in English and Hindi, highlighting WizardMath-7B as a top-performing model with enhanced accuracy on the IndiMathQA dataset, and proposing datasets like GPSM4K and MathQuest to improve LLMs' capabilities in geometry and mathematical problem-solving through structured solution formats and bilingual training approaches.

**Chapter 6** presents the DCL-GEC model for grammar error correction, leveraging dynamically detected contextual information to outperform traditional static methods by integrating broader context across sentences, achieving significant improvements in error types related to tense, agreement, and verb forms, particularly in scientific and complex paragraph-level corrections across diverse datasets.

**Chapter 7** introduces the M-CTG model for generating contextually accurate citation texts, using a specialized dataset from S2ORC and fine-tuning models like LLaMA, Alpaca, and Vicuna. Vicuna excelled, especially with advanced prompting methods (Zero-Shot CoT, Few-Shot CoT, Auto-CoT) and the integration of knowledge graphs for richer context. Future research will focus on author-specific embeddings and enhanced reasoning, showing the potential of combining LLMs and knowledge graphs for citation generation.

**Chapter 8** explores future works and advancements in AI models, particularly in education, engagement detection, geometry, grammatical error correction, citation generation, and controlled text synthesis, highlighting contributions to domain-specific reasoning and multimodal processing with promising directions for future research.

# CHAPTER 2

## Related Works

### 2.1 Introduction

The rapid evolution of AI and machine learning has led to transformative applications across diverse fields, yet there remain significant challenges when adapting these models to the specialized demands of both educational and scientific domains. In education, models are required to handle complex, multimodal information and to support nuanced forms of reasoning, question answering, and engagement tracking—essential skills for fostering effective learning experiences in subjects like physics, geometry, and mathematics. Equally critical is the need to adapt AI tools to online learning environments, where traditional, in-person engagement cues are absent. Here, models must not only predict student engagement levels accurately but also respond to emotional states like confusion or frustration, supporting a more adaptive and personalized virtual learning experience.

In scientific domains, the demands on AI models are similarly specialized, requiring models that can generate contextually accurate text, navigate domain-specific nuances, and uphold stringent standards of scholarly integrity. Tasks such as citation text generation, grammatical error correction, and controlled text synthesis require a level of contextual awareness and precision that general-purpose models struggle to achieve. The ability to manage scientific language and to produce structured, coherent content is especially crucial in supporting accurate and reliable academic writing, an area where precision is paramount.

The sections that follow are organized into two core areas: the educational domain and the scientific domain, each addressing unique yet interrelated challenges. By first examining the educational domain, this chapter delves into foundational works that have shaped AI's role in learning, from multimodal question-answering systems to student engagement analysis. These works illustrate how AI can augment educational outcomes, enabling models to act as personalized assistants in various subject areas.

Next, the focus shifts to the scientific domain, where AI's contributions to grammatical accuracy, citation generation, and structured scientific text are explored. This section highlights the foundational methodologies and recent advancements that are paving the way for more robust AI tools capable of supporting scientific research and academic communication.

The inclusion of both domains in this thesis is intentional, as the educational and scientific contexts are not only critical in their own right but are also closely related within the scope of AI's impact on knowledge dissemination and skill acquisition. In exploring the related works across these domains, this chapter establishes a comprehensive foundation for advancing AI applications that bridge educational support and scientific communication, enabling a seamless integration of context-sensitive reasoning, multimodal problem-solving, and controlled text generation across these vital areas.

## **2.2 Educational Domain**

The educational domain and its advancements in AI play a critical role in shaping the capabilities of language models for scientific applications. As educational AI research focuses on enhancing reasoning skills and multimodal comprehension, the methodologies and models developed in this area naturally benefit scientific AI tasks that require complex reasoning and the integration of textual and visual information. For instance, improving question answering and engagement prediction in educational contexts strengthens AI's ability to analyze, interpret, and respond accurately within structured knowledge frameworks, skills which are transferable to scientific applications such as problem-solving and research analysis. Thus, exploring related works in educational AI not only supports advancements in learning environments but also provides foundational techniques that enhance the application of AI in scientific contexts, creating a symbiotic relationship between the two fields.

### **2.2.1 Physics Question Answering**

The application of AI in education has seen substantial advancements, particularly in areas like intelligent tutoring, automated grading, and adaptive learning. These systems have significantly enhanced educational processes. In the context of question genera-

tion and augmentation, AI models have shown great potential. [Kumar et al. \(2022\)](#); [Mandlecha et al. \(2023\)](#) introduced a data augmentation technique that has become foundational for subsequent research, emphasizing the role of AI in improving educational outcomes.

Several datasets have been developed for mathematics and science domains, yet there is a notable gap in high-quality, challenging physics question datasets. Existing datasets like JEEBench [Arora et al. \(2023\)](#) and GSM8K [Cobbe et al. \(2023\)](#) contain questions tailored to specific educational levels, but they either surpass or fall short of the difficulty required for high school physics. For example, JEEBench contains advanced questions unsuitable for typical high school students, while GSM8K is more appropriate for middle school levels. SciQ [Welbl et al. \(2017\)](#) offers a large number of science questions, but they are limited to multiple-choice formats, lacking the complexity needed for in-depth learning.

In the realm of LLMs, [Bryant \(2023\)](#) demonstrated their efficacy in academic settings, particularly in physics. However, challenges such as replication issues were highlighted, which could be mitigated through techniques like prompt engineering. Few-shot prompting [Zhao et al. \(2021\)](#), a method where models are guided by a few examples, has shown promise in generating content aligned with specific educational curricula, such as high school physics. CoT prompting [Wei et al. \(2023\)](#) further enhances model performance by breaking down reasoning tasks into sequential steps, making it particularly useful for generating structured educational content.

Multimodal datasets like ScienceQA [Mandlecha et al. \(2023\)](#), GeoQA [Chen et al. \(2022\)](#), and TQA [Welbl et al. \(2017\)](#) have expanded the scope of educational datasets by incorporating images and text, though they often lack challenging questions suitable for high school levels. Recently, SceMQA [Liang et al. \(2024\)](#) addressed this gap by introducing a multimodal benchmark specifically targeting the college entrance level, covering subjects such as physics with a mix of multiple-choice and free-response formats. It includes knowledge-point annotations and varied question forms to better evaluate reasoning depth in AI models. Large multimodal models, such as GPT4-V [Yao et al. \(2023\)](#) and PaLM-E [Zhang et al. \(2020b\)](#), have pushed the boundaries of educational AI by integrating visual features, enabling more sophisticated question-answering tasks. The CoT paradigm has revolutionized reasoning in language models, evolving into more complex forms like Tree-of-Thoughts and Graph-of-Thoughts. Despite these

advancements, the application of multimodal CoT [Cobbe et al. \(2023\)](#) faces limitations, particularly in scenarios requiring diverse and robust training datasets. This is further validated by the PHYSICS benchmark [Feng et al. \(2025\)](#), which rigorously evaluates foundational models on university-level physics problems. This demonstrated that even with applying diverse prompting strategies, and Retrieval-Augmented Generation (RAG)-based knowledge augmentation, models fall short in deep physics reasoning.

The field has also seen efforts to enhance domain-specific question answering through retrieval-augmented generation techniques. By incorporating relevant context from external knowledge sources like Wikipedia, models can improve their accuracy in answering complex questions. This approach has achieved SOTA results in open-domain benchmarks such as Natural Questions [Chen et al. \(2022\)](#) and TriviaQA [Welbl et al. \(2017\)](#), and has potential applications in physics and math domains where domain-specific knowledge is crucial.

In summary, while significant progress has been made in educational AI, particularly in question generation and multimodal datasets, there remains a need for more specialized datasets and techniques that address the specific challenges of high school-level education, particularly in physics. Our work builds upon these foundations, introducing novel methodologies to fill these gaps and enhance the effectiveness of AI in education.

### 2.2.2 Online Learning

Student engagement prediction has been a significant focus in educational research, with various approaches and datasets developed to address the challenge. Early systems, utilized Kinect sensors [Zaletelj and Košir \(2017\)](#) to estimate attention levels through visual features like gaze and body posture. Similarly, [Goldberg et al. \(2021\)](#) analyzed engagement using gaze and facial action units, achieving a correlation of 0.44 with manual annotations.

Several datasets have been created to support engagement detection research. The DAiSEE dataset by [Gupta et al. \(2016\)](#) is widely used and includes videos of 112 participants, annotated across four levels of engagement, boredom, confusion, and frustration. The HBCU dataset [Whitehill et al. \(2014\)](#) includes data from 34 individuals participating in cognitive skills training, with annotations focusing on engagement. [Kaur et al.](#)

(2018) introduced the EngageWild dataset, featuring videos from 78 individuals with crowd-sourced annotations classifying engagement into four levels. Sathyanarayana *et al.* (2014) presented the SDMATH dataset, offering richly labelled data from one-to-one math tutoring sessions, including video and audio modalities. The EngageNet dataset Singh *et al.* (2023) focuses on classifying engagement into four levels, comprising 31 hours of video from 127 participants, segmented into 10-second clips.

Beyond dataset creation, research has explored advanced models for interpreting complex engagement signals. Kumar *et al.* (2024) introduced a cross-modality dual-pipeline approach specifically integrating visual and audio features using transformer encoders and a cross-modal fusion mechanism. Their method demonstrated competitive performance in estimating engagement levels by effectively combining information from different modalities.

Despite these efforts, challenges remain, particularly the scarcity of large, publicly available datasets and issues like class imbalance, which hinder the analysis of disengaged students. Most datasets focus on short video segments, limiting the assessment of engagement over longer periods. Additionally, while video-based methods are commonly used, they may not be practical for low-resource educational settings, highlighting the need for more efficient approaches, such as single-frame-based prediction methods.

Recent advancements in neural cellular automata (NCA) Chopard and Droz (1998); Bays (2010) have shown promise in various tasks, including image generation. NCAs, inspired by biological growth, use neural networks to define update rules, making them adaptable and lightweight. This approach contrasts with conventional deep learning models that analyze entire images at once, as NCAs focus on individual pixels and their interactions with neighbours.

In summary, while substantial progress has been made in developing tools and datasets for student engagement prediction, significant gaps remain, particularly in handling long-term engagement and creating more accessible methods for diverse educational environments.

### 2.2.3 Geometric Question Answering

Research in geometry problem-solving and multimodal datasets has significantly advanced with the development of various specialized datasets and LLMs.

The field of Geometry Multimodal Datasets has seen significant advancements with resources like the PGPS9K dataset [Zhang \*et al.\* \(2023\)](#), which is a comprehensive collection featuring 9,022 geometry problems with corresponding diagrams, curated from grades 6-12 textbooks. This dataset is unique in providing both diagram annotations and solution programs, covering a wide range of plane geometry problems with a strong emphasis on theorem-based problem-solving. Similarly, Geometry3K [Lu \*et al.\* \(2021\)](#) is a large-scale benchmark derived from high school textbooks and online libraries, focusing on complex problems, including irregular polygons. However, it lacks annotated theorem application sequences due to the intensive resource requirements for such annotations. Geo170K [Gao \*et al.\* \(2023\)](#) further advances the field by synthesizing geometric visual-text data from existing datasets, offering an expansive collection of 60,000 geometric image-caption pairs and over 110,000 question-answer pairs, significantly surpassing previous datasets like GeoQA++ in scale and coverage. To specifically evaluate the capabilities of LLMs and multimodal models on geometry problem-solving, [Zhang \*et al.\* \(2024\)](#) introduced GeoEval, a benchmark featuring diverse subsets, including problems requiring backward reasoning. Their comprehensive evaluation highlighted significant performance gaps on challenging and unseen problems, underscoring the necessity for robust evaluation methods that go beyond standard training data.

The domain of Multimodal LLMs includes advanced models like LLaVA, which integrates LLMs, such as GPT-4, with vision encoders like CLIP [Radford \*et al.\* \(2021\)](#), designed to process and act on multimodal instructions. This model enhances interactivity and adaptability by utilizing both visual and textual inputs, setting it apart from models primarily focused on image description. Building upon this, GLLaVA incorporates LLAMA-2 [Touvron \*et al.\* \(2023a\)](#) for language understanding and a Vision Transformer (ViT) [Radford \*et al.\* \(2021\)](#) as the image encoder, further strengthening its ability to solve geometric problems through a two-phase training process involving geometric visual-language alignment and instruction tuning. Further advancing MLLM architectures, MoVA [Zong \*et al.\* \(2024\)](#) addresses the limitations of single vision encoders on diverse image content, including geometry diagrams. It utilizes a Mixture

of Vision Experts adaptively routed and fused based on multimodal context to enhance generalization ability across various visual inputs.

The Need for Dataset Creation in geometry arises as existing datasets often fall short in complexity, particularly in handling numerical problems and theorem-based questions commonly encountered in high school geometry. These datasets typically lack the detailed, step-by-step solutions necessary for learning and error correction, underscoring the need for specialized datasets that cover a broader spectrum of geometric complexities. This includes advanced numerical tasks, theorem-based queries, and comprehensive solutions to better support educational and research efforts in geometry problem-solving.

This combined research highlights the ongoing developments in geometry problem-solving and the growing importance of multimodal models and datasets that integrate both visual and textual data to enhance problem-solving capabilities.

## **2.3 Scientific Domain**

The scientific domain leverages advances in AI from educational contexts to solve complex tasks, ranging from citation generation to grammar correction in academic writing. The techniques and datasets developed in education, such as engagement prediction and multimodal question-answering, underpin many scientific applications where context-aware and structured reasoning are essential. By improving educational AI tools, such as those for engagement and question answering, the scientific domain can adapt these methodologies for tasks like accurately generating citations, enhancing grammatical precision, and controlling the intent and context of generated text. This connection underscores the importance of exploring related works in scientific AI, as advancements in education directly contribute to more robust, domain-specific AI tools for research, further bridging the gap between learning environments and scientific rigour.

The educational domain is fundamentally connected to the scientific domain, as advancements in educational AI directly support scientific applications that demand structured knowledge and deep contextual understanding. Techniques like multimodal reasoning, developed for educational tasks such as physics question answering and student engagement prediction, are essential for scientific domains where interpreting complex,

context-rich data is crucial. For instance, methods used in question answering for educational content can enhance citation generation and grammatical error correction in research by improving context sensitivity and synthesis of domain-specific information. By building on the innovations in educational AI, this thesis extends their impact into scientific applications, creating AI tools that enhance both the learning and research processes, thereby bridging the educational and scientific domains in a mutually reinforcing manner.

### 2.3.1 Grammatical Error Correction

Since the CoNLL' 14 Shared Task, advancements in GEC have significantly accelerated. In 2016, [Rozovskaya and Roth \(2016\)](#) marked notable progress by integrating error-specific classifiers with a Phrase-Based Machine Translation (PBMT) model trained on the Lang-8 dataset. That same year, [Junczys-Dowmunt and Grundkiewicz \(2016\)](#) further enhanced GEC capabilities by combining a PBMT model with comprehensive language modeling and bi-text features.

A few years later, the LaserTagger method by [Malmi \*et al.\* \(2019\)](#) redefined text generation as a text-editing task, utilizing a BERT encoder with an auto-regressive transformer decoder. This approach improved grammatical correction tasks significantly, offering faster inference times compared to traditional seq2seq models. Building on these advancements, GECToR emerged as a method utilizing language models and linear classifiers for token-level corrections. Employing iterative sequence tagging [Awasthi \*et al.\* \(2019\)](#) and pre-trained transformers like RoBERTa [Liu \*et al.\* \(2019\)](#), XLNet [Yang \*et al.\* \(2019\)](#), and BERT [Devlin \*et al.\* \(2019a\)](#), GECToR leverages personalized grammatical transformations to better capture error-context relationships. Another recent approach, DeCoGLM [Li and Wang \(2024\)](#), proposed an integrated detection-correction structure based on a General Language Model (GLM). Their method enables simultaneous error detection and correction through multi-task learning within a single LLM, improving their performance on GEC tasks.

Moving towards document-level strategies, [Yuan and Bryant \(2021a\)](#) explored methods to enhance GEC performance by integrating broader context within NMT-based GEC. They introduced a three-step training approach to optimize parallel data usage at both sentence and document levels. [Chollampatt \*et al.\* \(2019\)](#) emphasized the impor-

tance of context preservation at both sentence and document levels, highlighting that many grammatical fixes rely heavily on context often overlooked in single-sentence approaches.

Evaluating GEC system performance is a critical step, [Kobayashi \*et al.\* \(2024\)](#) introduced SEEDA, a new dataset specifically designed for meta-evaluation. This benchmark features human corrections rated at both edit-based and sentence-based granularities for numerous state-of-the-art systems, including LLMs, aiming to address biases and outdated setups in conventional meta-evaluation methodologies.

The current challenges in GEC models largely stem from their reliance on single-sentence approaches or limited context handling. Effective grammar correction requires methods to maintain context in longer texts, which could be improved by segmenting sequences into manageable segments without losing essential context. This strategy would enhance both the efficiency and speed of GEC processes, particularly for document-level corrections.

### 2.3.2 Citation Generation

CTG intersects with several key areas, including citation recommendation, scientific document understanding, and summarization. Citation recommendation systems [Bhagavatula \*et al.\* \(2018\)](#) guide researchers to valuable sources, complementing CTG by providing references for specific documents or text excerpts. Additionally, summarization systems [Yasunaga \*et al.\* \(2019\)](#) condense information, aiding in the quick comprehension of research content. Notably, citation information is beneficial for scientific paper summarization and in multi-document summarization tasks [Chen and Zhuge \(2014, 2019\)](#).

[Koncel-Kedziorski \*et al.\* \(2019\)](#) advanced multi-sentence text generation from information extraction systems by incorporating knowledge graphs through Graph Attention Networks. [Chen \*et al.\* \(2021\)](#) addressed context-aware text generation in scientific domains with the SciXGen dataset, while [Zhu \*et al.\* \(2023\)](#) leveraged LLMs for constructing and reasoning with knowledge graphs, proposing the AutoKG system for this purpose. These studies highlight the importance of context in scientific text generation, a theme central to our work.

Two parallel works have focused specifically on generating citation texts from re-

search papers. [Luu et al. \(2020\)](#) pioneered this task, generating citation texts using both the source and cited documents. [Xing et al. \(2020a\)](#) expanded on this by exploring the relationships between scientific documents with a larger dataset, employing an implicit citation extraction algorithm enhanced by GPT-3. These studies provide a foundation for our approach to CTG.

In the domain of multi-reference citation generation, [Jung et al. \(2022\)](#) introduced a framework for controllable citation generation, using BART and T5 transformers and focusing on different citation intents such as background, methods, and results. [Gu and Hahnloser \(2022\)](#) proposed a pipeline for controllable citation generation, which includes an attribute recommendation module and a conditional citation generation module, demonstrating controllability across various characteristics. These works underscore the ongoing advancements in CTG and its applications.

Extractive and abstractive methods have both played significant roles in CTG. Early extractive approaches, such as keyword-based trees [Hoang and Kan \(2010\)](#) and Probabilistic Latent Semantic Analysis (PLSA) [Hu and Wan \(2014\)](#), focused on sentence selection from cited papers. However, these methods were limited by intellectual property concerns and a lack of engagement with the source material. Abstractive methods, pioneered by studies like [Hu and Wan \(2014\)](#); [Li et al. \(2018\)](#), aimed to generate more original and contextually relevant summaries. A notable example is the multi-source, pointer-generator network with cross-attention by [Xing et al. \(2020b\)](#), which allowed for nuanced interactions between citing contexts and cited abstracts.

More recent work, such as [Ge et al. \(2021\)](#), integrates background knowledge from citation networks and content from both citing and cited papers, aligning closely with our focus on multi-source inputs. [Gu et al. \(2022\)](#) introduces a novel extractive summarization approach using a multi-step episodic Markov Decision Process to reduce redundancy and generate compact summaries, providing valuable insights for our model development. Finally, [Li et al. \(2024\)](#) addresses the limitations of previous methods by retrieving cited text spans (CTS) from the full text, emphasizing the importance of grounding citations in comprehensive content, a principle that informs our framework’s diverse input use.

This body of work provides a robust foundation for the development of more sophisticated CTG models, emphasizing the importance of context, multi-source inputs, and innovative methodologies in advancing the field.

Recent advancements in citation generation focus on enhancing control, relevance, and contextual coherence in generated citation texts. Our approach introduces a novel framework designed to optimize citation generation accuracy by integrating user-defined keywords and intent-based constraints. Specifically, we employ a refined extraction model that builds upon the prior work of (Gu and Hahnloser, 2022), who demonstrated the effectiveness of keyword extraction via KeyBERT and sentence relevance using SentenceBERT, achieving enhanced precision in identifying key phrases for citations. Extending this, our model utilizes a fine-tuned BERT-based architecture with custom loss functions, which facilitates more nuanced alignment with user-specified citation contexts.

Furthermore, unlike prior methods that often rely on static models or limited citation categories (Cohan *et al.* (2019)), we incorporate a dynamic generation process that leverages both structured and unstructured scientific data to better adapt to varied citation intents, such as "Background," "Method," and "Result." This flexibility addresses noted limitations in previous systems, such as those by (Xing *et al.*, 2020a) and (Wu *et al.*, 2021), which do not accommodate multi-study summarization or customizable intent. Our model's evaluation highlights a significant improvement in citation relevance and interpretability, demonstrating a promising direction for future citation generation research.

### 2.3.3 Scientific Controllable Text Generation

Controllable Text Generation (CTG) has emerged as a crucial area in Natural Language Processing (NLP), particularly for applications that require precise customization of generated text to align with specific stylistic, structural, or semantic criteria. CTG enables models to generate text with properties such as intent, style, formality, and topic, which is especially valuable in scientific and technical fields where content coherence, accuracy, and relevance are critical. Research in CTG has proposed various frameworks and methodologies to guide model outputs effectively. For instance, (Goel *et al.* (2023)) introduced a modular schema for CTG that structures the text generation pipeline into seven interlinked components. This framework integrates external inputs, encoding and decoding strategies, and training objectives, each contributing to generating contextually aligned and user-specific text. Such a controlled approach, where modules are

fine-tuned across different stages of the generation process, proves versatile in scientific CTG.

A central challenge in CTG is guiding the initialization process to meet desired output properties. [Liu and Lapata \(2018\)](#) tackled this by decomposing encoder representations into subspaces to separate content and structure, which has been applied in tasks like abstractive summarization [Balachandran et al. \(2020\)](#). In the scientific domain, [Xing et al. \(2020a\)](#) employed dual encoders to process citing and cited papers independently, thereby creating more accurate citation texts by maintaining distinct hidden states for each document. Further advancing this approach, [Gu and Hahnloser \(2022\)](#) incorporated local context, defined as specific text windows surrounding citations, preserving critical information essential for accurate citation generation.

Encoding operations play a pivotal role in aligning model outputs with user-defined criteria throughout the generation process. Techniques like Prefix Tuning [Li et al. \(2021\)](#) and P-Tuning [Liu et al. \(2021\)](#) allow controlled generation with minimal computational overhead by using a small set of parameters prepended to input text. Auto Prompt, introduced by [Shin et al. \(2020\)](#), is a gradient-based method for automating prompt creation, achieving controlled task-specific outputs across domains and enhancing model flexibility in meeting diverse scientific writing requirements. Discriminators also provide valuable guidance in CTG by offering feedback to adjust the generator's latent space. Adversarial loss functions, extensively used to regulate encoding and ensure outputs align with attributes like style or sentiment, are particularly beneficial. [Romanov et al. \(2018\)](#) demonstrated the effectiveness of adversarial loss in producing stylistically neutral outputs, which is critical for the objectivity required in scientific writing. Additionally, [Gu and Hahnloser \(2022\)](#) used cross-entropy loss to fine-tune SCIBERT [Beltagy et al. \(2020\)](#), employing triplet loss to align citations with user-defined keywords, thereby improving precision and relevance.

Decoding strategies significantly impact the structural and semantic integrity of generated content. [Holtzman et al. \(2019\)](#) introduced Top-P (Nucleus) Sampling, a method that dynamically adjusts candidate tokens based on cumulative probability, enhancing contextual alignment. Additionally, [Gu et al. \(2017\)](#) proposed Trainable Decoding, wherein a reinforcement learning (RL) agent optimizes decoding to meet specific objectives, which has proven useful in CTG applications that require strict formatting or intent-based constraints. Tailored loss functions also play a critical role in CTG, with

standard cross-entropy often supplemented by custom functions to meet unique generation goals. For example, Unlikelihood Loss [Welleck \*et al.\* \(2019\)](#) reduces repetition by penalizing frequent tokens, while Coverage Loss [See \*et al.\* \(2017\)](#) addresses redundancy issues in extended sequences, aiding in the production of clearer scientific documentation.

Fine-tuning remains essential for adapting general-purpose models to domain-specific tasks. RL fine-tuning, as discussed by [Ranzato \*et al.\* \(2015\)](#), optimizes non-differentiable rewards, such as BLEU and ROUGE, enhancing summarization and translation performance. Reinforcement learning with human feedback, particularly effective for aligning generation with human preferences [Christiano \*et al.\* \(2017\)](#); [Ziegler \*et al.\* \(2019\)](#), could notably enhance citation accuracy and readability in scientific documents. Pre-trained large language models (LLMs) have opened new possibilities in CTG for scientific contexts. For instance, Galactica [Taylor \*et al.\* \(2022\)](#) is specifically tuned for scientific knowledge generation, while LLAMA and its variants [Touvron \*et al.\* \(2023a\)](#) address broader NLP tasks like summarization and citation generation. Techniques like Chain-of-Thought (CoT) prompting [Wei \*et al.\* \(2022\)](#) facilitate structured reasoning, beneficial for generating coherent scientific text. Prompt-tuning strategies, further refined by auto-CoT and Mixture of Soft Prompts (MSP) [Chen \*et al.\* \(2023\)](#), support complex reasoning without extensive reconfiguration.

Overall, advancements in CTG methodologies—from encoding and decoding strategies to custom loss functions and prompt-based fine-tuning—are progressively enabling large language models to produce precise, contextually appropriate scientific content. This growing body of work establishes a robust foundation for further research in CTG, with a strong focus on adaptable, controlled generation frameworks tailored to the unique demands of scientific and technical writing.

# CHAPTER 3

## Physics Question Answering

### 3.1 Motivation

Physics education relies on the effective integration of both textual and visual elements, as students learn concepts through a combination of descriptions, equations, and diagrams. While Large Language Models (LLMs) have shown potential in processing natural language, their application in high school physics remains limited, especially where multimodal understanding is required. Current models often fall short in this domain, as physics problems frequently demand not only accurate comprehension of technical language but also an ability to interpret visual information and follow structured reasoning steps.

To address these challenges, this work focuses on developing resources and techniques tailored for physics education, introducing a multimodal dataset specifically designed for high school physics questions. By fine-tuning LLMs with carefully curated content, this research aims to improve the capacity of these models to support physics learning, ultimately creating tools that can enhance students' problem-solving experience and deepen their conceptual understanding.

### 3.2 Introduction

The rapid advancement of transformer-based LLMs has brought significant improvements in tasks such as text generation, question answering, and document summarization. Recent models like GPT-3.5, PaLM, and LLAMA have demonstrated remarkable capabilities in understanding and generating human-like text. However, when applied to domain-specific challenges, such as solving physics problems, these models often fall short of achieving the desired accuracy. Addressing this gap, our work 'SciPhyRAG' [Anand \*et al.\* \(2023b\)](#) focuses on enhancing the performance of LLMs in physics question answering. By fine-tuning a model on a specially designed high school physics

dataset and incorporating retrieval-augmentation techniques, this research aims to improve the reasoning capabilities of LLMs, particularly in handling the complex and knowledge-intensive nature of physics problems. The study highlights the creation of a high-quality physics corpus based on NCERT textbook content, which serves as a benchmark for evaluating the model's performance and pushes the boundaries of LLM accuracy in computational problem-solving.

Building upon the theme of improving physics education through advanced models, our work [Anand \*et al.\* \(2023a\)](#) delves into the application of LLMs in educational contexts, specifically in high school physics. With the increasing demand for innovative teaching methodologies, this thesis contributes by developing and augmenting a high school-level physics question dataset. This dataset, derived from NCERT exemplar solutions, was expanded from an initial 766 questions to approximately 8000, using controllable text generation techniques. The dataset's creation is aimed at enriching the educational content available for AI models, enabling these models to provide more personalized and accurate solutions to physics problems, thereby enhancing the learning experience. The research also evaluates the effectiveness of the Llama2-7B model fine-tuned on this dataset, exploring the potential of LLMs in adapting to the diverse needs of students.

Moreover, Another one of our works [Anand \*et al.\* \(2024c\)](#) addresses the complexities of multimodal physics question answering, an area where LLMs and Large Multimodal Models (LMMs) have shown promise but still face challenges. Physics problems often involve both textual and visual components, such as diagrams, which add layers of complexity that models must interpret and understand. To tackle this, the research introduces the MM-PhyQA dataset, a novel multimodal dataset designed to evaluate LMMs' performance on physics questions. Additionally, the study explores the use of CoT Prompting in a multimodal context, developing a unique method to incorporate multiple images during the CoT prompting process. This approach aims to enhance the model's problem-solving abilities by providing a more comprehensive understanding of the question's context, further pushing the capabilities of LLMs and LMMs in educational applications.

Collectively, these three papers contribute to the broader field of physics question answering by exploring different facets of LLM and LMM performance enhancement, from retrieval-augmentation techniques and dataset creation to multimodal problem-

solving approaches. The insights gained from these studies have the potential to revolutionize physics education, paving the way for more intelligent tutoring systems and personalized learning experiences that can effectively support students in mastering complex scientific concepts.

Key contributions of this work are as follows:

- Presented the largest and most comprehensive high school physics question dataset based on NCERT exemplars, specifically designed to fine-tune LLMs for educational tasks in physics.
- Developed and fine-tuned a physics-specific retrieval pipeline (SciPhy-RAG) that incorporates contextually relevant passages to improve the model’s accuracy and explainability for complex physics question-answering tasks.
- Employed innovative data augmentation techniques to expand a base set of 766 questions to over 8,000 diverse physics problems, ensuring the dataset covers a wide array of high school physics topics and difficulty levels.
- Explored the effectiveness of CoT prompting, showing that the method significantly improves the model’s performance on both accuracy and explanation quality, verified by multiple evaluation metrics.
- Provided an open-source dataset with manually annotated LaTeX equations and explanations, setting a benchmark for future research in STEM-based question answering and AI-driven education technology.

## 3.3 Experiments

### 3.3.1 Novel Physics Education Dataset

#### 3.3.1.1 Dataset Description

Our original dataset consists of 766 questions generated from physics problems. After augmentation, our dataset consists of 7983 questions. Table. [3.1](#) shows the distribution of questions across question types and topics. A majority are subjective questions, as augmenting subjective problems was simpler. The questions in the dataset cover a wide range of topics in physics, including mechanics, electromagnetism, thermodynamics, optics, and atomic physics. The questions are designed to test students’ understanding and problem-solving skills in physics.

### 3.3.1.2 Sample Question in Physics Dataset

**Instruction:** A circular current loop of magnetic moment  $M$  is in an arbitrary orientation in an external magnetic field  $B$ . The work done to rotate the loop by  $300^\circ$  about an axis perpendicular to its plane is:

**Input:** (a)  $MB$ , (b)  $\frac{\sqrt{MB^2}}{2}$ , (c)  $\frac{MB}{2}$ , (d) Zero

**Output:** The work done to rotate the loop by  $300^\circ$  about an axis perpendicular to its plane is zero, as the correct answer is, **(d) Zero**.

Table 3.1: Distribution of Questions by Chapter

Chapter	Number of Questions
Units and Measurements	95
Motion in a Straight Line	121
Motion	183
Laws of Motion	336
Work, Energy and Power	148
System of Particles and Rotational Motion	176
Gravitation	260
Mechanical Properties of Solids	176
Mechanical Properties of Fluids	169
Thermal Properties of Matter	527
Thermodynamics	624
Kinetic Theory	1227
Oscillations	739
Waves	933
Electric Charges and Fields	214
Electrostatic Potential and Capacitance	216
Current Electricity	427
Moving Charges and Magnetism	116
Magnetism and Matter	323
Electromagnetic Induction	129
Alternating Current	110
Electromagnetic Waves	78
Ray Optics and Optical Instruments	105
Wave Optics	90
Dual Nature of Radiation and Matter	96
Atoms	127
Nuclei	101
Semiconductor Electronics: Materials, Devices and Simple Circuits	137
Communication Systems	107
Total	7983

Figure 3.1 illustrates the distribution of question lengths throughout the corpus. The

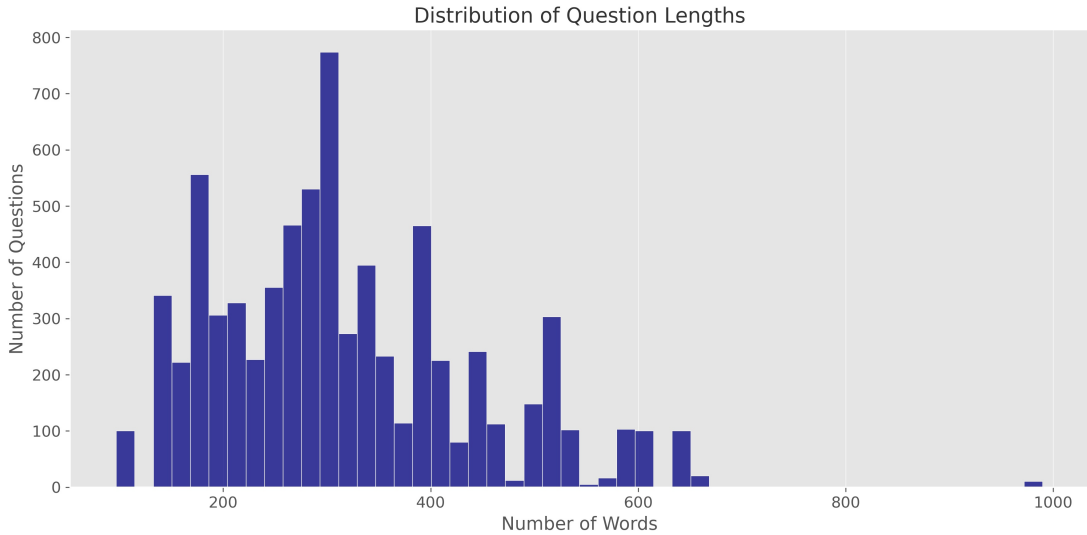


Figure 3.1: Question length distribution

average question prompt length of approximately 350 words allows LLMs to hold the question prompts in their Context Window without much issue.

### 3.3.1.3 Data Collection

To build a dataset of over 8000 physics questions, we employed a semi-automated process as follows:

- Physics question banks from Indian high school textbooks (such as NCERT Exemplar) were scraped and formatted using open-source OCR tools, with all questions standardized into TeX format.
- The formatted questions were organized into a key-value pair structure.
- A combination of the Self-Instruct method [Wang \*et al.\* \(2023\)](#) and a custom Python script was used to perform data augmentation for selected questions.

The primary dataset source was the NCERT exemplar solutions, chosen for their extensive coverage of essential physics topics at the high school level. Each question entry in the dataset’s JSON file contains an instruction (the question text), an input (options if a multiple-choice question or otherwise left blank), and an output (the answer).

Data augmentation expanded the dataset using a methodology based on [Kumar \*et al.\* \(2022\)](#); [Mandlecha \*et al.\* \(2023\)](#), which enabled an increase in question count from 766 to 7983.

#### **3.3.1.4 Student Survey**

To assess the dataset's effectiveness and alignment with the "Chain of Thought" (CoT) strategy, a survey of approximately 500 participants was conducted. The survey aimed to evaluate students' familiarity with high school physics and determine how well the first three questions in a sequence prepared them for a final question.

#### **3.3.1.5 Survey Design**

The survey contained the following components: A question to rate the participants' familiarity with high school physics on a scale from 1 (Weak) to 5 (Strong). Five sets of questions where participants were shown an image (representing a series of physics questions) and asked to rate on a scale of 1 to 5 how well the first three questions in the set provided a basis for answering the final question.

#### **3.3.1.6 Survey Findings**

**Familiarity with High School Physics:** The majority of participants (approximately 65 percent) rated their familiarity with high school physics between 3 and 4, indicating a moderate to strong grasp of the subject.

**Basis for Answering the Final Question:** On average, participants rated the effectiveness of the first three questions in providing a basis for the final question at 3.8. This suggests that most students found the initial questions effective in preparing them for the concluding question.

#### **3.3.1.7 Analysis and Implications**

The survey results demonstrate that the dataset's structure, particularly the CoT approach, effectively supports students' logical reasoning. The moderate to strong familiarity among participants strengthens the feedback's reliability. These findings highlight the dataset's potential in teaching and assessment, focusing not only on knowledge retention but also on fostering analytical progression.

### 3.3.1.8 Data Augmentation

Data augmentation is key to enhancing dataset diversity. The technique used from [Kumar \*et al.\* \(2022\)](#); [Mandlecha \*et al.\* \(2023\)](#) provided the foundation for this process, expanding the dataset from 766 to 7983 questions.

Custom augments scripts were developed for each physics chapter. These scripts used a template-based approach, populated with randomized elements from predefined lists. The augmentation introduced variability in:

- **Values Variation:** Randomization of numerical values to prevent bias towards specific answers.
- **Conceptual Variation:** Rephrasing to offer different perspectives on the same concept.
- **Structural Variation:** Changing linguistic patterns to avoid memorization.

This structured augmentation ensured comprehensive coverage across various topics while maintaining scientific accuracy and contextual relevance. It was designed to strike a balance between exposing the model to diverse data and adhering to core curriculum principles.

### 3.3.1.9 Construction of a CoT Dataset

Central to our experimental setup is the CoT method [Wei \*et al.\* \(2023\)](#), aimed at enhancing the model’s reasoning through a sequence of related questions. This was achieved using BERT [Devlin \*et al.\* \(2019b\)](#) embeddings and cosine similarity. The steps are as follows:

1. **Data Loading:** All questions and related metadata were loaded into the Python environment, each question serving as a potential CoT link.
2. **BERT Initialization:** We used the `bert-base-uncased` model from the Hugging Face Transformers library, chosen for its contextual richness.
3. **Embedding Generation:** Each question is converted into a BERT embedding, capturing its semantic essence. These embeddings serve as the foundation for determining the logical and contextual proximity between questions.
4. **Retrieval of Similar Questions:** Harnessing the power of cosine similarity, we identify the top three questions that are most closely aligned with the query in terms of their chain of thought. These questions act as the guiding prompts for the model.

5. Prompt Generation: With the three closest questions in hand, they are ordered and presented as a sequence of prompts preceding the main query:

```
q1: Closest question in terms of embedding.  
q2: Second closest question.  
q3: Third closest question.  
query: The main query.
```

This structured sequence, rooted in the CoT methodology, primes the model to approach the main query with a specific logical progression in mind. The hope is that the model, informed by the answers to the closest questions, will produce a more accurate and contextually relevant response to the main query. In essence, our approach crafts a logical pathway for the model to follow, ensuring that its responses are not just accurate but also logically consistent with the provided chain of thought.

The culmination of this process is a set of few-shot learning prompts, each designed to guide the model’s reasoning for a specific query. This methodology promises more accurate and contextually relevant model responses.

### 3.3.1.10 Evaluation of COT

In addition to our main evaluation, we also compared CoT model performance to non-CoT models, as detailed in previous work [Wei et al. \(2023\)](#), using metrics such as BERTScore, METEOR, ROUGE, and BLEU.

### 3.3.1.11 Results

After fine-tuning the Llama2-7B model, we evaluated the model using various metrics, including BERTScores [Zhang et al. \(2020b\)](#), METEOR score [Banerjee and Lavie \(2005\)](#), ROUGE scores [Lin \(2004a\)](#), and BLEU scores [Papineni et al. \(2002a\)](#). The scores obtained are detailed in the following sections and given in Table [3.2](#).

### 3.3.1.12 Evaluation Metrics

**METEOR:** The METEOR score obtained is **0.282** and **0.221**, indicating that the generated content aligns to some extent with the reference data. However, certain domain-specific terms or phrasings common in physics might still be missing.

**BERTScore:** BERTScores compute BERT-based embeddings of the ground truth answer in the annotated dataset and the answer generated by the model. The scores obtained indicate that the model’s generated text is relatively close to the reference text in terms of content and meaning.

**BLEU:** BLEU scores measure the overlap of n-grams between the generated content and reference content. The evaluation suggests that certain n-gram overlaps are missing, which means the generated content might lack some important multi-word physics terms or phrasings.

**ROUGE:** ROUGE scores assess the quality of the summaries generated by the model. The various ROUGE scores show the degree of overlap between the generated summaries and reference summaries in terms of n-grams and sequences. The F1 scores suggest that while there’s a decent overlap of terms with the reference, there is still room for improvement.

### 3.3.1.13 Discussion

Table 3.2: Comparison of Baseline and CoT Results

Metric	Score		% Change
	Baseline	CoT	
<b>BERT Scores</b>			
F1	0.8333	0.8943	+7.32%
Precision	0.8363	0.8968	+7.24%
Recall	0.8318	0.8934	+7.40%
<b>METEOR Score</b>			
Score	0.2825	0.4992	+76.68%
<b>BLEU Scores (1-gram)</b>			
Precision	0.1123	0.3647	+224.89%
Recall	0.1571	0.3019	+92.24%
<b>ROUGE-1 Scores</b>			
F1	0.2258	0.4845	+114.65%
Precision	0.2581	0.5528	+114.18%
Recall	0.3746	0.5372	+43.44%
<b>ROUGE-L Scores</b>			
F1	0.2732	0.5307	+94.23%
Precision	0.2980	0.5880	+97.32%
Recall	0.4118	0.5757	+39.79%

Taking into account the relatively modest scale of the evaluated model, it is reason-

able to anticipate that employing a more substantial and advanced model could yield improved results. The consistency in the BERTScore implies that our model's ability to convey semantics is notable, even with current limitations. The variability in BLEU and ROUGE scores highlights specific areas for growth, particularly in capturing longer n-grams and linguistic subtleties. A larger model could potentially address these challenges and lead to enhanced performance across a wider spectrum of linguistic contexts.

Evaluation metrics, in general, can be divided into those that focus on token-level matches and those that emphasize semantic-level understanding. While BLEU and METEOR combine both these perspectives, it's important to interpret their scores with this duality in mind.

**BERT Score Discussion:** The BERT scores indicate a consistent and positive impact from utilizing CoT Data compared to the baseline. Across all three metrics (F1, Precision, and Recall), there are improvements of +7.32%, +7.24%, and +7.40%, respectively. This uniform increase suggests that the CoT Data enhances the model's ability to retain semantic meaning and precision in its responses, without sacrificing recall. It's noteworthy that the BERTScore, which is based on cosine similarity, can be influenced by good grammar. Therefore, a model that generates grammatically correct sentences can achieve a high BERTScore even if the semantic content isn't fully aligned with the reference.

**METEOR Score Discussion:** A significant enhancement is observed in the METEOR score, which leaps from 0.2825 to 0.4992, constituting a +76.68% increase when employing CoT Data. METEOR, considering both precision and recall and being particularly sensitive to synonymy, indicates that the CoT-enhanced model might be effectively utilizing synonyms and aligning phrases in a semantically relevant manner, thus achieving a higher score.

**BLEU Score Discussion:** For BLEU scores, particularly focusing on 1-gram, the Precision metric experiences an extraordinary increase of +224.89%, while Recall also experiences a significant boost of +92.24%. This suggests that CoT Data aids in producing responses that are considerably more aligned with reference sentences, especially in utilizing appropriate 1-grams. It's inherent in the nature of BLEU that unigram matches

are the most frequent, while bigram and higher n-gram matches become progressively rarer. This points towards the need for further analysis into higher n-grams to provide insights into more intricate linguistic alignments.

**ROUGE Scores Discussion:** In the context of ROUGE-1 scores, there is a remarkable increase across all metrics, with F1, Precision, and Recall improving by +114.65%, +114.18%, and +43.44% respectively. This indicates a substantial enhancement in producing unigram overlaps with reference sentences, which is crucial for maintaining semantic and informational fidelity.

Similarly, ROUGE-L (longest common subsequence) scores also witness substantial improvements, where F1, Precision, and Recall are boosted by +94.23%, +97.32%, and +39.79% respectively, when utilizing CoT Data. This underscores the model's enhanced capability to maintain coherent and informationally equivalent responses over longer subsequences of text, indicating a potential strength in preserving sentence-level informational and structural fidelity.

It's worth noting that Language Learning Models (LLMs) typically struggle with some of these metrics, which could explain the initially low scores observed. As research progresses and models evolve, we anticipate improvements in these areas.

### 3.3.2 SciPhyRAG: Retrieval Augmentation to Improve LLMs on Physics Q&A

#### 3.3.2.1 Dataset Description

In order to enhance the application of language models in physics, this study introduces a comprehensive and rigorous dataset. The research [Anand \*et al.\* \(2023b\)](#) releases two open-source datasets :

- **PhyQA** encompasses a collection of 9.5K high school-level physics questions and answers, each accompanied by step-by-step explanations. The dataset covers a broad range of topics typically studied by high school students aged between 15-19 years. These topics include Alternating Current; Atoms and Nuclei; Communication Systems; Electric Charges, Fields, and Current; Electromagnetic Induction; Electromagnetic Waves; Capacitors; Dynamics and Rotational Mechanics; Units, Dimensions, and Kinematics; Ray and Wave Optics; Thermodynamics and Heat; Gaseous State; Waves, Sound, and Oscillations. The dataset is further divided into subtopics under each primary topic. Both datasets are formatted carefully to facilitate training and evaluation on open-source language models such as

LLAMA [Touvron *et al.* (2023a)], Alpaca [Taori *et al.* (2023)], and Vicuna [Chiang *et al.* (2023)].

- The **RetriPhy Corpus** is a curated collection derived from the NCERT Physics textbooks for 11<sup>th</sup> and 12<sup>th</sup> grades. It is manually annotated with LaTeX representations of equations and examples. This dataset encompasses 14 chapters each for both grades, 11<sup>th</sup> and 12<sup>th</sup>. Upon receiving a question, the entire corpus of chapters serves to generate context passages using a retrieval pipeline.

These datasets fulfill multiple objectives. Firstly, they allow language models to be trained on complex reasoning and computational problems specific to physics, requiring a deeper comprehension of scientific principles and the relationships between variables. Secondly, these datasets can serve as benchmarks for evaluating the performance of existing language models in solving physics problems, enabling meaningful comparisons, and promoting advancements in natural language understanding for physics. Lastly, our retrieval corpus supports developing and benchmarking retrieval systems for extracting quality passages for physics question-answering tasks.

### 3.3.2.2 PhyQA

The dataset comprises 9.5K Physics questions, with each chapter having nearly equal representation as illustrated in Figure 3.3. Each data point in the dataset is associated with 2 keys, i.e., "instruction" and "output," to organize the information. They are described as follows:

- **Instruction:** Key containing the question which needs to be answered by the model
- **Output:** Key containing the corresponding numerical answer along with a detailed explanation of how that answer was obtained.

An example of a question from the chapter Newton's Laws of Motion is given below:

---

**Instruction:** An aircraft of mass 176 kg executes a horizontal loop at a speed of 249 m/s with its wings banked at 80 degrees. What is the radius of the loop?

**Output:** Answer:  $35162.514\text{ m}$  <sep> Explanation: *To calculate the radius, we use the formula*

$$\frac{(v^2)}{(g * \tan(\frac{param*\pi}{180}))}$$

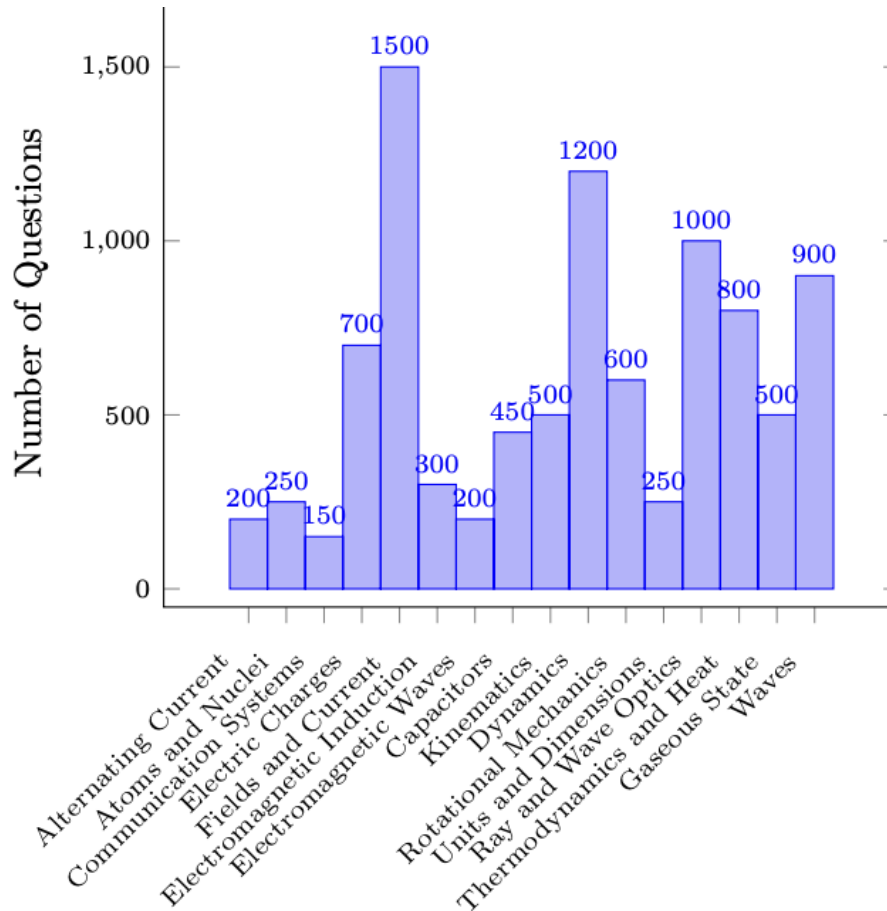


Figure 3.3: Insight into the number of questions at chapter-wise granularity in PhyQA

where  $v$  is the speed ( $249m/s$ ),  $g$  is the acceleration due to gravity ( $10m/s^2$ ), and  $param$  is the angle of banking  $80^\circ$ . Substituting the values,

$$R = \frac{249 \cdot 249}{10 \cdot \tan(\frac{80 \cdot \pi}{180})} = 35162.514m$$

### 3.3.2.3 RetriPhy Corpus

The RetriPhy Corpus comprises content extracted from NCERT books of Physics subjects for 11<sup>th</sup> and 12<sup>th</sup> grades. These NCERT (National Council of Educational Research and Training) books are known for their concise, accurate, and easily comprehensible presentation of concepts. The corpus contains theorems, equations, numerical problems, and explanations spanning various topics such as electric charges, gravitation, optics, atoms, nuclei, etc.

Containing material from all 14 chapters of the 11<sup>th</sup> and 12<sup>th</sup> grade NCERT Physics

books, the corpus comprises **28 documents**, each corresponding to a chapter. Each chapter contains around **30-32 paragraphs**, with approximately 400 tokens per paragraph. In total, the corpus extracts a total of **927 paragraphs**. We have also included an overlapping of 20 tokens in these paragraphs to ensure consistency.

#### 3.3.2.4 Data Collection and Augmentation

**PhyQA** : Data collection started by improving upon SCIMAT’s science problems [Kollepara \*et al.\* \(2021\)](#). Additional data was collected by scraping standard Indian High School open-source physics textbooks of classes 11 and 12.

Solvers have improved performance in answering math questions when the fine-tuning data set undergoes data enhancement transformations [Kumar \*et al.\* \(2022\)](#). We extend this to physics problem-solving by taking base problems from each sub-topic and applying these transformations to include a richer variety of questions. These transformations are two in nature:

- Substitution: Changing the values of constants in a question.
- Paraphrasing: Paraphrasing the problem  $q$  using a model to generate  $N$  candidate questions that differ from the one in which they were written.

**RetriPhy**: We have used NCERT textbook content to create the RetriPhy (Retrieval-based Physics) corpus. Our motivation for using NCERT textbook content is the concise and easy-to-understand explanation of concepts in these books. Also, PhyQA consists of problems related to the topics of grade 11<sup>th</sup> and 12<sup>th</sup> physics; hence, the focus is on the content from the 11<sup>th</sup> and 12<sup>th</sup> grade NCERT Physics textbooks.

These chapter-wise documents on physics are accessible on the official NCERT website. Our methodology involved using these documents to retrieve textual content from the chapters. To ensure an accurate representation of mathematical symbols and equations, we employed LaTeX annotations, thereby eliminating the potential for ambiguity in text interpretation.

#### 3.3.2.5 Data Annotation

**PhyQA**: Our team consisted of five dataset annotators, each having graduated high school and studied physics until class 12. Upon self-evaluation on a scale of 5, the

### 6.3 MOTION OF CENTRE OF MASS

Equipped with the definition of the centre of mass, we are now in a position to discuss its physical importance for a system of  $n$  particles. We may rewrite Eq.(6.4d) as

$$M\mathbf{R} = \sum m_i \mathbf{r}_i = m_1 \mathbf{r}_1 + m_2 \mathbf{r}_2 + \dots + m_n \mathbf{r}_n \quad (6.7)$$

Differentiating the two sides of the equation with respect to time we get

$$M \frac{d\mathbf{R}}{dt} = m_1 \frac{d\mathbf{r}_1}{dt} + m_2 \frac{d\mathbf{r}_2}{dt} + \dots + m_n \frac{d\mathbf{r}_n}{dt}$$

or

$$M \mathbf{V} = m_1 \mathbf{v}_1 + m_2 \mathbf{v}_2 + \dots + m_n \mathbf{v}_n \quad (6.8)$$

MOTION OF CENTRE OF MASS

Equipped with the definition of the centre of mass, we are now in a position to discuss its physical importance for a system of  $n$  particles. We may rewrite Eq.(6.4d) as

$$M \mathbf{R} = \sum m_i \mathbf{r}_i = m_1 \mathbf{r}_1 + m_2 \mathbf{r}_2 + \dots + m_n \mathbf{r}_n$$

Differentiating the two sides of the equation with respect to time we get

$$M \frac{d\mathbf{R}}{dt} = m_1 \frac{d\mathbf{r}_1}{dt} + m_2 \frac{d\mathbf{r}_2}{dt} + \dots + m_n \frac{d\mathbf{r}_n}{dt}$$

or

$$M \mathbf{V} = m_1 \mathbf{v}_1 + m_2 \mathbf{v}_2 + \dots + m_n \mathbf{v}_n \quad (6.8)$$

Figure 3.4: **Left:** A snippet from the Physics textbook of a section about "Motion of Centre of Mass"; **Right:** A snippet of the annotation file corresponding to the textbook snippet

annotators rated themselves as 3, 3, 4, 4, and 5. This self-evaluation was verified by administering a small test on basic questions to assess fundamental understanding.

These annotators used their in-depth understanding of physics concepts to annotate the solutions, providing relevant formulas and concepts. The annotated questions were then shuffled among the remaining annotators to evaluate the dataset's quality. During this process, the annotators who rated themselves as 3 could solve only 55% of the problems given to them, while the annotator rated 5 could solve 80% of the problems. This indicates that the dataset contains high-quality questions that can also be challenging for humans.

**RetriPhy:** The annotation process for the RetriPhy corpus focuses on the accurate representation of mathematical symbols and equations present in NCERT textbooks. For every mathematical notation or equation in the text, LaTeX was used for annotation. A  $\$$  symbol was added as the start and end delimiter to identify the beginning and end of LaTeX content in the text as shown in Figure 3.4. The annotation team consisted of three annotators, each contributing approximately one-third of the annotations. A shared segment of annotations was distributed to all annotators to validate the annotations, enabling cross-evaluation. This process revealed an impressive accuracy rate of 87% among the annotations.

Furthermore, the accuracy of annotations for all chapters was verified by both the annotators and an expert in the domain.

### 3.3.2.6 Inter-Annotator Agreement for Data Validation

In the data validation process, a team of five data annotators, all proficient college students in high school Physics, was employed. The dataset was equally divided among the annotators to ensure a balanced workload. Within this team, one annotator possessed expert-level knowledge, while the remaining four had intermediate expertise.

Rigorous attention to detail was maintained throughout the data annotation process to ensure the accuracy of the annotations. Each segment annotated by one annotator underwent a verification stage involving assessment by two other annotators. This multi-layered validation strategy was adopted to enhance the reliability of the annotated data.

The Fleiss' Kappa score of this annotation process was 0.65. By combining the expertise of the annotators, the cross-validation process, and the application of the Fleiss' Kappa coefficient, a robust framework for data validation was established, ensuring the accuracy and integrity of the annotated high school Physics dataset.

### 3.3.2.7 Fine-Tuning Using LoRA

**Model and Hyperparameters:** Vicuna [Chiang et al. \(2023\)](#) is an LLM prepared by fine-tuning a Llama base model on 70k user-shared conversations. We used a Vicuna model with 7 billion parameters as our baseline. NLP consists of pre-training language models on general text and fine-tuning model parameters on domain-specific data. However, as the model size increases, it is computationally expensive to fully fine-tune models, which involves retraining all parameters. We thus adopt the **Low Rank Adaptation (LoRA)** [Hu et al. \(2021\)](#), which proposes freezing model weights and injecting lower rank matrices into transformer layers that can be trained. This reduces training time substantially and reduces the hardware needed to keep model accuracy intact. We hypothesize that fine-tuning **Vicuna-LoRA** on our annotated **PhyQA** physics dataset will greatly improve the model's capabilities to answer physics questions.

**Experiment:** For this task, the PhyQA dataset is split into *8000 training samples* and *1500 test samples*. The training set is used to fine-tune the model. The Vicuna-LoRA model is run with different model weight representations, i.e. an **8-bit** representation and a **16-bit** representation. The LoRA rank  $r$  is set to **8** with a LoRA-dropout of

**0.05** for preparing the 8-bit fine-tuned model and is set to **16** for preparing the 16-bit fine-tuned model. The fine-tuning is run for **3 epochs** with a batch size of **128** and the learning rate equal to **3e-4** i.e. the Karpathy constant on an NVIDIA RTX A6000 GPU.

### 3.3.2.8 Rationale Behind Retrieval:

Our second experiment hypothesizes that providing relevant context to our language model about the question as input will greatly improve the explanation and precision of the answers. This is based on physics being driven by concepts and their interpretation rather than simply applying formulae. Given a query  $q$ , the retrieval-based system is prompted to find  $N$  relevant passages. Each passage is retrieved and appended with a  $\langle sep \rangle$  token. Query  $q$  is prepended to the  $N$  retrieved passages to form the final user query  $q_f$ , which can be described as:

$$q_f = q + \langle sep \rangle + \sum_{i=1}^n (N_i + \langle sep \rangle) \quad (3.1)$$

$q_f$  is then prompted to the fine-tuned Vicuna LoRA 7-billion model to get the answer.

### 3.3.2.9 Retrieval System Design:

The retrieval system combined with Vicuna-7b, LoRA fine-tuned on **PhyQA** on is called **SciPhy-RAG**. The RetriPhy Corpus is used to create passages of **400 tokens** each.

After creating the passages, the next step is to perform indexing, which addresses the challenge of memory storage. We perform the retrieval of relevant passages by using similarity matches between the indexed passages and the user query. However, as the corpus grows with more passages, this task becomes progressively more time-consuming. To efficiently store and index the passages, we adopt a method of representing them as dense vectors [Karpukhin \*et al.\* \(2020\)](#).

VectorStores, like Pinecone, are used for indexing and storing vector embeddings of text data for fast retrieval. It uses Approximate Nearest Neighbour (ANN) search in higher dimensions [Andoni \*et al.\* \(2018\)](#), which allows for handling large numbers of queries. ANN proposes, when given a set  $P$  of  $n$  points (in this case, a set of  $n$  queries),

a metric ball  $B_D(q|r)$  in metric space  $(X, D)$ . It creates a data structure  $S$  such that for any query  $q \in X$ ,  $S$  returns a point  $p$  that satisfies:

$$D(p, q) \leq r \tag{3.2}$$

$$\forall p \in B_D(q, cr) \cap P \tag{3.3}$$

It minimizes this for some  $c \geq 1$  and returns the point  $p$  at the minima. In our approach, each of the passages is converted into a 384-dimensional dense vector embedding using the **all-MiniLM-L6-v2 model** [Miao \*et al.\* \(2021\)](#). These passages are then stored in the VectorStore described above.

**Experimentation:** When prompted with a user query  $q$ , the system applies ANN search to identify top K relevant passages where K is specified by the user. The passages are returned by the system and appended to  $q$ . This final prompt with the passages  $q_f$  is prepared. We use the technique described in Section (B) to fine-tune Vicuna-7b using LoRA on PhyQA. The prompt  $q_f$  is inputted into the model specified above to obtain our final output.

### 3.3.2.10 Evaluation Metrics

We perform a two-tier evaluation of the fine-tuned models. We first choose the evaluation metrics such as BERT-Score [Zhang \*et al.\* \(2020b\)](#), METEOR [Banerjee and Lavie \(2005\)](#) and ROUGE-L, ROUGE-1, and ROUGE-2 [Lin \(2004a\)](#). These help us assess the quality and correctness of the explanations generated by the models. We then sample 100 questions from each chapter and prompt the model to give a "One-Word Answer". This gives the numeric answer, and we treat it as a classification task measuring the accuracies with the ground-truth answers of the test set. We call this metric as **Final Answer Accuracy (FAA)**. We repeat this with 10 randomly chosen samples and report the lowest accuracies achieved out of the 10.

### 3.3.2.11 Experimental Results and Analysis

Upon fine-tuning the baseline Vicuna-7B model and attaching the retrieval system, we get 2 models from the experimental setups described above. Table [3.3](#) shows their results on the evaluation metrics described above. Our fine-tuned SciPhy-RAG 16-bit

model (i.e. our retrieval pipeline + Vicuna-LoRA 16-bit) shows a **16.67%** increase on BERT-F1 scores over the base Vicuna-7b model. This shows that fine-tuning with **PhyQA** and using **RetriPhy** as our retrieval corpus increases the quality of explanations by a significant amount over the base model.

Table 3.3: Table showing the BERT, ROUGE and METEOR scores of the 2 Fine-Tuned Vicuna-LoRA models

Metric	Vicuna-7B	SciPhy-RAG (8-bit)	SciPhy-RAG (16-bit)
BERT (F1)	0.768	0.887	<b>0.899</b>
BERT (Precision)	0.744	<b>0.876</b>	0.865
BERT (Recall)	0.784	0.886	<b>0.895</b>
METEOR	0.285	0.347	<b>0.352</b>
ROUGE-L	0.321	0.371	<b>0.389</b>
ROUGE-1	0.315	0.358	<b>0.363</b>
ROUGE-2	0.147	0.181	<b>0.195</b>

On METEOR, our 8-bit SciPhy-RAG shows a **22.8%** increase. On the ROUGE evaluation metrics (ROUGE-L, ROUGE-1 and ROUGE-2), 16-bit SciPhy-RAG shows a much higher improvement (**19.4%**, **22.2%** and **35.3%** respectively) over the base Vicuna-7b model. We hypothesize that METEOR scores are calculated based on unigram matching between the reference and candidate sentences [Saadany and Orăsan \(2021\)](#), and the retrieval models drive the model output generation slightly away from the ground truth explanation. The same hypothesis holds for lesser improvements in ROUGE scores. However, the increase in BERT scores validates that the explanations are semantically similar and high quality.

Table [3.4](#) shows the Final Answer Accuracy (FAA) for the base Vicuna-7b model (i.e., before fine-tuning) and the SciPhy-RAG (16-bit) model after fine-tuning and applying retrieval. Note that the lowest scores have been reported for both models, demonstrating a substantial increase in accuracy across chapters, despite skewness in the training data. The lower accuracies observed are attributed to the smaller model sizes used due to resource constraints. However, our hypothesis suggests that similar improvements can be extended to larger models with parameter sizes > 50B and to newer architectures.

Table 3.4: Final Answer Accuracy before and after Fine-Tuning with PhyQA

Chapter	Before	After
Alternating Current	21.3	26.2
Atoms and Nuclei	20.8	27.1
Communication Systems	15.3	21.2
Electric Charges	22.9	26.5
Fields and Current	23.5	28.6
Electromagnetic Induction	22.2	29.1
Electromagnetic Waves	23.8	26.3
Capacitors	25.1	28.3
Dynamics & Rotational Mechanics	24.7	29.2
Units, Dimensions & Kinematics	22.4	27.3
Ray and Wave Optics	23.6	26.5
Thermodynamics and Heat	21.5	27.2
Gaseous State	19.6	24.3
Waves, Sound and Oscillations	20.6	28.4

### 3.3.3 MM-PhyQA: Multimodal Physics Question-Answering With Multi-Image CoT Prompting

#### 3.3.3.1 Dataset Description

There is a lack of multimodal datasets that comprise physics questions and are catered to high school students. While there are a few datasets available that consist of questions at a high school level, the quality of the questions does not belong to the highest standards. We curated a novel MM-PhyQA Dataset from publicly available resources [Anand et al. \(2024d\)](#). The resources are geared toward individuals who prepare for competitive exams throughout India, ensuring a higher difficulty level than that of an average high school physics question.

#### 3.3.3.2 Original Dataset Creation

Around 300 questions were manually created. As shown in Figure [3.5a](#), each question consists of a question, four options, the correct answer to the question, and an explanation that shows the reasoning by giving steps to approach the correct answer to select the correct answer.

### 3.3.3.3 Data Augmentation Procedure

For augmenting the data ChatGPT was given a prompt to create other variations of the text while ensuring that the meaning remained the same, bringing the total count of the questions in the dataset to 4500. Figure 3.5b shows the heatmap of the cosine similarity scores of the augmented questions w.r.t the original one for some of the questions. The questions were altered in two ways:

- **Numerical Value Variation:** During augmentation, numerical values in the original questions are adjusted to diversify the solutions, ensuring the model's impartiality. Python functions were developed for each question to get the correct answers after changing the values.
- **Structural Variation:** To avoid pattern memorization, the questions' structure was intentionally altered by rephrasing with ChatGPT and sometimes manual adjustments. Options were kept the same but randomly rearranged.

Initially, attempts to rephrase the entire query sometimes failed to properly shuffle the questions. Manual adjustments were made to correct these errors. While including the entire query didn't consistently result in a rephrased version, prompting ChatGPT to generate separate variations for the question and explanation improved results. However, some questions still required manual rephrasing, involving adjustments to the question, explanation, options, and correct answer.

### 3.3.3.4 Chain of Thought Variant

To facilitate the model to generate better reasoning, two questions were added corresponding to each question. These questions were based on the same topic and care was taken that similar concepts were utilized as seen in Figure 3.5a. All three questions consist of figures.

### 3.3.3.5 MM-PhyQA Dataset Topics

The dataset consists of topics that are present in high school physics curricula throughout India. The topics and the corresponding subtopics are listed in Table 3.5

### 3.3.3.6 Methodology

Figure 3.6 shows the pipeline that was utilized for data processing, input processing, and output generation. Each element in the dataset consists of the question ID, the

**Question:** A ball of mass ( $m$ ) 0.5 kg is attached to the end of a string having length ( $L$ ) 0.5m. The ball is rotated on a horizontal circular path about vertical axis. The maximum tension that string can bear is 324 N. The maximum possible value of angular velocity of ball (in rad/s) is:

- Options:**
- a) 9
  - b) 18
  - c) 27
  - d) 36

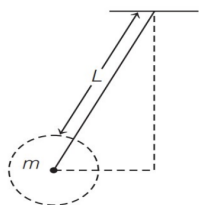


Figure 1: Image

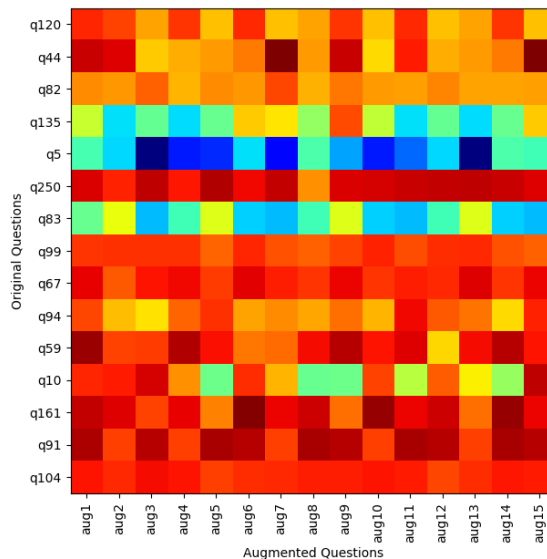
**Answer:** (d)  
**Explanation:** From the figure,  $T \sin \theta = mL \sin \theta \omega^2$ .

$$324 = 0.5 \times 0.5 \times \omega^2$$

$$\omega^2 = \frac{324}{0.5 \times 0.5}$$

$$\omega = \sqrt{\frac{324}{0.5 \times 0.5}} = \frac{18}{0.5} = 36 \text{ rad/s}$$

(a) Sample question of MMPhy-QA dataset



(b) Heatmap of text similarity between 15 randomly sampled original and augmented questions

Figure 3.5: MMPhy-QA Dataset questions

question, the label consisting of the corresponding answer and the reasoning, and the image filename. A function was used to convert each element to a prompt which can be fed to the model for generating the answer. For the Chain of Thought variant of the dataset, the structure was modified. As shown in Figure 3.7, the question was preceded by two similar questions with their correct answers and reasoning. All the three questions were separated by a delimiter consisting of hyphens. The filenames of the three images were stored in a comma-separated fashion.

### 3.3.3.7 Multi-image Chain-of-Thought (MI-CoT)

Different versions of LLaVA were utilized to evaluate the performance of CoT-Prompting. For the model to extract information from all the images corresponding to a list of questions, we came up with a novel approach, namely a Multi-Image chain of thoughts (MI-CoT). Under this technique, the three images were stacked on top of each other. The rationale for employing multi-image prompting was driven by the anticipation that the LLM would effectively distinguish and identify the specific image to be utilized for

Table 3.5: Topics and subtopics in the MM-PhyQA dataset

<b>Topic</b>	<b>Subtopics</b>
<b>Kinematics</b>	Velocity-Time, Acceleration, Rotational Motion, Gravitation, Motion in a Straight Line, Motion in a Plane, Periodic Motion, Wave Motion.
<b>Mechanics</b>	Law of Motion, Work, Power, Force, Law of Motion
<b>Electrostatics and Current Electricity</b>	Current, Voltage, Resistance, Electric Field, Ohm’s Law, Kirchhoff’s Laws, and Their Applications, Series and Parallel Combinations of Resistors
<b>Thermodynamics</b>	Laws of Thermodynamics, Thermal Equilibrium, Heat Transfer, Temperature, Reversible and Irreversible Processes, Kinetic Theory of Gases.
<b>Optics</b>	Reflection, Mirrors, Lenses, Wave Optics, Magnification.
<b>Magnetism</b>	Magnetic Field, Hysteresis, Permeability, Electromagnets.
<b>Electronic Devices</b>	Semiconductors, Logic Gates, Diode.
<b>Atoms</b>	Nuclei, Isotopes.

each question within a single prompt. Consider the images corresponding to the two prompt questions  $X_p$  and  $X_q$ , and the image for the main question  $X_r$ . LLaVA utilizes the CLIP visual encoder to get the visual feature  $Z_v$ :

$$Z_v = g(X_v) \quad (3.4)$$

where,

$$X_v = X_p \cdot X_q \cdot X_r \quad (3.5)$$

The filenames were passed as a list in the same order in which they were stacked. To make sure that the dimensions were correct for feeding the resultant concatenated image  $X_v$  into the CLIP encoder, the size of the images was reduced along one dimension using an autoencoder after basic pre-processing (normalization and padding) of the images. A basic neural-network-based autoencoder was employed and was trained on the train split for this purpose.

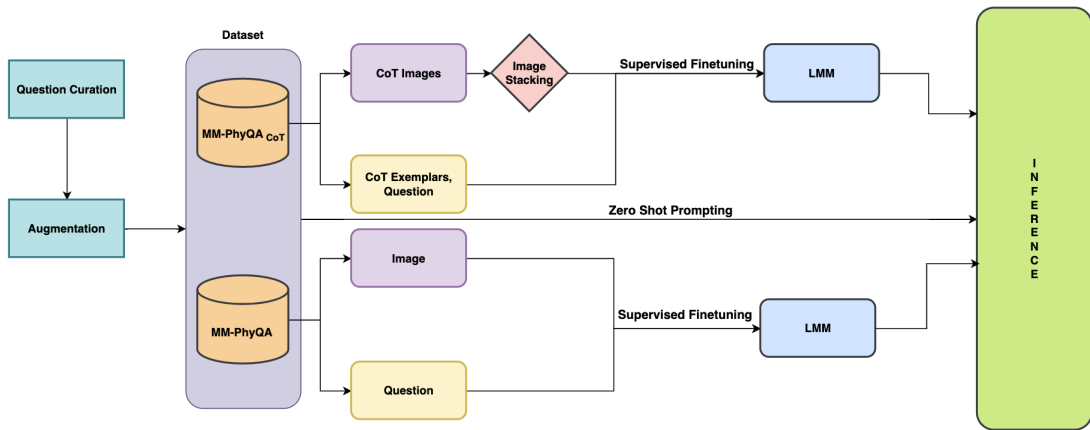


Figure 3.6: Schematic Pipeline of Multimodal Question Answering

question	label	image
<p>Q30_1- The velocity - time graph of a particle moving in a straight line is shown in figure. The mass of the particle is 2kg. Work done by all the forces acting on the particle in time interval between <math>t = 0</math> to <math>t = 10</math> s is. , select the correct option from - a)300J ; b)-300 J ; c) 400 J ; d)-400 J and give me the reason behind the selected option?  the correct option is a and the reason is From work-energy theorem, <math>W = \Delta KE = K_f - K_i = \frac{1}{2} m (v_f^2 - v_i^2) = \frac{1}{2} \times 2 [(-20)^2 - (10)^2] = 300 \text{ J}</math></p>	<p>b- Initial Velocity of particle, <math>v_i = 20 \text{ ms}^{-1}</math>  Final velocity of the particle, <math>v_f = 0</math>  According to work-energy theorem,  <math>W_{(net)} = \Delta KE = K_f - K_i</math>  <math>m(v_f^2 - v_i^2) = \frac{1}{2} \times 2 [0^2 - 20^2]</math>  <math>= -400 \text{ J}</math></p>	<p>q30_1.png,  q30_2.png,  q30.jpg</p>
<p>Q30_2- The <math>v</math>-<math>t</math> graph of a particle moving along the <math>x</math>-axis is shown in the figure. The mass of the particle is <math>4 \text{ kg}</math>. The work done by all the forces acting on the particle between <math>t = 3</math> s to <math>t = 6</math> s is. , select the correct option from - a)12 J ; b) 24J ; c) 8 J ; d) 32 J and give me the reason behind the selected option?  the correct option is b and the reason is By the equation of a line <math>y = mx + c</math>, we get, <math>v = -2t + 8</math> At <math>t = 3</math> s, <math>v = 2 \text{ m/s}</math>, <math>K_{E_i} = 8 \text{ J}</math> At <math>t = 6</math> s, <math>v = -4 \text{ m/s}</math>, <math>K_{E_f} = 32 \text{ J}</math> <math>W = K_{E_f} - K_{E_i} = 24 \text{ J}</math></p>		
<p>Q3- Velocity-time graph of a particle of mass 2 kg moving in a straight line is as shown in figure. Work done by all forces on the particle is. , select the correct option from - a)400 J ; b)-400 J ; c) -200 J ; d)200 J</p>		

Figure 3.7: Multi-Image Chain of thought (MI-CoT) Prompted text provided as input to LMMs during training. The main question to be answered is preceded by two exemplars, with the three questions separated by a delimiter. The image is a sequence of three comma-separated file names and the label is the ground truth

### 3.3.3.8 Models

We conducted a variety of experiments with both text and multimodal LLMs to gauge the difference in performance that comes about due to the change in the modality. LLaMA2-7b and Mistral-7b are the current SOTA open-source LLMs for textual input. These models were tested with text-only inputs. We use these LLMs to highlight the difference in the level of performance between fine-tuned models versus using them straight out of the box, aka through zero-shot prompting. For the ablation study, we also experimented with GPT-4, which is the current SOTA model for multimodal question-answering.

LLaVA and LLaVA-1.5, being multimodal, were provided with the figures along with the textual input. All the models were trained on A100 GPU and were fine-tuned for 5 epochs with a batch size of 8. Weighted Adam optimizer was utilised and the learning rate was set to  $2e-4$ .

Table 3.6: Performance of text-only and multimodal (MM) models. Model training specifications such as LoRA Rank and whether MI-CoT Prompting was used have been mentioned. All models were fine-tuned except for GPT-4, for which the answers were extracted using zero-shot prompting

Model	MI-CoT	Modality	Accuracy	Rouge1	Rouge2	RougeL	LoRA Rank
LLaMA2-7b	×	Text Only	0.25	0.380	0.187	0.315	8
Mistral-7b	×	Text Only	0.428	0.460	0.256	0.391	8
GPT-4	×	MM	0.331	-	-	-	-
LLaVA-13b	×	MM	0.293	0.551	0.383	0.501	64
LLaVA-1.5 7b	×	MM	0.533	<b>0.712</b>	0.579	<b>0.676</b>	64
LLaVA-1.5 13b	×	MM	0.527	0.672	0.532	0.634	64
LLaVA-1.5 13b	×	MM	0.531	0.621	0.490	0.586	128
LLaVA-13b	✓	MM	0.291	0.383	0.184	0.306	64
LLaVA-1.5 7b	✓	MM	0.354	0.496	0.343	0.444	64
LLaVA-1.5 13b	✓	MM	0.653	0.686	<b>0.585</b>	0.656	64
LLaVA-1.5 13b	✓	MM	<b>0.716</b>	0.677	0.582	0.650	128

We also experimented with different LoRA values in the case of the LLaVA-1.5 model. LoRA or Low-Rank Adaptation [Hu et al. \(2021\)](#), is a method to represent the weight changes during the training process in lower-ranked matrices. This is especially useful while fine-tuning general-purpose LLMs, as it speeds up the training process. A lower LoRA rank means fewer parameters are learned during the adaptation process, however, it results in a faster training process as well. We tested the 7b (7 billion) and 13b (13 billion) variants of LLaVA which correspond to the number of learning parameters. The different LLaVA configurations also formed the basis of our comparison of the performance of (MI-CoT) Prompting. For fine-tuning, open-source base model checkpoints from huggingface were utilized.

### 3.3.3.9 Results and Discussion

### 3.3.4 Model Performance

The results of the experiments with their accuracy scores on the test dataset are listed in Table [3.6](#). Mistral-7b and LLaMA2-7b, being text-only models, only take into account the textual data, which means that they are bound to miss critical information in some questions. We observed an accuracy score of 25.95% and 42.83% for LLaMA2-7b and Mistral-7b, respectively. Thus, we conclude that text-only LLMs are not capable of providing the right answers for a large number of multimodal questions which require multiple steps with complex reasoning to reach the final answer.

LLaVA is a model that can potentially answer complex questions due to its ability to

Table 3.7: Performance of text-only LLMs using zero-shot prompting and fine-tuning

Model	Task	Modality	Accuracy (in %)	Rouge 1	Rouge 2	Rouge L
LLaMA2-7b	Zero Shot Prompting	Text Only	14.22	0.301	0.096	0.201
	Supervised Fine-tuning	Text Only	25.95	0.380	0.187	0.315
Mistral-7b	Zero Shot Prompting	Text Only	23.32	0.259	0.083	0.180
	Supervised Fine-tuning	Text Only	<b>42.83</b>	<b>0.460</b>	<b>0.256</b>	<b>0.391</b>

process images. While the older LLaVA version with 13 billion parameters exhibited a lower accuracy than Mistral-7b, a text-only LLM, LLaVA-1.5 was able to perform significantly better than Mistral-7b. The best performance was seen when LLaVA, trained with 13 billion parameters, was fine-tuned with a LoRA rank of 128 and employed Chain of Thought Prompting with an accuracy score of 71.65%. A higher LoRA rank means that the model can learn more parameters during fine-tuning which makes it ideal for task-specific situations, such as answering complex physics questions. LLaVA-1.5 13b performs better than the 7b variant with an equal LoRA rank of 64 when multi-image prompting was utilized. This is because the larger number of trainable parameters allowed the model to learn and generalize better.

### 3.3.4.1 Zero Shot Prompting Vs Supervised Fine-Tuning

Table 3.7 shows the performance of LLaMA2-7b and Mistral-7b with zero-shot prompting and supervised fine-tuning. There is a marked improvement in the accuracy, Rouge1, Rouge2, and RougeL scores for both the models when fine-tuned on the dataset. This proves the assertion that current LLM models, in their out-of-the-box configurations, are not able to answer physics questions satisfactorily, and there is a need to fine-tune the models on domain-specific datasets to get better performance.

Zero-shot inferencing was done using the GPT-4 model. In most instances, GPT-4 failed to give correct answers and was not able to extract the entire information from the image. In some failure cases, GPT-4 needed more context than questions to make progress toward the solution.

### 3.3.4.2 Effect of Chain of Thought Prompting

For all variants of LLaVA that were tested, there was an increase in the accuracy score when MI-CoT Prompting was employed as seen in Figure 3.8a except in the case of LLaVA-1.5 7b model. A smaller number of trainable parameters meant that the model was not able to process the more complex multi-image input, leading to a sharp dip in

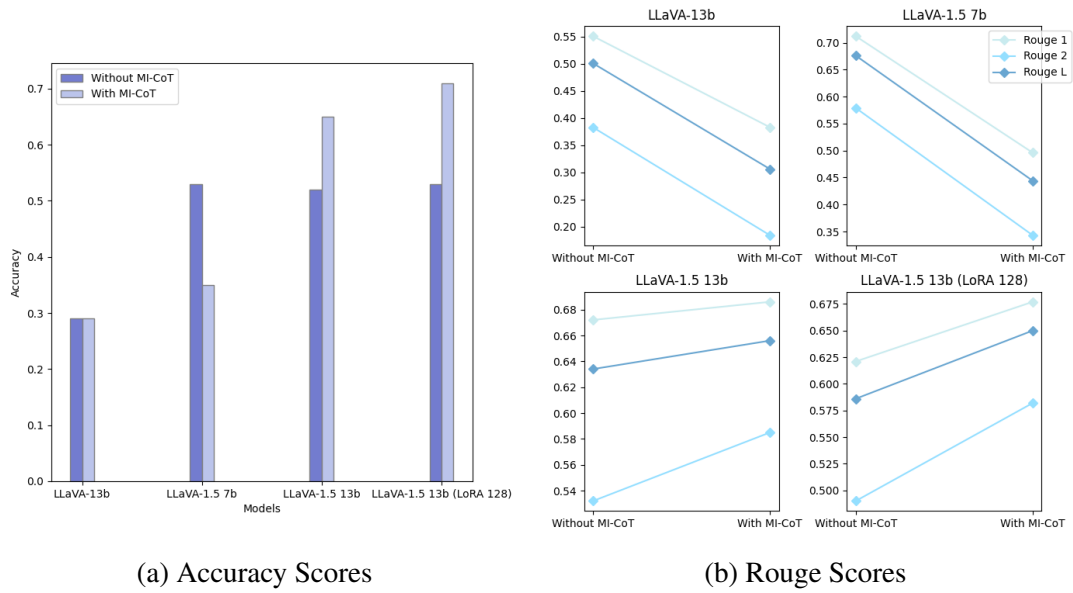


Figure 3.8: Comparison of the accuracy and rouge scores of different LLaVA variants when trained using (MI-CoT) Prompting vs their non-CoT prompted supervised fine-tuned (SFT) counterparts

the performance. The difference was the most significant in the case of LLaVA-1.5 13b trained with LoRA as 128, which also gave the best performance out of all the models tested when trained using MI-CoT Prompting. The MI-CoT Prompting trained version also exhibited high rouge scores as seen in Table 3.6. It can be observed from Figure 3.8b that the rouge scores were higher in the LLaVA-1.5 13b CoT variants, showcasing the fact that models that were able to leverage the MI-CoT prompt also showed a bump in the reasoning capabilities. A marked improvement in all metrics, when multiple images were provided in the prompt in the case of LLaVA-1.5 13b variants, provides evidence that the models were able to segregate and recognize the image that has to be used for each question present in a single prompt.

### 3.3.4.3 Error Analysis

Different types of errors were explored in Arora and Singh (2023). We investigated the error cases that were thrown by the best-performing model, LLaVA-1.5 13b. Figure 3.9 shows the different types of errors that were encountered. Their descriptions are as follows:

- **Conceptual Error:** The model is not able to identify the concepts that have to be involved correctly. For instance, in Figure 3.9a the model fails to identify that Kirchhoff’s loop rule has to be applied.
- **Grounding Error:** The model is able to identify the concept that has to be ap-

plied but does not formulate the correct equation. In the case of Figure 3.9b, the model is not able to apply the correct equation to get the centripetal acceleration of the cyclist.

- **Computational Error:** The model makes an algebraic mistake. In Figure 3.9c, the concept and the equations are correct, but the computation of the final answer is incorrect.

## 3.4 Conclusion

In our research, we have made significant strides in the application of AI to education, particularly in the domain of high school physics. Our work includes the creation of the largest dataset of high school physics questions based on NCERT exemplars, specifically designed for fine-tuning AI models. This dataset lays a strong foundation for the development of advanced AI tools, with current models focusing on text-based questions. However, recognizing the multimodal nature of physics problems, future work will extend these models to handle diagrams, graphs, and other visual components. Additionally, our dataset's varying representation of NCERT chapters offers an intriguing area for further investigation, particularly in how these variations impact AI model performance.

As we look ahead, several exciting avenues emerge for future exploration. Incorporating advanced prompting techniques like Chain of Thought and Tree of Thought into our datasets holds the potential to enhance explanation richness and accuracy. Moreover, expanding our benchmarks to other STEM fields, such as Chemistry and Biology, will drive the development of multimodal models and further advancements in using LLMs and AI to accelerate research and improve education quality across disciplines. Our research, particularly the PhyQA and RetriPhy systems, represents a significant step towards more intelligent, interactive, and personalized learning experiences, with the potential to influence AI-driven educational tasks like Automatic Evaluation.

Lastly, our introduction of the MM-PhyQA dataset has provided a valuable benchmark for LLM performance in educational contexts, revealing the challenges and potential of various models in handling complex reasoning tasks. Notably, fine-tuning models with techniques like MI-CoT prompting has shown promise, especially when combined with multimodal capabilities, as demonstrated by the LLaVA-1.5 model. Future research will focus on further enhancing model alignment through techniques like

### Prediction

correct option is (b) and reason is Considering the circuit, we find that  $R = 2 \times 10^3 \Omega$  and  $Q = 2 \times 10^6 \Omega$ . The potential difference across the cells is  $V = \frac{18}{0.01} = 180 \text{ V}$ .

### Reference

correct option is d and reason is From kirchhoff's loop rule,  $-i(2) + 18 - i(6.6) - 12 - i(1) = 0$   $i = \frac{6}{9.6} \text{ A}$  Terminal Voltage of 18V battery is  $V = \epsilon - ir_1 = 18 - i \times 2 = 18 - 0.625 \times 2 = 17 \text{ V}$  For 12V battery,  $V = \epsilon + ir_1 = 12 + 1 \times 0.625 = 16.626 \approx 17 \text{ V}$

### (a) Conceptual Error

#### Prediction

correct option is b and reason is Acceleration of the cyclist at point R = centripetal acceleration  $a_c = \frac{v^2}{r} = \frac{101^2}{1} = 101 \text{ ms}^{-2}$ , along RO

### Reference

correct option is a and reason is Acceleration of the cyclist at point R = centripetal acceleration  $a_c = \frac{v^2}{r} = \frac{10^2}{1000} = \frac{100}{1000} = 0.1 \text{ ms}^{-1}$ , along RO

### (b) Grounding Error

#### Prediction

correct option is c and reason is Capacitors in Series:  $2C_1 \times V_1 = C \times V_2$ ,  $2V_1 = V_2$  is the first equation and  $V_1 + V_2 + V_1 = 60$  is the second equation. After solving these two equations, we get  $V_2 = 20 \text{ V}$ .

### Reference

correct option is d and reason is Capacitors in Series  $2C_1 \times V_1 = C \times V_2$ ,  $2V_1 = V_2$  is equation first and  $V_1 + V_2 + V_1 = 60$  is equation second. After solving these two equations we get  $V_2 = 30 \text{ V}$

### (c) Computational Error

Figure 3.9: Types of errors encountered by LLaVA-1.5 13b

RLHF and extending MI-CoT prompting to additional multimodal tasks, pushing the boundaries of what AI can achieve in education. Together, these efforts pave the way for a new era of AI-driven education, where learning is more tailored, interactive, and effective.

# CHAPTER 4

## Online Learning

### 4.1 Motivation

Accurately assessing student engagement is vital in online learning, as it directly impacts educational outcomes and the overall learning experience. Traditional face-to-face classrooms rely on in-person cues to monitor engagement, yet remote learning environments lack equivalent methods for effectively gauging students' focus and emotional states. In this context, AI-driven engagement prediction models offer a promising solution by analyzing visual cues from video data to detect states like confusion, boredom, and frustration.

However, developing robust and scalable models for engagement prediction remains challenging due to the dynamic nature of emotions and the limited availability of datasets that represent real-world learning scenarios. This chapter addresses these gaps by advancing engagement detection methodologies, introducing novel models like NeuralGaze and GLAMOR-Net, and expanding datasets such as DAiSEE, EngageNet, and ECLIPSE. By leveraging innovative architectures and integrating multimodal features, this work seeks to improve engagement prediction accuracy, ultimately supporting more responsive and personalized educational interventions in online settings.

### 4.2 Introduction

For online as well as traditional education to be considered successful, student participation is important. However, the lack of non-verbal signals like body language and facial expressions in the virtual environment can make it more challenging to evaluate participation. Engagement detection in online learning extends beyond traditional face-expression classification by interpreting sustained, context-aware behavioral cues rather than isolated emotion labels. While expression classifiers map discrete facial movements to predefined emotions, engagement models must fuse micro-expressions,

gaze direction, head pose, posture shifts, and hand gestures often subtle and gradual into a cohesive measure of attention over time. Moreover, live deployment in virtual classrooms imposes stringent real-time constraints: models must maintain low latency (ideally under 100 ms per frame) and minimal computational footprint to run on students’ or instructors’ devices without interrupting the learning flow. Addressing these requirements necessitates lightweight architectures, parameter sharing, and early spatial downsampling to balance inference speed, resource usage, and predictive accuracy. Teachers can effectively identify and address disengagement by automating engagement assessments. Although video-based datasets such as the Belfast Database, AffectNet, EngageNet, and DAiSEE assist engagement prediction, they present computational difficulties for real-time applications and often contain imbalanced data sets with high levels of participation. Because it requires less computational power, image-based analysis delivers increased real-time applicability.

We present our dataset, ECLIPSE (Extended Classroom Learning Insights via Prolonged Student Engagement), that addresses these limitations. 250 participants’ photographs from a 55-minute online lecture from ECLIPSE, whereas 224 participants’ images were from a 50-minute lecture. Our datasets focus on low-engagement states that are underrepresented in existing datasets and have annotations for boredom, engagement, confusion, and frustration.

We use this dataset to investigate single-frame engagement recognition. Utilizing OpenFace face landmarks, eye gaze, head behaviour, and facial action units as indicators, we benchmark modern models that include EfficientNet, GLAMOR-Net, Residual Attention Networks (RAN), and ViTs. We offer NeuralGaze, which combines gaze vectors and facial landmarks with NCA to effectively represent images and classify engagement, and CG-ViT, a content-guided ViT model that includes contextual signals from video lectures. Our methods greatly outperform current methods by resolving class imbalance through weighted sampling and targeted loss.

#### Key Contributions:

- Presentation for the ECLIPSE dataset, which was created to identify disengaged students in online environments by tracking engagement trends over a longer amount of time with a focus on low-engagement settings.
- The ECLIPSE dataset is used to benchmark modern algorithms for single-frame affect recognition, such as EfficientNet, RAN, and GLAMOR-Net. In order to enhance predicted accuracy, dataset personalization is introduced.

- CG-ViT is a new architecture that significantly enhances engagement classification over baseline ViT models by including content guidance into the ViT framework and using contextual signals from video lectures.
- The creation of the NeuralGaze framework, which combines face action units, local and global picture context, and NCAs for image embedding generation, produced competitive results on the DAiSEE dataset with less computational overhead.

## 4.3 Experiments

### 4.3.1 Unveiling Learner Dynamics: The ECLIPSE Dataset and Neural Gaze Framework for Prolonged Engagement Assessment in Online Learning

#### 4.3.1.1 Dataset Description

#### 4.3.1.2 Data Collection

In this research [Anand et al. \(2024e\)](#), first-year undergraduate students aged between 17 and 18 were presented with online lectures on human-computer interaction, with each lecture having a duration of 55 minutes, 30 minutes, and 20 minutes. To effectively monitor fluctuations in student attention over these extended video lengths, images of the participants were captured at intervals of 45 seconds. This strategy was implemented as the analysis of individual images requires considerably lower computational power compared to full video analysis. As established in prior studies, significant alterations in facial expressions and emotional states generally manifest over time frames of one minute or more, implying that shorter sampling intervals might not provide sufficient context and could miss the temporal nuances of student affect [Dhall et al. \(2020\)](#); [Whitehill et al. \(2014\)](#). Therefore, the choice of 45-second intervals was considered optimal for this purpose. Furthermore, our dataset introduces a novel contribution by tracking student affective states for durations that exceed twenty minutes, directly addressing existing research indicating that attention levels drop substantially beyond this time frame [Bradbury \(2016\)](#). Notably, no current public dataset for engagement analysis encompasses images collected over such prolonged periods. This innovative methodology enables a richer and more detailed exploration of how student engagement patterns

shift during extended instructional sessions.

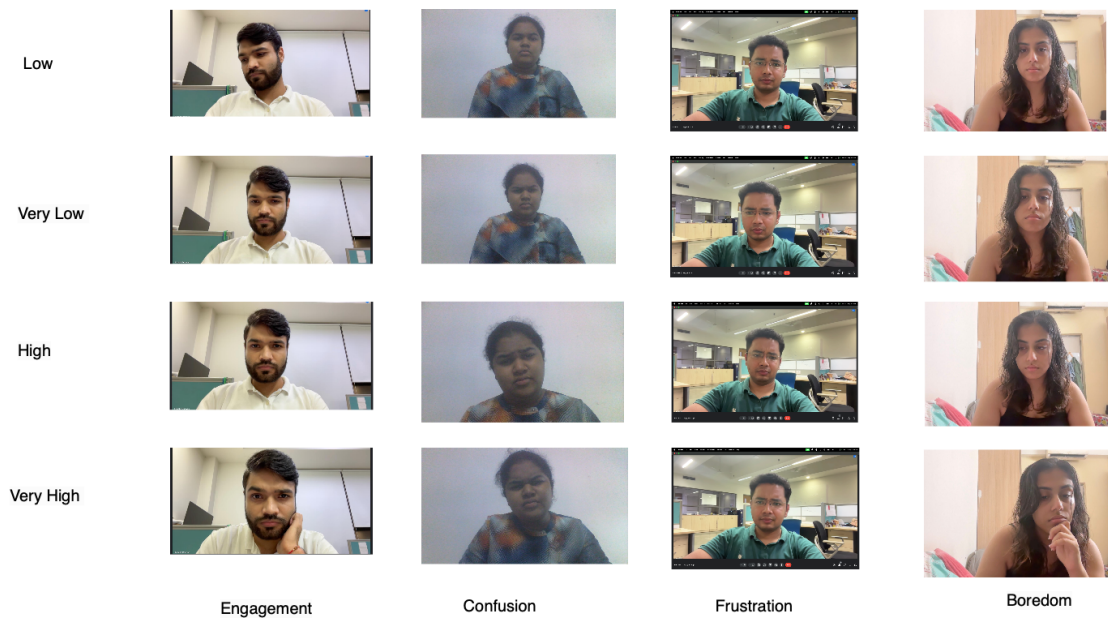


Figure 4.1: Sample of images in ECLIPSE for different affects at different levels

#### 4.3.1.3 Data Annotations

Our dataset labels four affective states relevant to user engagement: engagement, annoyance, bewilderment, and boredom, similar to the DAiSEE dataset. Each state is classified using a four-level scale: (1) extremely low, (2) low, (3) high, and (4) very high, following a method akin to that used by [Whitehill \*et al.\* \(2014\)](#) in their study on facial expressions. This labelling method intentionally omits a "neutral" state. Initial trials indicated that crowd annotators have a tendency to choose the "neutral" label when uncertain, which can undermine the creation of a robust and reliable dataset. The specified levels encourage annotators to provide a precise evaluation of the affective state, thereby enhancing the reliability of the dataset. The purpose of the four-level scale was to ensure that annotators make specific and accurate selections regarding affective states, thus improving the dataset's dependability. A group of three annotators worked collaboratively to assign engagement classifications to each frame of the participants. Given the subjective nature of affective states such as user frustration, interpretations of these emotions can vary based on different perspectives and biases. To address this issue, we employed a majority voting approach to determine the final label assignments. Additionally, a cross-labeling strategy was implemented to verify and correct any errors in the annotations. The annotators categorized participants in the images into classes

based on their levels of engagement, boredom, confusion, and frustration as shown in Figure 4.1.

Table 4.1: Weighted Cohen’s kappa for the three annotators

State	Labeler 1 vs 2	Labeler 1 vs 3	Labeler 2 vs 3
Boredom	0.839	0.735	0.749
Engagement	0.891	0.981	0.875
Confusion	0.855	0.719	0.702
Frustration	0.813	0.805	0.708

We used weighted Cohen’s Kappa with quadratic weights as the performance metric to assess the consistency among annotators. This metric measures the degree of agreement across multiple annotators who categorize the same data points, thus evaluating interrater reliability. A Kappa value between 0.70 and 0.80 is considered satisfactory. The Kappa coefficients for our dataset are provided in Table 4.1.

#### 4.3.1.4 Dataset Statistics

The dataset utilized in this study consists of 250 students, with video content being analyzed at intervals of every 45 seconds. For videos with a duration of 55 minutes, this results in 73 snapshots per student. In the case of 30-minute videos, each student has 40 snapshots, while for 20-minute videos, the number of snapshots per student decreases to 26. When focusing on specific groups, we observe that a cohort of 50 students, each with 55-minute videos, yields a total of 3,650 snapshots. Likewise, 90 students with 30-minute videos generate 3,600 snapshots, and another group comprising 110 students with 20-minute videos produces a total of 2,860 snapshots. The combined total from all these groups amounts to 10,110 images, offering a robust dataset for detailed analysis and research. The class distribution in the dataset is presented as follows: Highly Engaged (15.05%), Engaged (34.79%), Barely Engaged (21.53%), and Not Engaged (28.61%), as illustrated in Figure 4.2.

This dataset is distinct from those used in earlier studies, such as those by Gupta *et al.* (2016) and Singh *et al.* (2023), which experienced substantial class imbalance, complicating the accurate prediction of disengagement. In contrast, our dataset demonstrates a more balanced class distribution, with both "Not Engaged" and "Barely Engaged" classes being well-represented. The distribution of frames corresponding to each level of Engagement is shown in Figure 4.2.

Table 4.2: ECLIPSE Dataset: Affective State Label Composition

Affective State	Very low	Low	High	Very High
Engagement	28.61%	21.53%	34.79%	15.05%
Boredom	32.96%	16.46%	28.78%	21.78%
Confusion	62.40%	14.38%	17.73%	5.47%
Frustration	63.38%	18.68%	14.13%	3.8%

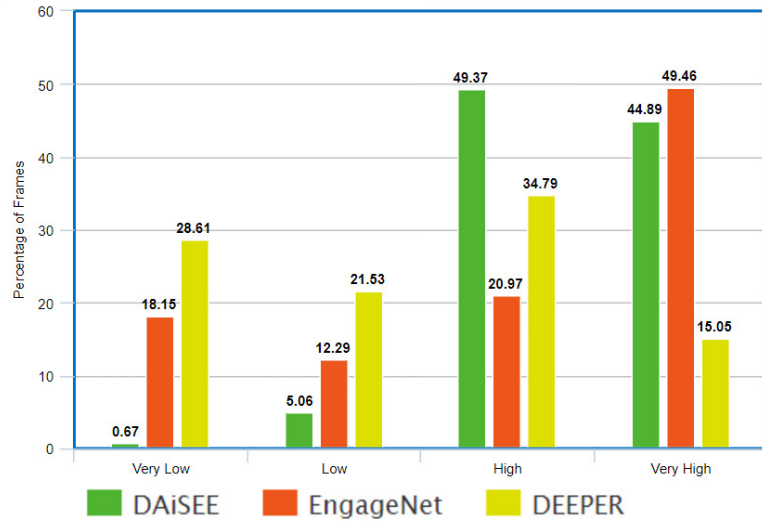


Figure 4.2: Engagement affect's class distribution for comparison between EngageNet, DAiSEE & DEEPER

#### 4.3.1.5 Research Ethics and Participant Protections

All participants provided written informed consent prior to data collection, and participation was entirely voluntary with the option to withdraw at any time. To ensure privacy, all data were anonymized using unique identifiers unlinked from any personal information.

The dataset is intended solely for non-commercial academic research. It is not publicly available and can only be accessed upon request, subject to approval and compliance with ethical use guidelines. Requesters must agree not to redistribute or use the data for commercial purposes, in line with institutional ethical standards.

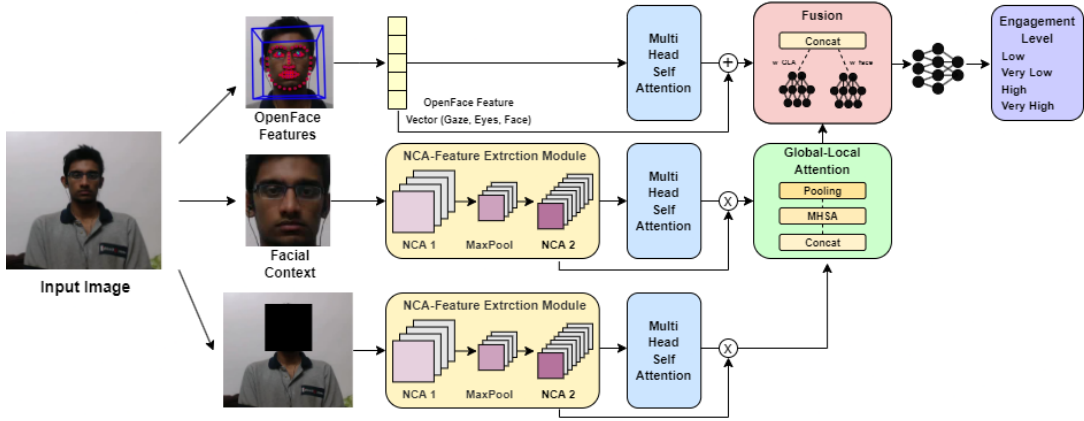


Figure 4.3: Model architecture with NCA-based feature embedding module combined with Multi-head self-attention and facial landmarks

#### 4.3.1.6 Proposed Methodology

Our objective is to merge localized facial features, such as expressions, key facial landmarks, and gaze directions, with broader image elements, including environmental context, body posture, and hand dynamics. While facial expressions and gaze orientation offer crucial emotional and attentional information, body posture and hand movements supply a wider contextual understanding of engagement. By combining these aspects, our approach enables an analysis that incorporates both nuanced facial expressions and broader behavioral indicators. To accomplish this, we design a model architecture and methodology, drawing inspiration from GLAMOR-Net [Le et al. \(2022\)](#), which we adapt specifically for engagement prediction tasks. Figure [4.3](#) provides a visual representation of our model’s architecture with the NCA-based feature extraction module for engagement classification.

#### 4.3.1.7 NCA-based Feature Extraction Module

We propose an NCA-based encoding module, following the approach described in [Mordvintsev et al. \(2020\)](#), to extract meaningful image embeddings. In this model, hidden channels for each pixel capture global information by propagating local information from the pixel’s vicinity across multiple time steps. This enables NCAs to effectively capture both local and global image information, making them a robust choice for creating image embeddings.

The complete encoding process employs two NCA models sequentially. Initially,

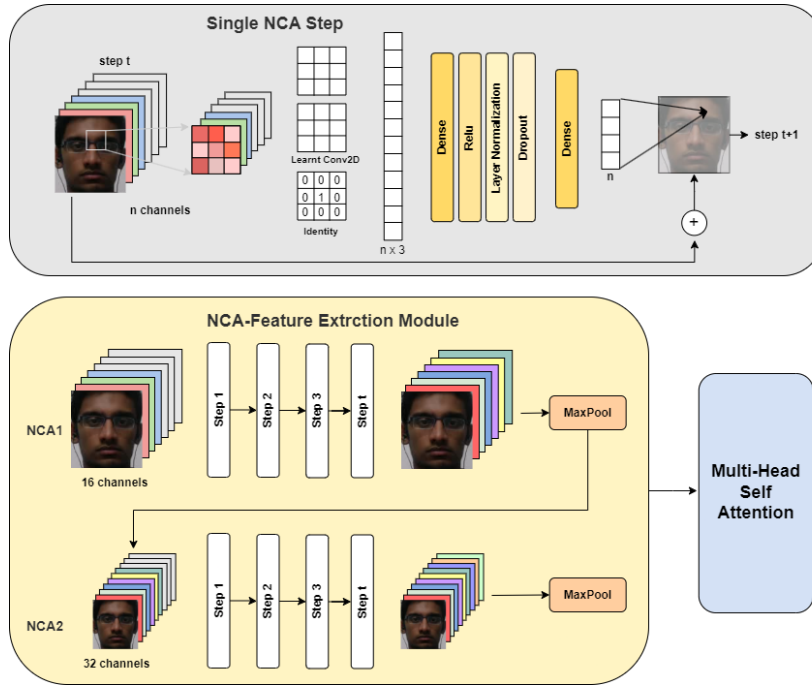


Figure 4.4: NCA-based feature extraction module consists of 2 NCA models. For NCA1, the original image is concatenated with 16 additional hidden channels. The output of NCA1 is downsampled using MaxPooling and given as input to NCA2 after the addition of 32 additional channels

the output from the first NCA model is processed through a MaxPool layer with a pool size and stride of (4,4). This downsampled embedding is then fed into a second, smaller NCA model, denoted as NCA2, which features an increased number of channels to enhance its capacity for capturing complex information. To facilitate this, empty channels are concatenated with the reduced embedding from NCA1 and forwarded to NCA2. The final embeddings generated by NCA2 are further downsampled using another MaxPool layer with the same pool size and stride of (4,4). For the Global content feature module, an additional MaxPool operation with a pool size of (2,2) and stride 1 is applied. This additional operation ensures that the dimensionality of the facial and global context features aligns properly for subsequent processing.

The NCA models learn feature embeddings by focusing on crucial pixel neighbourhood regions and propagating this local information across the image. MaxPooling operations are strategically employed to extract the most relevant features, while the second NCA model refines these features through additional processing cycles.

After this, the Multi-head self-attention (MHSA) mechanism is applied to the output from NCA2, enhancing the learned embeddings by emphasizing the most relevant features. The attention mechanism prioritizes key features based on their contextual

importance. The entire NCA-based attentive feature extraction pipeline is replicated separately for both the Facial and Global Context modules to derive the respective embeddings. For the facial context, the student's face is cropped, and for the global context, the facial region is masked, leaving the rest of the image intact. This design, inspired by the GLAMOR-Net framework, enables the context module to selectively gather valuable environmental data while avoiding redundant facial features. Refer to Figure 4.4 for a comprehensive illustration of the NCA feature extraction module and a detailed depiction of a single NCA step.

#### 4.3.1.8 Global-Local Attention module

The concept of utilizing a Global-Local Attention (GLA) module, inspired by GLAMOR-Net, has been adapted to merge global and local feature embeddings effectively. In our model, the outputs generated by the facial and context NCA-based feature extraction modules are first concatenated and then forwarded to an additional MHSA block. The resultant output from this MHSA block is subsequently integrated with the previously concatenated features, resulting in an enhanced attention-aware representation. This representation undergoes layer normalization followed by global average pooling to produce the final vector embedding, which encapsulates both global and local feature information. It is essential to note that the implementation of the GLA module in our approach diverges significantly from the one employed in GLAMOR-Net.

#### 4.3.1.9 Head Pose and Facial Features

The extracted head pose and facial features, including facial landmarks and gaze vectors, are obtained using OpenFace [Baltrusaitis et al. \(2018\)](#). These features are subsequently combined with the GLA feature vector within the fusion module, preparing them for the final classification. A multi-head self-attention mechanism is then applied to the combined features to refine the representation. The resulting output from this step is integrated with the original feature vector, enhancing the overall representation for further processing.

Inspired by the findings in [Singh et al. \(2023\)](#), where experiments were performed on GLAMOR-Net both with and without OpenFace features, it was observed that incorporating these facial features led to improved performance. Consequently, we integrate

the OpenFace facial features in the final architecture proposed in this work.

#### 4.3.1.10 Fusion Module and Classification Network

The final features for downstream classification are obtained by integrating the outputs from the Global-Local Attention (GLA) module, which fuses the attention-aware combined and condensed embeddings of both Facial and Contextual data, alongside the enhanced OpenFace Facial Features. Distinct neural networks are employed to generate individual scores for the outputs of the GLA and OpenFace modules. These scores undergo normalization via the softmax function to derive their respective weights. The resulting weighted vectors are subsequently concatenated and fed into a Fully Connected Neural Network for the concluding classification task.

#### 4.3.1.11 Dataset Splits

We propose a new data-splitting technique that has demonstrated improved accuracy and reduced loss.

**Generalization:** To validate the adaptability of our model across various classroom scenarios, the dataset was split into distinct training, validation, and testing sets, each encompassing different groups of students. This strategy supports the development of robust models by precisely evaluating their performance. The testing set provides an unbiased measure of the model’s ability to perform with entirely new student data. The participants were divided among training, validation, and testing sets using an 80:10:10 ratio.

**Personalization:** We achieve subject-specific personalization by segmenting each participant’s 55-minute, 30-minute, and 20-minute video image sequences into dedicated training, validation, and testing sets. Unlike the generalization method, each of these sets contains specific frames for every student. The model is trained to recognize the unique patterns and temporal variations in each student’s expressions of emotional state. This personalization technique lays the groundwork for future enhancement of models using minimal additional classroom data, allowing quick adaptation to new students and evolving interaction patterns. The images for each participant were divided into training, testing, and validation sets using the same 80:10:10 ratio.

We utilized the complete EngageNet dataset, as outlined in Table [4.3](#), to examine

and contrast the effectiveness of various models in predicting engagement levels across multiple and binary class categories. The results presented pertain to the validation split of the initial dataset. To evaluate the impact of dataset generalization and personalization on the EngageNet dataset, a subset of 3000 images was selected. This subset was curated to ensure equal representation of each interaction level, as detailed in Table 4.6.

#### 4.3.1.12 Baseline Models

We compare the benchmark results obtained using EfficientNet Tan and Le (2019), ViTs Dosovitskiy *et al.* (2020a), Residual Attention Networks Wang *et al.* (2017), and GLAMOR-Net Le *et al.* (2022) with the original baselines provided for the DAiSEE and EngageNet datasets. Furthermore, we evaluate the performance of our proposed framework, NeuralGaze, which employs a Content-guided ViTand NCA-based Self-Attention mechanism to effectively capture both global and local features.

**EfficientNet:** EfficientNet Tan and Le (2019) is a compact convolutional neural network that achieves scalability and accuracy through a compound scaling approach. This method balances the depth, width, and resolution of the network for improved efficiency. In this study, we fine-tune the final classification layer of the EfficientNet-B0 model, which is pre-trained on the ImageNet dataset Deng *et al.* (2009), adapting it for the specific engagement classification task.

**Residual Attention Networks:** Residual Attention Networks Wang *et al.* (2017) combine residual connections, which help optimize gradient flow, with attention modules to improve both model optimization and feature learning in visual tasks. These networks are particularly effective in emphasizing crucial facial features such as eyes and mouth, which are relevant in engagement assessment tasks. We train the RAN model end-to-end on our datasets to maximize its learning capacity.

**Global-Local Attention (GLAMOR-Net):** GLAMOR-Net Le *et al.* (2022) employs separate CNN branches to extract features from facial expressions and the surrounding context, integrating them through a global-local attention mechanism. This mechanism prioritizes significant aspects of the extracted features, which are then fused to predict emotional states. In our version of GLAMOR-Net, we extend the model by including OpenFace-based features such as Head Pose, eye gaze, and Facial Action Units (FA). Additionally, we incorporate Focal Loss (FL) into the training regime. Ab-

lation studies presented in Table 4.3 demonstrate the effectiveness of these modifications in enriching contextual information and enhancing overall performance.

**Vision Transformer (ViT):** The Vision Transformer (ViT) model (Dosovitskiy *et al.* (2020a)) encodes images as sequences of patches, transforming them into embeddings that capture spatial and appearance information. A transformer encoder then applies self-attention to analyze these embeddings for both local and global contextual relationships, followed by a multi-layer perceptron (MLP) head for classification. We fine-tune the final classification layer of the ViT model using the pre-trained weights from ImageNet-21k (Deng *et al.* (2009)) to predict engagement levels.

#### **Content-guided ViT model (CG-ViT):**

We extend the ViT model to integrate the content that the user has viewed over the last 45 seconds. This is achieved in the following manner: (1) For an image  $x_i \in I$ , where  $I$  is the set of images, we extract the feature vector  $f_i^1$  from the final layer of the frozen ViT Encoder ( $E_1$ ). (2) For a video segment  $v_i \in V$ , where  $V$  is the set of 45-second video segments, we extract the feature vector  $f_i^2$  from the final layer of the frozen TimeSformer (Bertasius *et al.* (2021)) Encoder ( $E_2$ ). (3) We concatenate these two feature vectors and feed them into an MLP head for engagement classification.

$$f_i^1 = E_1^{freeze}(x_i); \quad f_i^2 = E_2^{freeze}(v_i)$$

$$f_i = \{f_i^1 || f_i^2\}; \quad o_i = MLP(f_i)$$

For this model, we first pre-train the ViT model on the DAiSEE dataset and then freeze the pre-trained layers of both the ViT and TimeSformer models, training only the classification layer on the ECLIPSE dataset. Table 4.7 presents the results, highlighting the substantial improvement achieved by incorporating content information into the pre-trained ViT model.

#### **4.3.1.13 Implementation Details**

The models are trained for a maximum of 60 epochs with a batch size of 8. During training, the best model based on validation performance is saved, and its corresponding results on the test dataset are recorded. The Adam optimizer employs a learning rate scheduler, which reduces the learning rate at specified intervals. The initial learn-

ing rate is set to  $5 \times 10^{-4}$ , with a decay factor of 0.1 applied at each interval. For the NCA feature extraction module, specific hyperparameters are chosen: the NCA1 layer has a hidden feature size of 16 channels, and an additional 32 channels are assigned to NCA2. A fire rate of 0.5 is used, and the dense layer is set with a hidden dimension of 128. At initialization, the additional hidden channels in the NCA model are set to zero. The detailed architecture of this model is illustrated in Figure 4.4. Each Multi-head Attention module in the model is configured with eight heads. All dense neural networks are implemented as single-layer Neural Networks with hidden layer dimensions of 128. The Focal Loss function is utilized in training the model. Classification accuracy is chosen as the evaluation metric due to its prevalent use in the relevant domain. In the four-class classification task, the original labels of "very low," "low," "high," and "very high" are maintained. However, for the two-class classification, the "very low" and "low" categories are merged into a single "low" category, while "high" and "very high" are combined into a single "high" category.

#### 4.3.1.14 Results

Table 4.3 presents a comparison of multi-class and binary classification results across three datasets: DAiSEE, EngageNet, and ECLIPSE, focusing on varying levels of the engagement affective state. As shown in Table 4.3, GLAMOR-Net emerges as the best-performing model across all datasets, particularly when it incorporates OpenFace facial features and focal loss to handle class imbalance. The integration of focal loss with GLAMOR-Net leads to a notable improvement over competing models, as depicted in Table 4.3. Specifically, for the EngageNet dataset, the combination of GLAMOR-Net with focal loss and the inclusion of facial features achieves an accuracy of 0.750, exceeding the benchmark of 0.676 set by Singh *et al.* (2023). Unlike Singh *et al.* (2023)'s transformer-based video analysis approach, our method employs single-frame analysis, which not only lowers computational requirements but also enhances classification efficiency.

Table 4.3: Engagement level Classification Results for DAiSEE, EngageNet, and our dataset ECLIPSE

Method	EngageNet	DAiSEE	ECLIPSE
EfficientNet-B0	0.606	0.526	0.394
ViT	0.604	0.456	0.382
RAN	0.604	0.541	0.387
GLAMOR-Net (4 class)	0.586	0.568	0.412
GLAMOR-Net + FL (4 class)	<b>0.750</b>	0.569	<b>0.750</b>
GLAMOR-Net + FA + FL (4 class)	<b>0.750</b>	<b>0.572</b>	0.389
NeuralGaze (4 class)	0.604	0.551	0.366
Transformer (benchmark for EngageNet)	0.6761	-	-
GLAMOR-Net (2 class)	0.871	0.950	0.532
GLAMOR-Net + FL (2 class)	0.878	0.950	0.500
GLAMOR-Net + FA + FL (2 class)	0.882	0.900	0.561
NeuralGaze (2 class)	0.871	0.953	0.471

Table 4.4 elaborates on the performance for classifying confusion, boredom, and frustration. Here, GLAMOR-Net, combined with OpenFace features and focal loss, achieves the highest performance, with our NeuralGaze model coming next for the DAiSEE dataset. Interestingly, in the binary classification tasks, particularly for Boredom and Frustration, NeuralGaze outperforms GLAMOR-Net. These specific categories demonstrate a more balanced distribution of data points across different engagement levels, highlighting the effectiveness of NeuralGaze in contexts with a well-balanced dataset.

Table 4.4: Confusion, Frustration, and Boredom results for DAiSEE and ECLIPSE dataset for various models

Method	Dataset	Confused	Bored	Frustrated
EfficientNet-B0	DAiSEE	0.671	0.462	0.777
	ECLIPSE	0.578	0.382	0.571
ViT	DAiSEE	0.672	0.481	0.777
	ECLIPSE	0.576	0.227	0.626
RAN	DAiSEE	0.672	0.468	0.777
	ECLIPSE	0.561	0.392	0.485
GLAMOR-Net	DAiSEE	0.682	0.484	0.782
	ECLIPSE	<b>0.586</b>	0.365	0.602
GLAMOR-Net + FL (4 class)	DAiSEE	0.688	<b>0.492</b>	0.789
	ECLIPSE	0.572	0.360	<b>0.657</b>
GLAMOR-Net + FA + FL (4 class)	DAiSEE	<b>0.691</b>	0.491	<b>0.793</b>
	ECLIPSE	0.584	<b>0.413</b>	0.636
NeuralGaze (4 class)	DAiSEE	0.688	0.479	0.790
	ECLIPSE	0.582	0.300	0.629
GLAMOR-Net (2 class)	DAiSEE	0.910	0.740	0.742
	ECLIPSE	0.719	0.564	0.784
GLAMOR-Net + FL (2 class)	DAiSEE	0.920	0.730	0.751
	ECLIPSE	<b>0.726</b>	0.596	0.783
GLAMOR-Net + FA + FL (2 class)	DAiSEE	<b>0.940</b>	0.740	0.755
	ECLIPSE	0.725	<b>0.654</b>	<b>0.786</b>
NeuralGaze (2 class)	DAiSEE	0.916	<b>0.790</b>	<b>0.967</b>
	ECLIPSE	0.718	0.420	0.784

On the ECLIPSE dataset, the results consistently point to GLAMOR-Net, integrated with OpenFace features and focal loss, as the top-performing model for classifying affective states such as confusion, boredom, and frustration, both in four-class and two-class classification scenarios. This consistency suggests the efficacy of OpenFace features like eye gaze, head pose, and facial action units in discerning affective states. Additionally, the use of weighted loss functions like focal loss proves beneficial by alleviating the undue influence of majority classes, leading to higher classification accuracy.

Our evaluation of video-based models on the DAiSEE dataset, as outlined in Table 4.5 emphasizes the opportunity to derive critical insights from video data. Nonetheless, computational efficiency and associated costs are pivotal challenges that need to be addressed through methodological improvements.

Configuration	Frames	Accuracy
EmotionNet (DAiSEE)	1	0.5107
NeuralGaze (Ours)	1	<b>0.554</b>
GLAMOR-Net + FA + FL	1	<b>0.572</b>
DFSTN	20	0.5884
Marlin + FA + BodyPose Features	30	0.59
ResNet + TCN	50	0.639
LRCN (DAiSEE)	250	0.579
BERN	300	0.60

Table 4.5: Engagement level prediction comparison with single-frame and video-based benchmarks. Our model gives competitive performance to video-based models and surpasses existing single-frame benchmarks. We set new benchmarks with GLAMOR-Net+FA+FL and NeuralGaze

The outcomes summarized in Table 4.5 benchmark our model’s performance against existing single-frame and video-based models. Our proposed NeuralGaze architecture and modifications to GLAMOR-Net that integrate Facial Action Units and Head Pose outperform the single-frame classification benchmarks and exhibit competitive results with video-based benchmarks that utilize multiple frames per video. An observed trend of increased accuracy with more frames processed per video points to the potential for further exploring temporal-domain classification techniques, especially with the inte-

gration of Local-Global Attention embeddings and resource-efficient video data analysis strategies in future research.

The Content-Guided Vision Transformer (CG-ViT) model investigates how the content being viewed correlates with participant engagement levels. As detailed in Table 4.7, the CG-ViT model shows a marked improvement in performance when content information is incorporated alongside participant expressions. These findings underscore the significance of instructional content in shaping student behaviour and highlight the opportunity for educators to strategically design lecture content that fosters an engaged classroom environment.

Table 4.6: Engagement level Classification Results for EngageNet and EngageNet & ExCEDA combined respectively

<b>Models</b>	<b>EngageNet</b>		<b>ExCEDA+EngageNet</b>	
	Gen.	Pers.	Gen.	Pers.
GLAMOR-Net (4 class)	54.94	64.23	53.29	56.32
GLAMOR-Net (2 class)	78.83	86.25	74.34	76.75
GLAMOR-Net+ OpenFace (4 class)	56.64	<b>68.72</b>	55.68	56.81
GLAMOR-Net+ OpenFace (2 class)	84.15	<b>92.12</b>	77.46	77.72
Transformer (Baseline)	67.61	-	-	-

Table 4.7: Results for binary classification of Content-guided ViT model on the ECLIPSE dataset vs ViT model without video features

Classes	F1-score		Improvement
	with content	without content	
Confusion	56.22%	48.81%	15.18%
Boredom	79.00%	72.23%	9.37%
Frustration	63.14%	55.42%	13.93%
Engagement	82.53%	64.83%	21.45%

## 4.4 Conclusion & Future Scope

This chapter set out to make engagement detection both more realistic and more affordable for online classrooms. By releasing the **ECLIPSE** image-centric dataset and its ExCEDA extension, we captured minute-by-minute shifts in attention across an entire lecture—data that existing video corpora simply do not provide . Building on these resources, we introduced two complementary single-frame approaches. **Neural-Gaze** couples facial landmarks, gaze vectors, and local–global context to achieve strong recognition with modest compute, while **CG-ViT** fuses what a learner is seeing with their facial cues to enrich low-engagement detection . Together, the dataset and models advance the goal of responsive, AI-driven support for distance learning.

Our experiments demonstrate that the image-first strategy can rival much heavier video pipelines. Careful fusion of facial detail with contextual signals consistently produced clearer distinctions among boredom, confusion, frustration, and on-task states. Just as important, the lightweight designs preserve real-time speed, making deployment on everyday hardware feasible. These outcomes confirm that accurate engagement analysis no longer requires processing hundreds of frames; a single, well-chosen snapshot, enriched with minimal context, can suffice.

Looking ahead, the thesis identifies two immediate priorities. First, personalising engagement models to each learner and adapting to user-specific expression patterns promises to raise accuracy and make feedback feel natural. Second, triggering real-time interventions based on detected engagement levels so that virtual and hybrid classrooms

can respond the moment attention begins to wane . Pursuing these avenues will further close the gap between human instructors and automated tutoring systems, reinforcing the broader thesis vision of domain-aware, context-sensitive AI for education.

# CHAPTER 5

## Mathematical Question Answering

### 5.1 Motivation

AI has transformed many educational domains, yet AI struggles with complex mathematical and geometric problems that require both visual and textual reasoning. Geometry, in particular, poses unique challenges in secondary education, where students must interpret images and textual information to solve problems—tasks that demand a deep understanding of geometric and mathematical principles. Existing datasets, such as GEOS and Geometry3K, advance this area but are often constrained by limited scalability and question diversity, particularly in multilingual contexts.

Most educational AI research has focused on English, leaving a gap for students in non-English-speaking regions, especially those who learn in Hindi. Millions of students lack robust, localized AI tools that support their learning needs, creating barriers in understanding complex subjects. To address these gaps, this work introduces datasets like GeoVQA, GPSM4K, MathQuest, and IndiMathQA, which enhance AI’s ability to solve bilingual and multimodal problems. By building these resources, this research aims to improve AI’s accessibility and effectiveness in education, offering support that aligns with diverse linguistic and cultural educational needs.

### 5.2 Introduction

Geometry problems, characterized by their intricate relationship between textual descriptions and visual representations, serve as a fundamental benchmark for evaluating multimodal numerical reasoning capabilities in AI. The inherent complexity and diversity of the information entailed in these problems pose significant challenges, necessitating the development of sophisticated datasets capable of addressing these complexities. Historically, datasets such as GEOS and Geometry3K have played pivotal roles in advancing the field by expanding the range of geometric shapes and variable operators, although their scalability was limited due to heavy reliance on manual annotation.

The advent of more sophisticated datasets like GeoQA has marked a significant step forward by including clear annotations of the problem-solving process, thereby enhancing the universality and interpretability of multimodal numerical reasoning. Despite these advancements, existing geometry datasets fall short in several critical areas, particularly in catering to secondary-level education. Our efforts to bridge these gaps led to the development of comprehensive multimodal geometry datasets: GeoVQA and GPSM4K. Each dataset contains approximately 4,000 geometric image-caption pairs and question-answer pairs, significantly enhancing the diversity and complexity of problems. These datasets introduce dual solution versions for each problem and incorporate underrepresented problem types like Numerical Answer Questions and Theorem Proving Questions, broadening their applicability in educational contexts and improving the capability of AI models to solve a broader spectrum of geometric problems.

On another front, mathematical problem-solving represents a multifaceted cognitive skill that spans the comprehension of problem statements, identification of pertinent concepts and formulas, and application of suitable strategies and algorithms. The rapid advancements in AI and NLP have seen LLMs emerge as formidable tools capable of generating natural language text across various domains. LLMs, grounded in transformer architecture, demonstrate impressive proficiency in mathematical reasoning by leveraging their inherent understanding of arithmetic operations and symbolic manipulation. However, they still grapple with substantial hurdles in tackling math word problems, particularly those necessitating intricate reasoning or domain-specific knowledge.

In response to these challenges, we introduce the MathQuest dataset, curated from 11th and 12th standard Mathematics NCERT textbooks, to equip LLMs with the ability to solve complex mathematical problems. This dataset encompasses various levels of mathematical complexity and serves as a robust foundation for fine-tuning SOTA LLMs. Through the introduction of datasets like GeoVQA, GPSM4K, and MathQuest, along with innovative approaches in dataset structure and model fine-tuning, we aim to push the boundaries of geometric and general mathematical problem-solving in AI, making significant strides towards developing models that can reason mathematically with a deeper understanding of both visual and textual data.

Recent advancements in open-source and closed-source LLMs have shown promise in improving mathematical problem-solving abilities. However, evaluations on datasets

like GSM8K and MATH reveal stark differences in these models’ capabilities, especially in handling complex, high-school-level problems that demand deeper contextual understanding and advanced reasoning. Furthermore, a notable gap persists in the performance of LLMs when applied to English versus non-English languages, particularly Hindi. Given that over 105 million students in India predominantly use Hindi in their educational contexts, enhancing LLMs to address this gap is crucial. Current research efforts in Hindi LLMs have not fully optimized these models for domain-specific tasks like mathematical problem-solving.

To address these challenges, we introduce several key contributions focused on enhancing small open-source LLMs (SLLMs) for complex mathematical tasks in Hindi. Our research emphasizes a novel decomposition strategy, structured solution approaches integrated with curriculum learning, bilingual combined training, and the development of a comprehensive IndiMathQA dataset, containing 7,823 questions across 14 mathematical domains. Comprehensive performance analyses on multilingual LLMs underscore the limitations and potential improvements achievable within constrained computational resources. By releasing datasets and methodologies publicly, we aim to promote transparency and encourage further research.

Through these advancements, our research continues to enrich the multimodal understanding and application of AI in educational settings, providing clearer insights into the problem-solving process and offering more personalized learning experiences facilitated by technological innovations.

Key contributions of this work are as follows:

- Introduced two multimodal geometry datasets, GeoVQA and GPSM4K, each containing around 4,000 image-caption and question-answer pairs. These datasets enhance complexity and diversity by incorporating dual solution versions, Numerical Answer Questions, and Theorem Proving Questions, addressing gaps in secondary-level geometry education.
- Created the MathQuest dataset from 11th and 12th standard NCERT textbooks to support LLMs in handling complex high-school-level mathematical problems, laying a robust foundation for fine-tuning models on advanced mathematical reasoning.
- Developed the IndiMathQA dataset with 7,823 questions across 14 domains to bridge the performance gap in mathematical problem-solving for Hindi-speaking students, supporting LLMs optimized for educational contexts in non-English languages.

- Conducted an in-depth analysis on multilingual LLMs, evaluating open-source and closed-source models on datasets like GSM8K and MATH, and highlighting the advancements possible with optimized datasets and training methodologies within computationally constrained environments.

## 5.3 Experiments

### 5.3.1 GeoVQA: A Comprehensive Multimodal Geometry Dataset for Secondary Education

The study introduced "GeoVQA" [Anand \*et al.\* \(2024a\)](#) an extensive geometry dataset curated from Grades 6 to 12, based on textbooks aligned with both State Board and NCERT standards. In contrast to traditional datasets predominantly composed of multiple-choice questions, GeoVQA integrates Numerical Answer Questions and theorem-proving questions, thus broadening its educational scope and depth. The dataset is stored in JSON format, including details such as chapter titles, questions, answers, associated images, and educational levels, effectively covering a wide range of geometric concepts.

Our dataset comprises 1,438 multimodal question-answer pairs, meticulously extracted from mathematics textbooks, and has been further expanded to a total of 4,440 problems. The dataset's statistical details are provided in Table [5.1](#). Each question is paired with two distinct solutions: the original solution from the textbook and a restructured version through Gemini Pro Vision. This restructured solution deconstructs the problem into fundamental components, promoting a systematic problem-solving approach and enhancing the interpretability of computational models.

#### 5.3.1.1 Dataset Description

##### Dataset Extraction

This process was carried out systematically by utilizing resources from high-school textbooks and previous years' question papers from a variety of Indian educational boards, such as the Central Board of Secondary Education (CBSE) and the Maharashtra Board. These educational materials were sourced in PDF format from online platforms and converted using Mathpix, a tool proficient in transforming PDF files into Mathpix

Markdown (MMD) and further into LaTeX documents. This transformation enabled the efficient extraction of geometric question-answer pairs, integrating the detailed capabilities of LaTeX with the straightforwardness of Markdown. Our primary focus was on geometry-related questions that included accompanying images. A team of human experts meticulously extracted these questions along with their corresponding solutions, images, educational standards (grade), and specific geometry topics. This thorough approach enabled the construction of a detailed and relevant dataset for our study. The chapter-wise distribution of this dataset across nine distinct geometric concepts is depicted in Figure 5.1

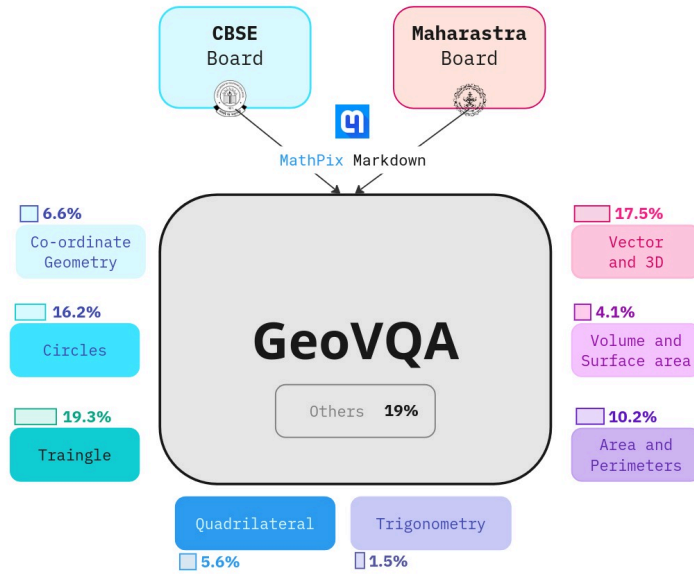


Figure 5.1: Data Extraction and topic-wise distribution

Table 5.1: GeoVQA Dataset statistics

Diff.	Type	Original	Augmented	Total
Easy	Num.	244	488	<b>732</b>
Medium	Num.	661	1322	<b>1983</b>
Hard	Num.	132	264	<b>396</b>
Easy	Prov.	104	208	<b>312</b>
Medium	Prov.	283	566	<b>849</b>
Hard	Prov.	56	112	<b>168</b>
Total		<b>1480</b>	<b>2960</b>	<b>4440</b>
Test set				150

## Dataset Augmentation

In the augmentation phase, our dataset was expanded through the creation of additional question-answer pairs corresponding to diagrams extracted from textbooks. This approach was designed to increase the diversity and depth of the dataset, thus providing a comprehensive resource for advancing multimodal question-answering research.

### **5.3.1.2 Diagram Description Generation**

During the Diagram Description Generation phase, we utilized the Gemini Vision Pro API to automatically generate descriptions of the diagrams. This step was crucial for enriching the dataset with essential visual information, which is integral to a comprehensive multimodal question-answering framework.

The methodology involved presenting the API with both a geometric question-answer pair and its associated diagram. The prompt was carefully crafted to ensure that the descriptions generated were centered on the visual elements of the diagram, avoiding any repetition of the question or answer content. The aim was to produce descriptions that were succinct and directly relevant to the key features of the diagram. This setup, comprising the geometric question, the answer, and the diagram, enabled the model to accurately capture the primary geometric characteristics of each diagram.

### **5.3.1.3 Question Generation**

Using Gemini Pro Vision [Team \*et al.\* \(2023\)](#), a multimodal LLM, new question-answer pairs were systematically generated based on the diagrams extracted from textbooks. Each initial diagram-QA pair served as a template to create two additional QA pairs. The generation process involved analyzing the visual information embedded within the diagrams in conjunction with the textual context provided by the original questions and answers. This augmentation approach significantly enriched the dataset by introducing variations in the structure of questions while retaining semantic consistency with the original content. By diversifying the question formats, the goal was to train multimodal models to handle a wider range of inquiries, thereby improving their capacity to interpret and respond to varied inputs.

### 5.3.1.4 Solution Regeneration

Following question generation, solutions were regenerated using Gemini Pro, a text-based LLM. A structured prompt guided this regeneration process, which systematically deconstructed and reconstructed solutions to reinforce problem-solving methodologies. This process involved breaking down problems into their core components, highlighting the relevant concepts and theories, and developing structured methodologies for resolving the problems. Each computational step was executed with precision, leading to the presentation of a cohesive final solution. The aim of this regeneration process was to instill a deeper understanding of mathematical principles and problem-solving techniques within the multimodal models, equipping them to address complex questions with enhanced accuracy and efficiency. Through this systematic regeneration, the models were trained to identify underlying patterns and relationships, thereby improving their comprehension and response abilities. Below is a structured prompt involving different solution steps used for solution regeneration.

```
Prompt 1 : To Regenerate the Solution of Geometric Problems
```

```
# Question : {Geometry Questions}
```

```
# Image : {Corresponding Image}
```

```
# Data Identification: "[List the key values and information provided in the question. Format this section as a JSON string.]"
```

```
# Problem Analysis: "[Break down the problem into its fundamental components. Present this analysis as a JSON string.]"
```

```
# Theoretical Framework: "[Outline the relevant concepts, theories, or principles that apply to the problem. Format this explanation as a JSON string.]"
```

```
# Methodology Development: "[Develop the formulae, algorithms, or theorems necessary to solve the problem, if applicable. Describe these methodologies in a JSON string format.]"
```

```
# Computation: "[Perform the necessary calculations or logical steps to arrive at the solution. Present these computations as a JSON string format.]"
```

```
# Solution: "[Present the final answer or solution to the problem in a JSON string format. Ensure clarity and conciseness in the explanation.]"
```

### 5.3.1.5 Model Evaluation and Fine-tuning

In this research, we analyze the performance of LLaVA versions 1.5 and 1.6, GLLaVA 7B, and Gemini Pro Vision on three distinct geometry datasets: GeoVQA (Ours), PGPS9K, and Geometry3K. To assess the models' adaptability and reasoning capabilities in unfamiliar contexts, we employ three methodologies: Zero-Shot, Few-Shot (encompassing both Static and Dynamic), and Fine-Tuning. For testing the LLaVA models, various configurations such as Vicuna 7/13B [Zheng et al. \(2024\)](#) and Mistral 7B [Jiang et al. \(2023\)](#), among others, are utilized. Comprehensive results and discussions are presented in the Results section, providing valuable insights into the efficacy of current methodologies for addressing educational geometry challenges.

### 5.3.1.6 Zero-shot Inference Analysis

During the zero-shot inference phase, our primary goal is to evaluate the generalization capabilities of the models. For this purpose, we use two models: LLaVA [Liu et al. \(2024\)](#) and G-LLaVA [Gao et al. \(2023\)](#). These models are tested with geometry-related questions from our dataset that were not part of their training data. The objective is to measure their ability to leverage broad knowledge and apply deductive reasoning to address new and specific geometry problems without prior exposure. This strategy is essential to understanding how well these models can infer and resolve problems beyond their training experience, indicating their practical utility and generalization strength. The zero-shot results are summarized in Table [5.2](#).

Table 5.2: Comparison of the Accuracy (%) on Different Models with Zero-Shot Inference on GeoVQA (Ours) Dataset and Benchmark Datasets: PGPS9K and Geometry3K

Model	GeoVQA	PGPS9K	Geometry3K
LLaVA-1.5-7B	3	4	1
LLaVA-1.5-13B	7	2	14
LLaVA-1.6-7B	3	1	11
LLaVA-1.6-13B	3	8	7
G-LLaVA-7B	18	10	21
Gemini-Pro	44	30	47
GPT-4	<b>60</b>	<b>36</b>	<b>51</b>

### 5.3.1.7 Few-shot Inference Analysis

**Static Few-shot Inference** Since LLMs are recognized as Few-shot learners [Brown et al. \(2020\)](#), the objective in the static few-shot inference configuration was to evaluate the models' performance when provided with only a minimal set of additional information. In this setup, each model received two example questions alongside their respective answers, which were randomly selected from the training portion of our dataset. Afterwards, a novel geometry problem was presented to the model. This methodology aims to simulate real-world scenarios where models are given limited yet pertinent examples to aid in solving new problems, thereby assessing their ability to generalize from given instances to novel questions.

#### **Dynamic Few-shot Inference:**

In contrast to the static method, the dynamic few-shot inference approach involves carefully selecting two sample questions and their corresponding answers based on their relevance and similarity to the new geometry question being posed. The main goal of this method is to determine whether the inclusion of contextually and conceptually aligned examples enhances the model's ability to reason and solve mathematical problems effectively. Preliminary results indicate that closely aligning the training examples with the context of the new question can lead to a notable improvement in the model's problem-solving performance. This suggests a promising strategy for advancing few-shot learning in the domain of geometric problem-solving.

Refer to Table [5.3](#) for results pertaining to few-shot inference outcomes.

### 5.3.1.8 Fine-Tuning

We conducted fine-tuning of the LLaVA 1.5 / 1.6 and GLLaVA models on four distinct datasets: the initial dataset containing 1480 question-answer pairs, a regenerated version of the same dataset using Gemini Pro vision, an expanded dataset comprising 2233 pairs, and an augmented version of this expanded dataset with regenerated solutions. For each model, the fine-tuning process was executed with a learning rate of  $3e^{-5}$ , utilizing a batch size of 4 per GPU, over a total of 2 epochs. The training environment consisted of an NVIDIA DGX server equipped with A-100 GPUs, each featuring 40 GB of RAM. To validate the models' performance after fine-tuning, we evaluated them on the PGPS9K and Geometry3K benchmark datasets. The results of these evaluations

Table 5.3: Comparison of Different Models’ Accuracy (%) with Few-Shot Inference on GeoVQA (Ours) Dataset

Model	GeoVQA Accuracy (%)
<i>Static-Few-shot Inference</i>	
LLaVA-1.5-7B	26
LLaVA-1.5-13B	31
LLaVA-1.6-7B	34
LLaVA-1.6-13B	27
G-LLaVA-7B	6
Gemini-Pro	<b>38</b>
<i>Dynamic-Few-shot Inference</i>	
LLaVA-1.5-7B	27
LLaVA-1.5-13B	34
LLaVA-1.6-7B	28
LLaVA-1.6-13B	25
G-LLaVA-7B	5
Gemini-Pro	<b>40</b>
GPT-4	78

can be found in Table 5.4

Table 5.4: Comparison on different model Accuracy (%) with Few-shot inference on GeoVQA (ours) dataset and Benchmark Dataset: PGPS9K and Geometry3K. The highest accuracy among each dataset is highlighted with bold and green, red, and yellow colour highlights across each model and dataset to represent the highest score on open source models.

Model	Settings	GeoVQA	PGPS9K	Geometry3K
<i>Instruction-Tuning: GeoVQA (1480) dataset</i>				
G-LLaVA-7B	SFT <sub>original</sub>	19	5	11
G-LLaVA-7B	SFT <sub>regenerated</sub>	14	2	17
<i>Instruction-Tuning: Augmented Dataset (4440)</i>				
LLaVA-1.5-7B	SFT <sub>original</sub>	13	8	17
LLaVA-1.5-13B		24	7	16
LLaVA-1.6-7B		21	8	22
LLaVA-1.6-13B		27	5	10
G-LLaVA-7B		20	17	7
LLaVA-1.5-7B	SFT <sub>regenerated</sub>	23	26	34
LLaVA-1.5-13B		26	40	38
LLaVA-1.6-7B		17	21	34
LLaVA-1.6-13B		30	29	29
G-LLaVA-7B		28	12	18
Gemini-Pro	-	44	30	47
GPT4	-	60	36	51

### 5.3.1.9 Experimental Results

Figure 5.3 and 5.4 show the step-by-step performance of solution steps generated by the base and fine-tuned variants of models on the GeoVQA dataset. In our experiments, we evaluated the effectiveness of prompting the LLaVA v1.6 13B model to generate solutions in a stepwise manner. The results of our test set showed that the LLaVA v1.6 13B model without a step-by-step prompt achieved an accuracy of 23%. In contrast, the model with a step-by-step prompt improved to an accuracy of 30%, representing a 7% increase. Based on these findings, we concluded that the step-by-step prompt yields better results, and therefore, this prompting technique was used throughout the rest of our experiments. We evaluated the solution generated by the model based on the following solution steps: Data identification(DI), Problem Analysis (PA), Theoretical Framework(TF), Methodology Development (MD), and Computation(C). While the majority of models demonstrated strong performance in data identification and problem analysis, but exhibit a sharp decline in the theoretical framework and computation step.

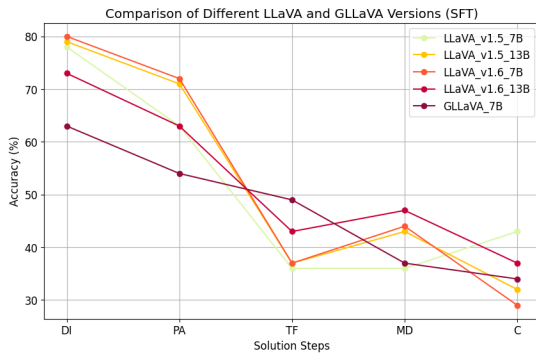


Figure 5.3: Fine-tuned models performance on solution steps

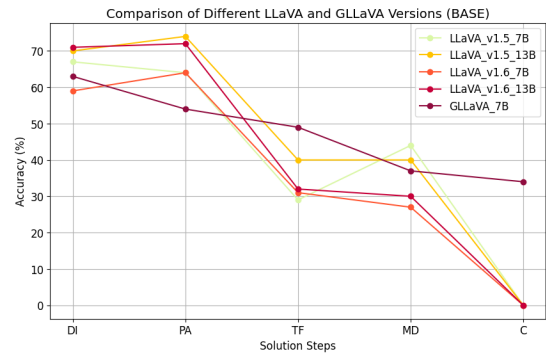


Figure 5.4: Base models performance on solution steps

The figure 5.5 shows top@1 accuracy of LLaVA-based and GLLaVA models on their base and finetuned variants. Fine-tuned variants of LLaVA show better performance as compared to their base variants.

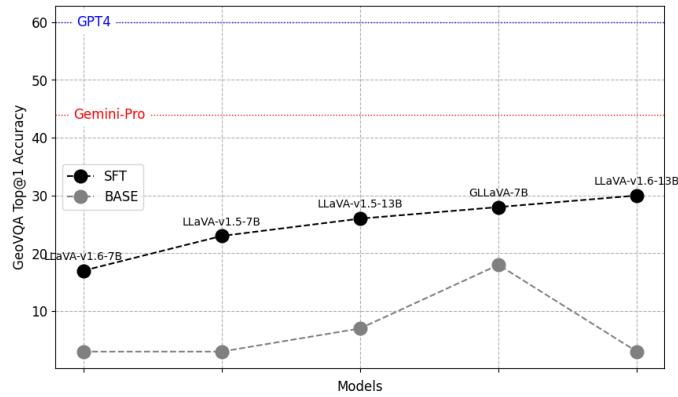


Figure 5.5: Top@1 accuracy of base and fine-tuned models

### 5.3.1.10 Evaluation Metrics

#### 5.3.1.11 Final Solution Evaluation using Gemini API

In our study, we use accuracy as the main evaluation metric. It is crucial to acknowledge that prior research has predominantly relied on Top-10 accuracy, which involves generating ten sequences and considering the first correct solution as the correct prediction. By contrast, our study specifically measures Top-1 accuracy. During the instruction tuning phase, we design the model to generate outputs in a fixed format. For the purpose of evaluation, we employ regular expressions to directly extract predicted responses from the generated answers. If the regular expression fails to extract a valid response, the answer is considered incorrect. The Prompt simplifies Gemini API-based accuracy evaluations by focusing exclusively on comparing final outputs within files. It constructs prompts to guide evaluators through a clear and simple process, where they compare the final answers recorded in "Ground Truth" with those in "Prediction" data. The evaluators are instructed to label the comparison as "Yes" if the answers match or "No" if they differ. Moreover, the prompt encourages providing a justification for the chosen label, promoting transparency and uniformity in human assessments.

#### 5.3.1.12 Step by Step Evaluation using Gemini API

**Reasoning Score Evaluation:** The prompt initiates a standardized Gemini-based framework for evaluating the accuracy of mathematical solutions. It designs prompts that lead evaluators through a structured two-step assessment. In the initial step, evaluators com-

pare the predicted solution with the ground truth by examining the mathematical concepts, computational steps, and principles applied. This involves analyzing the overall strategy, each individual step, and the logical reasoning used. In the second step, if relevant, evaluators check whether the final outcomes of both solutions align. After completing this analysis, the prompt instructs evaluators to provide a definitive justification (Yes or No) for their evaluation supported by clear reasoning.

### **5.3.2 Improving Multimodal LLM’s Ability In Geometric Problem Solving, Reasoning, And Multi-Step Scoring**

#### **5.3.2.1 Methodology**

#### **5.3.2.2 Dataset Extraction**

This process was systematically carried out by utilizing resources from high-school textbooks and previous years’ question papers from various Indian educational boards, such as CBSE (Central Board of Secondary Education) and the Maharashtra Board. These educational resources were sourced online in PDF format and processed using Mathpix, a tool specialized in converting PDF content into Mathpix Markdown (MMD) and then into LaTeX documents. This conversion allowed us to efficiently extract geometric question-answer pairs, combining the flexibility of Markdown with the advanced features of LaTeX. Our primary focus was on geometry-related questions that included images. A team of human experts manually extracted these questions, along with their solutions, corresponding images, educational standards (grade), and specific geometry topics. This detailed approach enabled us to develop a comprehensive and relevant dataset for our research. The chapter-wise distribution of our dataset, covering 9 distinct geometric concepts, is illustrated in Figure [5.6](#) and the corresponding statistics are summarized in Table [5.5](#).

Table 5.5: GPSM4K Finetuning Dataset Statistics

Difficulty	Question Type	Original	Augmented	Total
		Samples	Samples	Samples
Easy	Numericals	250	500	750
Medium	Numericals	700	1400	2100
Hard	Numericals	140	280	420
Easy	Theorems	109	218	327
Medium	Theorems	300	600	900
Hard	Theorems	70	140	210
Total Samples		<b>1480</b>	<b>2960</b>	<b>4440</b>

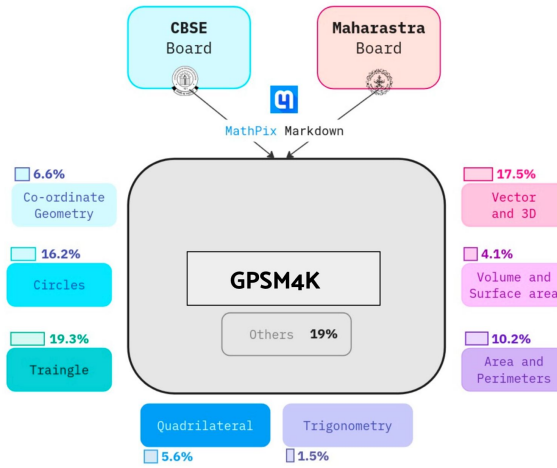


Figure 5.6: Data Extraction and topic-wise distribution

Table 5.6: GPSM4K Inference Dataset Statistics

Difficulty	Question Type	Original Samples
Easy	Numericals	70
Medium	Numericals	56
Hard	Numericals	34
Easy	Theorems	18
Medium	Theorems	12
Hard	Theorems	10
Total Inference Samples		<b>200</b>

### 5.3.2.3 Dataset Augmentation

**Diagram Description Generation.** In the Diagram Description Generation phase, the Gemini Vision Pro API was employed for the automated generation of diagram descriptions. This stage was crucial for augmenting the dataset with vital visual information, essential for a holistic multi-modal question-answering framework.

The methodology involved presenting the API with a geometric question-answer pair and the respective diagram. The prompt was meticulously designed to ensure that the generated descriptions focused on the diagram's visual elements, without echoing the content of the question or answer. The objective was to produce descriptions that were concise and directly pertinent to the diagram's features. Alongside this prompt, the geometric question and answer and the diagram were provided, enabling the model to generate a description that accurately captured the key features of the diagram.

**Question Generation.** Using Gemini Pro vision [Team \*et al.\* \(2023\)](#), a multimodal LLM, we systematically generated new question-answer pairs based on diagrams extracted from textbooks. Each original diagram-QA pair served as a template for the generation of two additional QA pairs. The generation process involved considering the visual information contained within the diagrams along with the textual context of the questions and answers. This augmentation method significantly enriched the dataset by introducing variations in question formulation while maintaining semantic coherence with the original content. By diversifying question formulations, we aimed to train multimodal models to handle a broader range of inquiries, enhancing their ability to comprehend and respond to diverse inputs.

**Solution Regeneration.** After question generation, solutions were regenerated using Gemini Pro, a textual LLM. Employing a structured prompt, solutions were systematically deconstructed and reconstructed to reinforce problem-solving methodologies. The regeneration process involved breaking down the problem into its fundamental components, outlining relevant concepts and theories, and developing appropriate methodologies for problem resolution. Computation steps were performed meticulously, leading to the presentation of a final solution. This regeneration process aimed to instill a deeper understanding of mathematical principles and problem-solving strategies within multimodal models, enabling them to tackle complex questions with greater accuracy and efficiency. Through systematic regeneration, multimodal models were

trained to discern underlying patterns and relationships, facilitating enhanced comprehension and response capabilities. Figure 5.7 shows the entire data augmentation pipeline as described above.

```
# Question : {Geometry Questions}
# Image : {Corresponding Image}

# Data Identification: "[List the key values and information provided
in the question.
Format this section as a JSON string.]"

# Problem Analysis: "[Break down the problem into its fundamental components.
Present this analysis as a JSON string.]"

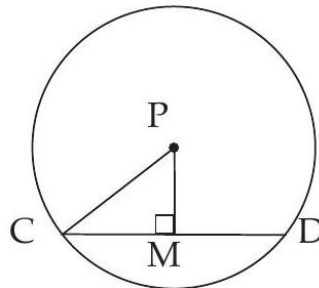
# Theoretical Framework: "[Outline the relevant concepts, theories, or
principles that apply to the problem. Format this explanation as a
JSON string.]"

# Methodology Development: "[Develop the formulae, algorithms, or theorems
necessary to solve the problem, if applicable. Describe these methodologies
in a JSON string format.]"

# Computation: "[Perform the necessary calculations or logical steps to
arrive at the solution. Present these computations as a JSON
string format.]"

# Solution: "[Present the final answer or solution to the problem in a JSON
string format. Ensure clarity and conciseness in the explanation.]"
```

Figure 5.7: Prompt used for Solution Regeneration



### Original Q&A Pair

---

**Question:** The diameter of a circle is 26 cm and the length of the chord of a circle is 24 cm. Find the distance of the chord from the center.

**Answer:** Given:

- (i) A circle with center 'P' and diameter 26 cm.
- (ii) Length of chord CD = 24 cm
- (iii) seg PM  $\perp$  chord CD, C – M – D

Diameter of the circle = 26 cm

...(Given)

$$\text{Radius} = \frac{\text{Diameter}}{2} = \frac{26}{2}$$

$$\therefore \text{Radius of the circle} = 13 \text{ cm}$$

$$\therefore PC = 13 \text{ cm}$$

Seg  $PM \perp$  chord  $CD$

...(Given)

$\therefore CM = \frac{1}{2}CD$ ...(Perpendicular drawn from the centre of the circle to the chord bisects the chord.)

$$CM = \frac{1}{2} \times 24 = 12 \text{ cm}$$

In  $\triangle PMC$ ,  $\angle PMC = 90^\circ$

...(Given)

$$\therefore PC^2 = PM^2 + CM^2 \dots (\text{Pythagoras theorem})$$

$$\therefore 13^2 = PM^2 + 12^2$$

$$\therefore 169 - 144 = PM^2$$

$$\therefore PM^2 = 25$$

$$\therefore PM = 5 \text{ cm}$$

...(Taking square roots)

---

### Diagram Description Generation:

---

A circle with center  $P$  and a diameter of 26 cm. A chord  $CD$  of length 24 cm is drawn in the circle.  $PM$  is perpendicular to  $CD$  and  $M$  is the mid-point of  $CD$ .

---

### Question Augmentation

---

#### Question 1:

The diameter of a circle is 13 cm, and the length of a chord of the circle is 12 cm. Find the distance of the chord from the center.

#### Answer 1:

Let's denote the radius of the circle as  $r$ .

Given that the diameter of the circle is 13 cm, so the radius,  $r$ , is half of the diameter, i.e.,  $r = \frac{13}{2} = 6.5 \text{ cm}$ .

Since  $PM$  is perpendicular to  $CD$ , it bisects  $CD$ , so  $CM = \frac{1}{2}CD$ .

$$CM = \frac{1}{2} \times 12 = 6 \text{ cm}$$

In  $\triangle PMC$ , as  $\angle PMC = 90^\circ$ , by Pythagoras' theorem:

$$PC^2 = PM^2 + CM^2$$

Substituting the values, we get:

$$6.5^2 = PM^2 + 6^2$$

$$42.25 = PM^2 + 36$$

$$PM^2 = 42.25 - 36 = 6.25$$

$$PM = \sqrt{6.25} = 2.5 \text{ cm}$$

Thus, the distance of the chord from the center is 2.5 cm.

#### 5.3.2.4 Overview of Experimentation and Model Evaluation

In this study [Anand et al. \(2024b\)](#), we conducted a series of experiments to evaluate the performance of Large Vision Language Models (LVLMs) using the newly developed GPSM4K dataset. Our experiments were designed to assess the models' capabilities across various scenarios, including zero-shot inference, fine-tuning, and the impact of multimodal enhancements. We tested different configurations of the LLaVA models (versions 1.5 and 1.6), GLLAVA models, and Gemini Pro Vision, comparing their performance across multiple datasets such as GPSM4K, PGPS9K, and Geometry3K. The LLaVA models were evaluated using Vicuna 7/13B [Zheng et al. \(2024\)](#) and Mistral 7B [Jiang et al. \(2023\)](#) configurations, among others. Additionally, we explored the effectiveness of image captioning and Retrieval-Augmented Generation (RAG) techniques in improving model accuracy. All accuracies are reported based on experiments conducted on a test set of 150 geometry problems. These experiments were structured to provide a comprehensive evaluation of the models' generalization abilities, the influence of visual encoders, and the advantages of integrating multimodal data in geometric problem-solving tasks. Detailed methodologies and results of each experiment are dis-

cussed in the following sections.

### 5.3.2.5 Zero-shot Inference Analysis

In the zero-shot inference phase, the primary goal is to evaluate the models' generalization abilities. For this, we utilize two models, LLaVA [Liu et al. (2024)] and G-LLaVA [Gao et al. (2023)], exposing them to geometry-related questions from our GPSM4K inference dataset detailed in Table 5.6, which were not part of their pretraining. The focus is to determine how well these models can employ overarching knowledge and deductive reasoning to address novel and specific geometry problems without having prior exposure to them. This approach is fundamental in gauging the models' potential to infer solutions and handle problems beyond their training domain, providing insight into their real-world applicability and generalization prowess.

The results, displayed in Table 5.7, indicate that GPT-4 achieved the highest accuracy across all tested datasets, showcasing its superior generalization ability. Specifically, GPT-4 attained 60% accuracy on the GPSM4K dataset, 36% on PGPS9K, and 51% on Geometry3K. Gemini-Pro also demonstrated strong performance, with 44% accuracy on GPSM4K, 30% on PGPS9K, and 47% on Geometry3K. Among the LLaVA models, LLaVA-1.5-13B and LLaVA-1.6-13B performed relatively better. G-LLaVA-7B also showed notable generalization, achieving 18% accuracy on GPSM4K, 10% on PGPS9K, and 21% on Geometry3K. These findings underscore the efficacy of GPT-4 and Gemini-Pro in zero-shot inference settings, while also pointing to areas for improvement in the GLLaVA-based models.

The strong performance of both GPT-4 and Gemini-Pro in these zero-shot inference tasks can be largely attributed to their extensive parameter counts and comprehensive training on vast datasets, which allows them to generalize more effectively to previously unseen problems.

### 5.3.2.6 Experiment 1

*"Why do we need different datasets for Geometry Problems?"*

Table 5.7: Comparison of the Accuracy (%) on different models with Zero-shot inference on GPSM4K (Ours) dataset and Benchmark Dataset: PGPS9K and Geometry3K

Model	GPSM4K	PGPS9K	Geometry3K
LLaVA-1.5-7B	3	4	1
LLaVA-1.5-13B	7	2	14
LLaVA-1.6-7B	3	1	11
LLaVA-1.6-13B	3	8	7
G-LLaVA-7B	18	10	21
Gemini-Pro	44	30	47
GPT-4	<b>60</b>	<b>36</b>	<b>51</b>

In this experiment, we utilized the PGPS9K dataset, which contains more than 9,000 plane geometry problems annotated with detailed diagrams and interpretable solution programs, to serve as a benchmark for our study. In addition to this dataset, we also fine-tuned LLaVA and G-LLaVA models on our custom-developed GPSM4K dataset to compare their performance on a test set consisting of 150 high school-level geometry problems. Fine-tuning was carried out with a learning rate of  $3e^{-5}$ , a batch size of 4 per GPU, and for 2 epochs in total. This process was performed using an NVIDIA DGX server equipped with A-100 GPUs, each having 40 GB of RAM. Our aim in this experiment was to assess how effectively these models can handle geometry problems that require step-by-step solutions, involving the application of geometric theorems and numerical reasoning.

The results presented in Table 5.8 show that models fine-tuned on the GPSM4K dataset generally surpassed those fine-tuned on the PGPS9K dataset. For instance, the base LLaVA 1.6 models with Vicuna and Mistral variants exhibited notable improvements in accuracy when trained on the GPSM4K dataset. Specifically, GPT-4 achieved the highest accuracy of 56.66%, followed by Gemini at 35.33%, highlighting the proficiency of these models on the GPSM4K test set.

The higher performance of models trained on the GPSM4K dataset can be explained by the dataset’s step-by-step solution annotations, which provide a richer learning experience. In contrast, the PGPS9K dataset primarily contains multiple-choice questions

(MCQs) with accompanying diagrams, which may not require the same level of detailed reasoning needed to solve complex problems. The structured solutions in GPSM4K offer models a more methodical approach to learning and applying geometric reasoning in problem-solving.

Table 5.8: Performance of LVLMs on GPSM4K test set

Models	Trained on	Accuracy (%) on GPSM4K test set
Base LLaVA 1.6 + Vicuna 7B	PGPS9K	8
Base LLaVA 1.6 + Vicuna 13B	PGPS9K	8.66
Base LLaVA 1.6 + Mistral 7B	PGPS9K	7.33
Base LLaVA 1.6 + Vicuna 7B	Ours	22.66
Base LLaVA 1.6 + Vicuna 13B	Ours	24
Base LLaVA 1.6 + Mistral 7B	Ours	24.66
Base LLaVA 1.6 + Vicuna 34B	-	27.33
Base LLaVA3 8B	-	22.6
GLLAVA 7B + LLaMa 1.5	GeoQA++	10.66
GLLAVA 13B + LLaMa 1.5	GeoQA++	22.66
GLLAVA 7B + LLaMa 1.5	Ours	17.33
GLLAVA 13B + LLaMa 1.5	Ours	25.33
Gemini	-	35.33
GPT-4	-	<b>56.66</b>

### 5.3.2.7 Experiment 2

*"Does the Visual Encoder model limit LLMs' capability to comprehend math, particularly geometry, because of its inability to accurately decode geometric figures, which lack the standard features found in most image datasets?"*

The integration of Visual Encoder models with Language Models (LMs) to enhance the comprehension of mathematical concepts, particularly in the realm of geometry, is an area that warrants thorough investigation. Cutting-edge vision models such as LLaVa Liu *et al.* (2024), often trained on widely-used datasets like COCO or ImageNet, tend to face challenges when it comes to interpreting geometric figures. These datasets typically do not contain the specific characteristics required for geometric imagery, which in turn can hinder the ability of LMs relying on these models to properly interpret and solve complex geometry problems. Thus, evaluating the impact of Visual Encoder models on the mathematical reasoning abilities of LMs, especially in geometry-related tasks, is crucial.

In our study, we use a test set containing 150 image examples, employing models like GIT Wang *et al.* (2022), BLIP Li *et al.* (2022), its enhanced version BLIP-2 Li *et al.* (2023), LLaVa Liu *et al.* (2024), ViT Dosovitskiy *et al.* (2020b), and Gemini Pro to generate image captions. These captions are then assessed against those created by human subject matter experts, utilizing the Jaccard index and Cosine similarity for comparison.

We opted for Jaccard and Cosine similarities instead of traditional evaluation metrics such as BLEU Papineni *et al.* (2002a) and ROUGE Lin (2004a), which prioritize exact n-gram matches. These traditional metrics can be overly rigid in image captioning tasks, where there can be multiple valid ways to describe an image. Cosine similarity, in particular, better accommodates synonyms and paraphrasing, making it more suitable for the varied descriptions possible in image captioning. Moreover, Jaccard and Cosine similarities are less computationally demanding compared to metrics like BERTScore, which rely on deep learning models, making them more feasible for large-scale experiments or situations with constrained computational capacity.

Figure 5.8 presents the captions generated by different models for a sample geometry diagram. Table 5.9 demonstrates that Gemini Pro, which appears to incorporate more advanced techniques for processing both visual and textual data, surpasses other models in performance according to both the Jaccard index and Cosine similarity metrics.

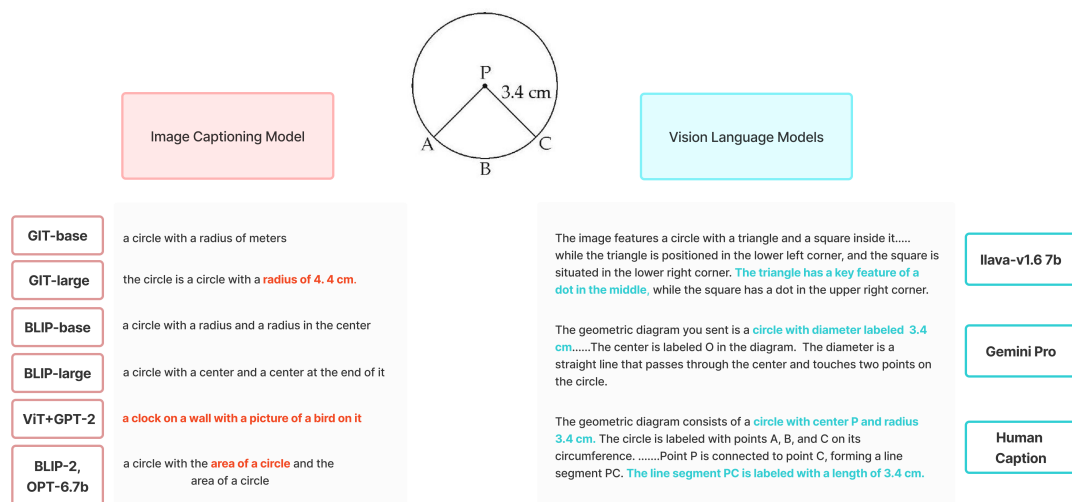


Figure 5.8: Image captioning example

Table 5.9: Image Captioning Result

<b>Models</b>	<b>Mean / Median Jaccard</b>	<b>Mean / Median Cosine similarity</b>
GIT-base	0.08 / 0.07	0.25 / 0.24
GIT-large	0.09 / 0.09	0.28 / 0.27
BLIP-base	0.11 / 0.10	0.23 / 0.22
BLIP-large	0.14 / 0.13	0.25 / 0.23
ViT+GPT-2	0.05 / 0.05	0.06 / 0.06
LLaVa v1.6 7b	0.12 / 0.12	0.43 / 0.42
BLIP-2 OPT-6.7b	0.12 / 0.12	0.33 / 0.32
<b>Gemini Pro</b>	<b>0.36 / 0.28</b>	<b>0.65 / 0.63</b>

### 5.3.2.8 Experiment 3

*"Can adding image captions to multimodal geometry figures improve LLMs' ability to solve complex geometry problems and theorems?"*

In this study, we investigated the impact of adding image captions on the performance of LVLMs in solving geometry problems. Drawing on the results from Experiment 2 (Table 5.11), where Gemini Pro outperformed other models in generating captions based on the Jaccard index and Cosine similarity metrics, we selected both Gemini Pro and human-generated captions for use in the test set. These captions were integrated with the problem prompts and then fed into the GLLAVA 7B and 13B models, both of which had been finetuned on the GPSM4K dataset.

The outcomes, presented in Table 5.10, demonstrate a notable improvement in model accuracy when captions were included, underscoring the value of merging textual and visual information. Notably, the LLaMa3 8B model, which only utilized textual input without image captions, achieved an impressive accuracy of 28.66%. This suggests that with well-crafted text descriptions, the model can successfully handle geometry problems, emphasizing the importance of high-quality, detailed captions.

Moreover, the accuracy of both GLLAVA 7B and 13B models showed an increase

with the inclusion of captions, with human-generated captions delivering a larger improvement compared to those generated by Gemini Pro.

Table 5.10: Model performance on GPSM4K test after adding Image captions

<b>Model Finetuned on GPSM4K + Captions</b>	<b>Accuracy (%) on GPSM4K test set</b>
GLLaVA 7B FT + Gemini	21.33
GLLaVa 13B FT + Gemini	22.66
GLLaVA 7B FT + Human	24.66
GLLaVa 13B FT + Human	26.66

### 5.3.2.9 Experiment 4

*"Can Multi-modal RAG enhance model performance on Geometry problems?"*

We present a robust multimodal RAG framework (see Figure 5.9) that combines multimodal retrieval functionality with in-context learning (ICL). This framework operates by first utilizing a retrieval module designed to fetch the most relevant multimodal question-answer pairs. These retrieved elements are then incorporated into the model’s reasoning process, guiding it through the rationales linked to each fetched item. After the retrieval phase, the model independently synthesizes coherent rationales that demonstrate an in-depth understanding and engagement with the problem at hand. Retrieval-augmented generation (RAG) improves the capabilities of language models by retrieving pertinent information from a vast corpus and using it to enhance the model’s generative output. This method proves particularly effective in producing informed and context-aware responses by leveraging both retrieval and generative mechanisms. In our experiments, we employed a vector database to store question-answer pairs, along with image descriptions from our training dataset. When encountering a new question during inference, cosine similarity was used to retrieve the most relevant questions from the vector database. The retrieved information was then used as context to help the model produce more informative and illustrative responses. We conducted 1-shot RAG experiments on the Gemini Pro and LLaMa3 models, with results

summarized in Table 5.11

Table 5.11: Model performance on GPSM4K test with RAG

Model	Accuracy (%) on GPSM4K test set
Gemini	35.33
Gemini + 1-shot RAG	38
LLaMa 3 8B	22.6
LLaMa 3 8B + 1-shot RAG	26

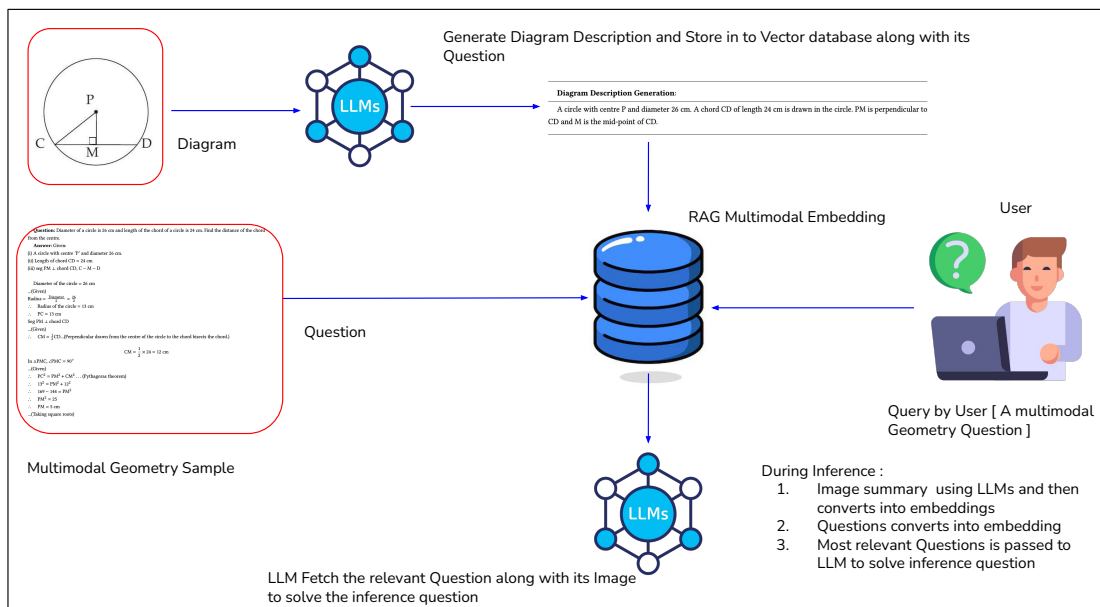


Figure 5.9: Multimodal RAG tackles this challenge by combining the strengths of images and descriptions. It analyzes diagrams to understand the problem and uses text descriptions to find the solution. In 1-shot multimodal RAG, a single problem-solution example is used to find similar problems from a vast dataset. These retrieved examples then help solve new inference questions, showcasing the model’s ability to generalize its knowledge

### 5.3.2.10 Evaluation

In this study, we employed the Gemini Pro API to evaluate the performance of our models. Two distinct approaches were adopted for this evaluation: final answer evaluation and step-by-step evaluation. The key metric used was Top-1 accuracy, which differs from the more commonly used Top-10 accuracy in prior research. Top-1 accuracy takes into account only the first generated sequence, whereas Top-10 accuracy considers any

of the ten generated sequences as correct if they successfully solve the problem. Additionally, to ensure the robustness of our results, we applied a majority count method, repeating the extraction and comparison process three times. This repetition minimizes the effects of anomalies or errors in any single iteration, resulting in more consistent and dependable accuracy measurements.

### Final Answer Evaluation

The evaluation of the final solution follows a two-step approach using the Gemini Pro API. In the first stage, the initial prompt examines both the ground truth solution and the predicted outcome, extracting the final answer from each of them. Then, in the second stage, the extracted answers are compared; if they match, they are labeled as 'Yes', whereas mismatches are marked as 'No'. We opted to use two distinct prompts rather than a single one because this approach allowed the model to generate more precise results.

Instructions: 1. Read the final answer in the Ground Truth. 2. Read the final answer in the Prediction. 3. Return the final answers in a structured format.  
 Ground Truth: <ground truth> Prediction: <prediction>  
 # Output should be in the format: # Ground Truth Answer: <extracted answer> # Prediction Answer: <extracted answer>

Prompt 2: Objective: Compare the final answers and label the comparison.  
 Instructions: 1. Compare the final answer from the Ground Truth and the Prediction. 2. Label the comparison as "Yes" if both answers are identical, and "No" if they differ.  
 Ground Truth Answer: <ground truth> Prediction Answer: <prediction>  
 # Output should be in the format: # Comparison Result: Yes/No **Prompt 1: Extract the final answer from the Ground Truth and the Prediction.**

**Instructions: 1. Read the final answer in the Ground Truth. 2. Read the final answer in the Prediction. 3. Return the final answers in a structured format.**  
**Ground Truth: <ground truth> Prediction: <prediction>**  
**# Output should be in the format: # Ground Truth Answer: <extracted answer> # Prediction Answer: <extracted answer>**

**Prompt 2: Objective: Compare the final answers and label the comparison.**  
**Instructions: 1. Compare the final answer from the Ground Truth and the Prediction. 2. Label the comparison as "Yes" if both answers are identical, and "No" if they differ.**  
**Ground Truth Answer: <ground truth> Prediction Answer: <prediction>**  
**# Output should be in the format: # Comparison Result: Yes/No**

Figure 5.10: Two Prompts Used in Final Solution Extraction and Comparison Respectively

### CoT Step-by-step-evaluation

When compared to general visual question-answering tasks, mathematical problem-solving with MLLMs demands a more refined and structured approach, particularly requiring step-by-step CoT reasoning. This need for detailed reasoning makes the simple binary 'Correct' or 'Incorrect' assessment, typically used in existing benchmarks, in-

Provide a label 'Yes' or 'No' based on the below criteria.

Criteria: 1. Mathematical Concept and Computation Steps: Are they the same in both prediction and ground truth? This includes the approach, steps followed, and principles applied. 2. Final Answer: If applicable, does the final answer match both prediction and ground truth?

Labeling: - Label 'Yes' if the mathematical concept, computation steps, and final answer (if applicable) are the same. - Label 'No' if there are differences in the concept, steps, or final answer (if applicable).

Provide a Reason for your label.  
Ground Truth: <ground truth>  
Prediction: <prediction>  
Label: Yes/No

**Prompt: Evaluate the prediction against the ground truth based on mathematical accuracy.**  
**Provide a label 'Yes' or 'No' based on the below criteria.**

**Criteria: 1. Mathematical Concept and Computation Steps: Are they the same in both prediction and ground truth? This includes the approach, steps followed, and principles applied. 2. Final Answer: If applicable, does the final answer match both prediction and ground truth?**

**Labeling: - Label 'Yes' if the mathematical concept, computation steps, and final answer (if applicable) are the same. - Label 'No' if there are differences in the concept, steps, or final answer (if applicable).**

**Provide a Reason for your label.**  
**Ground Truth: <ground truth>**  
**Prediction: <prediction>**  
**Label: Yes/No**

Figure 5.11: Prompt Used for Step-by-Step Evaluation

sufficient for evaluating the precision and complexity of multi-step reasoning processes. To better address this issue, we introduce a CoT evaluation strategy aimed at rigorously assessing mathematical CoT skills in visual contexts, utilizing a two-stage prompting process with Gemini Pro.

In our evaluation, the MLLM’s output alone is used to prompt Gemini Pro, intentionally excluding the original questions, diagrams, and ground-truth answers. A pre-defined key-step template is applied to each problem to ensure consistency and allow for comparability across evaluations. However, we also incorporate adaptive extraction strategies when necessary, tailored to the distinct output of each MLLM. This flexibility is essential as mathematical problems often permit multiple solution approaches, and different MLLMs may vary in terms of the length and style of their reasoning processes. Thus, the rigid use of a template is supplemented by adaptive methods to maintain precision in the evaluation of CoT.

For the step-by-step evaluation, we used a structured prompt, as illustrated in Figure 5.11, to assess the mathematical accuracy at each stage of the problem-solving process. This approach examines how closely the model’s predictions align with the ground truth, focusing on mathematical concepts, intermediate computational steps, and the final solution. By evaluating each step individually, this method enables a more comprehensive assessment of the models’ problem-solving abilities, ensuring that both the

<b>Data Identification</b>	✓+1	length of diagonal: 13 cm, area of rectangle: 60 cm <sup>2</sup>
<b>Problem Analysis</b>	✓+1	To solve for the length and breadth of the rectangle, we must use the area and the diagonal's length.
<b>Theoretical Framework</b>	✓+1	For the rectangle, the Pythagoras Theorem, links the length, breadth, and diagonal. The area formula is $A = lb$ .
<b>Methodology Development</b>	✓+1	We begin by expressing the length in terms of the breadth (or vice versa) using the area formula..... The perimeter is calculated by adding the length and breadth and multiplying by two.
<b>Computation</b>	✓+1	We start with the equation $lb = 60$ to express one dimension in terms of the other. Then, we use $l^2 + b^2 = 13^2$ to...will yield the length and breadth. Finally, we calculate the perimeter with $P = 2(l + b)$
<b>Solution</b>	✗0	Let the length be $l$ and the breadth be $b$ . From the area, we have $l \times b = 60$ . .....Solving these equations, we find $l = 5$ cm and $b = 12$ cm . Consequently, the perimeter, calculated as $2(l + b)$ , is $2(5 + 12) = 34$ cm.

Figure 5.12: The binary ‘Correct’ or ‘Incorrect’ evaluative approach of existing benchmarks is inadequate to examine the depth and precision of the multi-step reasoning process. To this end, We propose a CoT evaluation strategy to thoroughly assess their mathematical CoT skills in visual contexts, involving Multimodal LLMs.

final answer and the reasoning behind each intermediate step are evaluated for accuracy and correctness.

Gemini Pro is employed to review each key step, offering a detailed evaluation and generating an overall score. We feed the extracted key steps, along with the original questions, diagrams, and ground-truth answers, into Gemini Pro for analysis. Specifically, Gemini Pro assigns a binary score of either ‘1’ (correct) or ‘0’ (incorrect) to each intermediate step and then computes the final score based on the correctness of both the individual steps and the final answer. This CoT evaluation approach(see Figure 5.12) also provides in-depth error analysis for each step, offering crucial insights that can guide the development of MLLMs in solving mathematical problems.

### 5.3.2.11 Results and Discussion

The base model LLaVA 1.6, when fine-tuned on the PGPS9K dataset, consistently exhibited low accuracy rates, ranging between 7.33% and 8.66%. However, there was a marked improvement when this same model was fine-tuned on our dataset, with accuracy increasing significantly to between 22.66% and 24.66% across various configurations of the Vicuna and Mistral models. These results underscore the effectiveness of our dataset in enhancing model performance through fine-tuning.

The GLLaVA models, both in the 7B and 13B configurations, also demonstrated significant improvements when fine-tuned on our dataset compared to their performance on the GeoQA++ dataset. Specifically, the GLLaVA 7B model achieved an accuracy of

17.33% when fine-tuned on our dataset, a considerable improvement from the 10.66% accuracy observed with the GeoQA++ dataset. Similarly, the GLLaVA 13B model’s accuracy increased from 22.66% with the GeoQA++ dataset to 25.33% on our dataset. These improvements highlight the benefits of using our dataset for fine-tuning these models.

Larger models, such as Gemini and GPT-4, displayed superior performance, achieving accuracies of 35.33% and 56.66%, respectively. This indicates that models with larger capacities and parameter counts are better equipped to generalize across complex tasks, such as those involving geometric problem-solving.

The findings also suggest that human-annotated captions contribute more significantly to performance improvements than captions generated by the Gemini model. For example, the GLLaVA 7B FT model’s accuracy increased from 21.33% with Gemini-generated captions to 24.66% when human captions were used. This emphasizes the higher quality and relevance of human-generated captions. Additionally, a comparison of different model architectures shows that the Gemini + Human configuration outperformed the GLLaVA-based configurations, achieving an accuracy of 29.33%. Moreover, as indicated by the results in Table 5.11, the use of RAG appears effective in enhancing the model’s ability to handle more complex tasks. However, despite these advancements, the models still exhibit limitations in leveraging theorem-based knowledge to solve problems effectively.

### 5.3.3 Mathify: Evaluating LLMs on Mathematical Problem Solving Tasks

#### 5.3.3.1 Dataset

For our research [Anand *et al.* (2023d)], we utilized the Math-401 dataset [Yuan *et al.* (2023)], which consists of 401 mathematical problem samples. This dataset features a wide variety of mathematical operations such as addition (+), subtraction (−), multiplication (\*), division (/), exponentiation, trigonometric functions (sin, cos, tan), and logarithmic functions (log, ln). It also includes a mix of integers, decimals, and irrational numbers ( $\pi$ ,  $e$ ). Given that the limited size of this dataset would not be sufficient for training LLMs effectively, we applied data augmentation techniques, expand-

ing the dataset to 302,000 samples. To achieve this augmentation, we employed the *SymPy* Python library, which enabled us to generate mathematical equations along with their respective ground truth solutions. These generated equations included basic arithmetic operations like addition, subtraction, multiplication, and division, as well as more complex expressions involving brackets, reflecting the challenges often posed by real-world mathematical problems. Table 5.12 presents a detailed breakdown of the types of questions included in our augmented dataset. Additionally, we tested our model on four other datasets: GSM-8K Cobbe *et al.* (2021), DeepMind Saxton *et al.* (2019), NumGLUE Mishra *et al.* (2022), and SimulEq Koncel-Kedziorski *et al.* (2016).

Table 5.12: Distribution of Question Types in Augmented Math-401 Dataset

Type	Range	Decimal Places (1 - 4)	Variables	Count
Small Integer	[-20, 20]	×	(x, y)	65,000
Small Decimal	[-20, 20]	✓	(x, y)	35,000
Small Decimal + Integer	[-20, 20]	✓	(x, y)	39,000
Large Integer	[-1000, 1000]	×	(x, y)	39,000
Large Decimal	[-1000, 1000]	✓	(x, y)	25,000
Large Decimal + Integer	[-1000, 1000]	✓	(x, y)	25,000
3 Terms	[-100, 100]	✓	(x, y, z)	25,000
4 Terms	[-100, 100]	✓	(w, x, y, z)	49,000
<b>Total</b>	-	-	-	<b>302,000</b>

### 5.3.3.2 Our Dataset: MathQuest

We have carefully constructed a specialized dataset named MathQuest, which is derived from high school mathematics problems found in NCERT textbooks. MathQuest serves as a comprehensive resource, featuring word problems that vary in complexity and cover a broad spectrum of mathematical topics. Specifically, our dataset spans 14 primary mathematical domains, such as sets, trigonometry, and binomial theorem, among others. The distribution of problems across these different concepts is illustrated in Figure 5.14. MathQuest consists of a total of 223 samples. As shown in the charts, the "Sequence and Series" category has the largest number of problems within the dataset. To offer insight into the structure of our dataset, a sample from MathQuest is showcased in Figure 5.13.

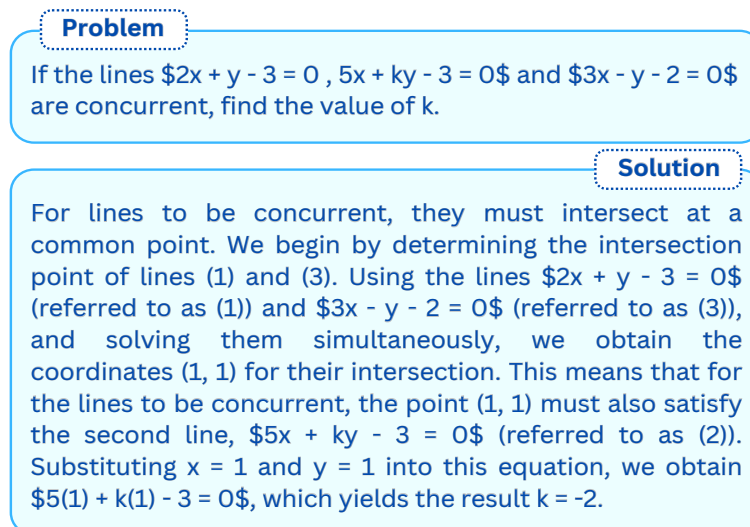


Figure 5.13: Our Dataset **MathQuest** Sample

### 5.3.3.3 Methodology

This study focuses on improving the mathematical problem-solving abilities of LLMs. Initially, we noticed that current open-source models such as LLaMA-2 [Touvron \*et al.\* \(2023a\)](#) and Vicuna [Chiang \*et al.\* \(2023\)](#) encountered difficulties when dealing with basic mathematical tasks, including simple operations like addition and subtraction. This observation spurred our investigation, aiming to enhance the models’ capabilities in understanding and solving mathematical problems with greater accuracy.

To pursue this objective, we adopted an instructive methodology akin to teaching mathematics to students. We began by focusing on fundamental mathematical operators, such as  $+$ ,  $-$ ,  $*$ ,  $/$ , before advancing to more complex operators and expressions. In the same vein, we worked to familiarize the LLMs with the significance and function of these operators and expressions. For this purpose, we utilized the Math-401 dataset [Yuan \*et al.\* \(2023\)](#), a critical resource comprising 401 samples of basic mathematical questions paired with their respective answers. However, due to the dataset’s modest size, we expanded it to include a broader range of questions, increasing both diversity and complexity, which enabled the models to comprehend and apply advanced mathematical concepts during training.

For the fine-tuning process, we utilized three widely recognized LLMs: LLaMA-2 [Touvron \*et al.\* \(2023a\)](#), WizardMath [Luo \*et al.\* \(2023\)](#), and MAMmoTH [Yue \*et al.\* \(2023\)](#). LLaMA-2 [Touvron \*et al.\* \(2023a\)](#) is an enhanced version of its predecessor, LLaMA, featuring improvements from training on a substantially expanded dataset.

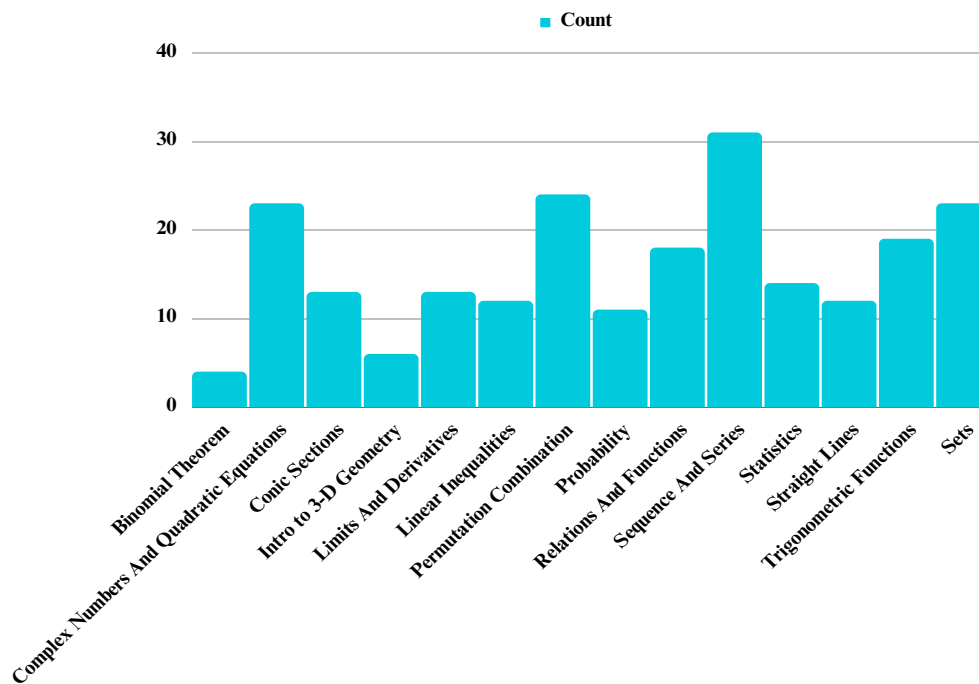


Figure 5.14: Distribution of Count of Samples of each Concept

Key upgrades include a 40% increase in the size of the pre-training corpus, doubling the model’s context length, and introducing grouped-query attention to optimize performance.

WizardMath [Luo et al. \(2023\)](#) presents a novel technique called Reinforcement Learning from Evol-Instruct Feedback (RLEIF). This technique integrates Evol-Instruct and reinforced process supervision methods to evolve the GSM8k and MATH datasets. The pre-trained LLaMA-2 model is then fine-tuned using this evolved data and reward models, culminating in the creation of the WizardMath model.

Finally, the MAmmoTH [Yue et al. \(2023\)](#) models are fine-tuned using the MathInstruct dataset, which was carefully designed for instructional tuning. MathInstruct consists of a collection of 13 mathematical datasets, with six newly curated rationale datasets. It integrates both CoT and program-of-thought (PoT) rationales, offering broad coverage of various mathematical domains. The overall fine-tuning procedure is depicted in Figure. [5.15](#).

Table 5.13: Exact Match Accuracy on 100 Samples from 5 Datasets and MathQuest (Before Fine-tuning on Math-401). (\*) Denotes the Math-401 Subset Used for Augmentation.

Model (Params)	GSM-8K	DeepMind	NumGLUE	SimulEq	MathQuest
<b>LLaMA-2 (7B)</b>	16.0	46.0	37.0	11.0	10.4
<b>LLaMA-2 (13B)</b>	22.0	50.0	42.0	15.0	14.1
<b>WizardMath (7B)</b>	61.0	51.0	54.0	27.0	14.6
<b>WizardMath (13B)</b>	65.0	55.0	70.0	36.0	14.3
<b>MAmmoTH (7B)</b>	43.0	49.0	54.0	23.0	12.2
<b>MAmmoTH (13B)</b>	44.0	48.0	56.0	26.0	18.1

### 5.3.3.4 Experimental Setup, Hyperparameters and Evaluation Metrics

In this section, we explore the experimental details, providing an overview of the setup and hyper-parameters used throughout the process. The core aim of our research was to generate a high school-level mathematical dataset, which includes a wide range of questions with varying levels of complexity, and to establish solid baselines for solving mathematical problems.

To accomplish this, we conducted experiments utilizing three well-known LLMs: LLaMA-2 [Touvron et al. \(2023a\)](#) and WizardMath [Yue et al. \(2023\)](#). The experiments were performed using both the 7B and 13B parameter versions of these models. Our methodology involved two distinct stages. In the first stage, we loaded the original model weights and ran inference on our pre-determined test set. In the second stage, we fine-tuned these models, using the Math-401 [Yuan et al. \(2023\)](#) dataset as an essential component of the fine-tuning process.

The Math-401 [Yuan et al. \(2023\)](#) dataset initially consisted of 401 basic mathematical equations along with their solutions. To increase the dataset’s diversity and scope, we applied data augmentation techniques by adding more complex equations involving operations such as addition (+), subtraction (−), multiplication (\*), division (/), and the use of parentheses (()). This augmentation aimed to develop a more generalized and flexible dataset. Following this, we proceeded to fine-tune the LLMs on this extended Math-401 [Yuan et al. \(2023\)](#) dataset.

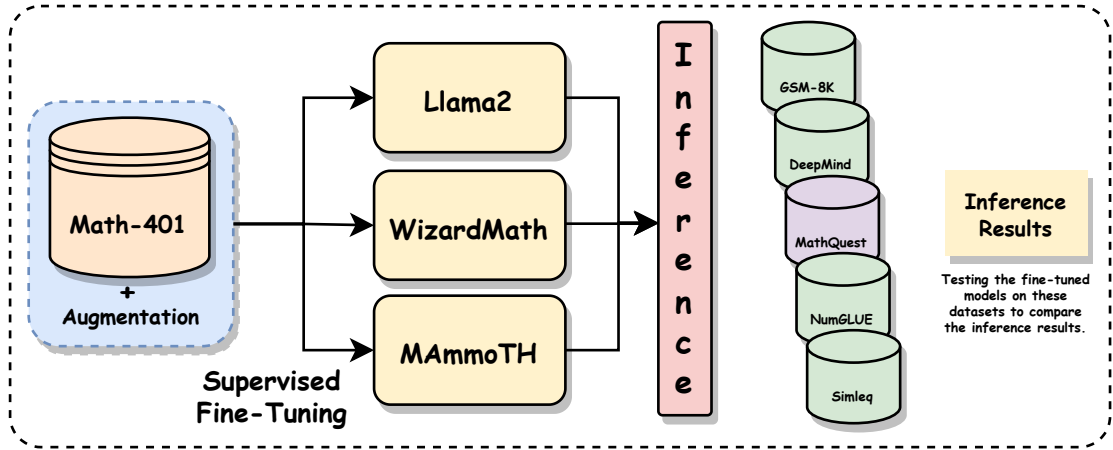


Figure 5.15: This figure shows the fine-tuning flow, the LLMs we use for fine-tuning, and the datasets we use for inference.

Table 5.14: Exact Match Accuracy on 100 Samples from 5 Datasets and MathQuest (After Fine-tuning on Math-401). (\*) Denotes the Math-401 Subset Used for Augmentation.

Model (Params)	GSM-8K	DeepMind	NumGLUE	SimulEq	Math-401*	MathQuest
LLaMA-2 (7B)	30.0	46.0	45.0	15.0	17.0	10.6
LLaMA-2 (13B)	42.0	51.0	54.0	16.0	24.0	20.3
WizardMath (7B)	64.0	55.0	52.0	29.0	15.0	16.01
WizardMath (13B)	68.0	56.0	70.0	38.0	10.0	20.1
MAmmoTH (7B)	56.0	50.0	62.0	24.0	16.0	18.5
MAmmoTH (13B)	67.0	51.0	64.0	34.0	18.0	24.0

The dataset was divided into three subsets: training (241,600 samples), validation (30,200 samples), and test (30,200 samples). To optimize the model performance, we utilized the AdamW optimizer, a well-established method known for its effectiveness in improving outcomes. This step played a pivotal role in achieving the results discussed in this research.

For the fine-tuning process, we adopted the QLoRA [Dettmers \*et al.\* \(2023\)](#) approach, which is specifically designed to enhance memory efficiency while reducing computational costs by employing 4-bit quantization on a pre-trained language model, thereby generating LoRA. Each model was fine-tuned over the course of 10 epochs, with a learning rate set at  $3 \times 10^{-4}$ . Following the fine-tuning, we evaluated the models

using the same test set that was used for pre-fine-tuning inference. The outcomes of this evaluation are shown in Table 5.14 clearly demonstrating the improvements in mathematical problem-solving abilities observed before and after the fine-tuning process.

### 5.3.3.5 Evaluation Metric

We conducted a comprehensive comparison of all model variants to evaluate the effectiveness of the solutions they generated. To gauge performance, we focused on the precision with which the generated solutions aligned with the correct answers across five open-source datasets: GSM-8K, DeepMind, SimulEq, NumGLUE, and Math-401. These datasets offer ground truth solutions, enabling us to compute exact match accuracy for each model.

### 5.3.3.6 Results & Discussion

In this section, we discuss the results of our experiments in the field of mathematical problem-solving. The experiments are based on both our custom-built dataset, MathQuest, and five other publicly available datasets. This thesis introduces baseline performance metrics specifically for the MathQuest dataset. To assess the performance of LLMs across these various datasets, we employ exact match accuracy as the key evaluation metric.

The results are organized into two primary categories: **before fine-tuning** and **after fine-tuning** the models. The main objective of this organization is to assess the model's ability to learn and adapt. Table 5.13 shows the exact match accuracy of three models, each with two variants (7B and 13B), prior to fine-tuning across five datasets, including MathQuest. As outlined in Table 5.13, the performance of all the models was notably lower on the SimulEq dataset and the augmented Math-401 dataset. This lower performance is likely due to the complexity of the problems in these datasets, which often involve additional external knowledge, such as in questions like "Number of red color cards in a deck of 52 cards." To illustrate the improvements made post-fine-tuning, Table 5.14 presents the accuracy results after this fine-tuning process. Across all models, there was a marked improvement in accuracy following fine-tuning, especially with our complex and diverse question-answer dataset. Furthermore, models with 13B parameters achieved higher accuracy than those with 7B parameters.

The key insights from Table 5.13 and Table 5.14 demonstrate that the top-performing model for our MathQuest dataset is **MAMmoTH-13B**, which achieved the highest accuracy post-fine-tuning, at 24.0%. Another important observation is that both MAMmoTH 7B and 13B models generated outputs with precision up to two decimal places, further indicating their accuracy. As highlighted in Table 5.15 the complexity and diversity of the MathQuest dataset present significant challenges, leading to lower accuracy when compared to other datasets.

### 5.3.4 Bridging the Language Gap in Mathematical Reasoning of LLMs

#### 5.3.4.1 Methodology

#### 5.3.4.2 Decomposition Strategy on HAWP Dataset

To enhance the computational precision of LLMs in handling arithmetic tasks with large numbers, we introduce a Decomposition Strategy aimed at multiplication and division operations (Anand *et al.* (2025)). In multiplication, the strategy involves breaking the multiplicand into its place value components—such as hundreds, tens, and ones—and multiplying each by the other multiplicand. These partial products are then combined to arrive at the final result. For division, a similar approach is adopted where the dividend is decomposed into segments, each divided by the divisor, with the resulting quotients summed to produce the final output. This strategy addresses the limitations in calculation skills observed in open-source language models. In this study, we emphasize the presentation and validation of the Decomposition Strategy using the HAWP dataset, which comprises basic mathematical word problems that require single-operation solutions. This allows us to clearly showcase the strategy’s efficacy in a controlled and straightforward environment. While applying this approach to more complex datasets is an exciting prospect, we have deliberately narrowed our focus to the HAWP dataset to ensure a rigorous and detailed evaluation of this innovative strategy.

We utilized 2,336 Hindi arithmetic problems from the HAWP dataset, encompassing fundamental operations such as addition, subtraction, multiplication, and division. As the original dataset lacked solutions, we enhanced it by generating corresponding question-answer pairs using GPT-4. These pairs were meticulously reviewed and cor-

rected by five human experts, resulting in the creation of the Enhanced HAWP dataset.

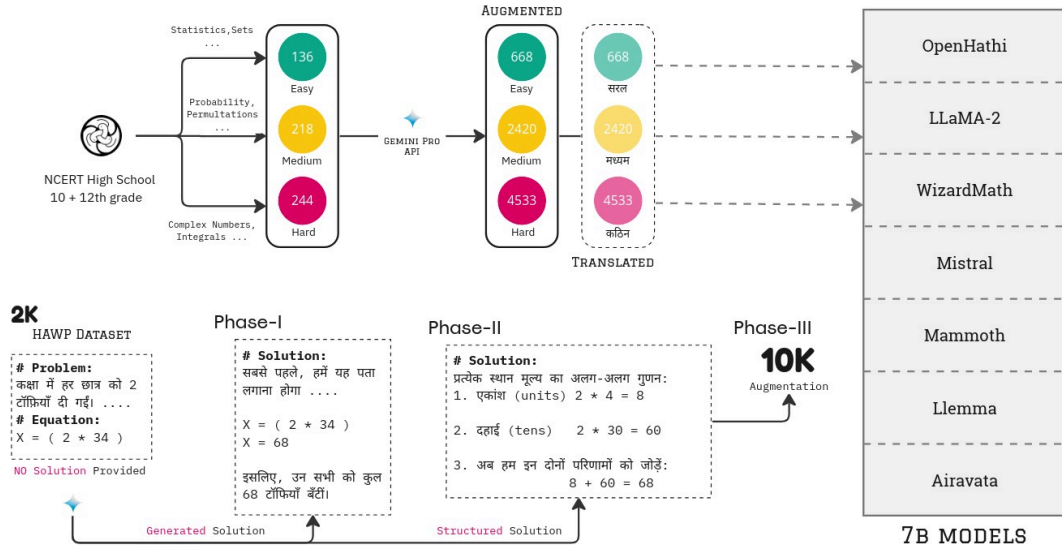


Figure 5.16: Strategic Decomposition: A Method for Simplifying Complex Math Problems into Smaller, Solvable Units to Improve SLLM Capabilities

To assess the effectiveness of the Decomposition Strategy, we applied it to the Enhanced HAWP dataset (refer to Figure 5.16). Manually solved examples based on this strategy were employed in few-shot learning with GPT-4 to refine the remaining solutions in the Enhanced HAWP dataset. These refined solutions, with a 70% training and 30% testing split, were subsequently used to fine-tune the OpenHathi 7B, WizardMath-v1.1 7B, and LLeMMA 7B models. The corresponding accuracy improvements achieved are displayed in Table 5.15.

In the final phase of our research, we explored the advantages of fine-tuning with an augmented dataset that had been previously prepared using the Decomposition Strategy. We expanded the initial 2,000 problems to 10,000 using a one-shot prompting method with GPT-4. These newly generated problems were thoroughly reviewed by five human experts to ensure accuracy, resulting in the HMQA (Hindi Math Questions-Answers) dataset. This expanded dataset was then employed to fine-tune the OpenHathi 7B, WizardMath-v1.1 7B, and LLeMMA 7B models, with the results compared to earlier settings as shown in Table 5.15.

Table 5.15: Performance of LLMs in Hindi Math questions using decomposition strategy. Bold values indicate improvements from the previous step. Underlined values show the highest performance of SLLMs for each operation.

Model	Add	Sub	Mul	Div
<i>Zero-shot Prompting</i>				
OpenHathi-7B	0.35	0.53	<u>0.44</u>	0.33
LLaMA-2-7B	0.39	0.55	0.33	0.5
LLeMMA-7B	0.49	0.63	0.22	0.17
Mistral-7B	0.49	0.55	0.22	0.25
WizardMath-7B	<u>0.63</u>	<u>0.67</u>	0.22	<u>0.67</u>
Gemini-pro	0.78	0.80	1.0	0.75
GPT-4	0.98	0.93	0.88	0.91
<i>Few-shot Prompting</i>				
OpenHathi-7B	<b>0.49</b>	<b>0.63</b>	0.34	<b>0.58</b>
LLaMA-2-7B	<b>0.53</b>	<b>0.72</b>	<b>0.44</b>	0.5
LLeMMA-7B	<b>0.82</b>	<b>0.9</b>	<b>1.0</b>	<b>0.67</b>
Mistral-7B	<b>0.78</b>	<b>0.77</b>	<b>0.56</b>	<b>0.83</b>
WizardMath-7B	<b>0.72</b>	<b>0.73</b>	<b>0.56</b>	0.67
<i>Instruction-Tuning (Enhanced HAWP)</i>				
OpenHathi-7B	<b>0.78</b>	<b>0.85</b>	0.22	0.50
LLeMMA-7B	0.78	0.83	0.67	0.67
WizardMath-7B	<b>0.96</b>	<b>1.0</b>	<b>0.78</b>	<b>0.75</b>
<i>Instruction-Tuning (Enhanced HAWP + Decomposition Strategy)</i>				
OpenHathi-7B	0.78	0.85	0.22	<b>0.67</b>
LLeMMA-7B	<b>0.80</b>	<b>0.925</b>	0.67	<b>0.83</b>
WizardMath-7B	<u>0.95</u>	<u>1.0</u>	<u>0.78</u>	<b>0.83</b>
<i>Instruction-Tuning (HMQA Dataset)</i>				
OpenHathi-7B	<b>0.82</b>	0.85	<b>0.44</b>	<b>0.75</b>
LLeMMA-7B	<b>0.86</b>	<b>0.97</b>	<u>0.67</u>	<u>0.75</u>

### 5.3.4.3 IndiMathQA

We have carefully constructed a comprehensive dataset of math problems, named IndiMathQA, by drawing problems from the official NCERT textbooks<sup>1</sup>, which are widely used in Indian schools. This dataset includes 598 meticulously curated math problems along with their corresponding solutions. These problems are specifically designed for students in grades 10, 11, and 12, covering a broad range of difficulties and encompassing 14 major mathematical topics, such as sets, trigonometry, and the binomial theorem, among others.

<sup>1</sup><https://ncert.nic.in/>

#### 5.3.4.4 LLM Enhancement in Bilingual Mathematics

In this section, we present the methods employed to enhance mathematical reasoning abilities in bilingual environments. The strategies we propose include Structured Solution Creation, Curriculum Learning, and Bilingual Training in both Hindi and English. The process of creating the Bilingual Training Dataset is divided into two main phases: (i) Classification according to Complexity, which is essential for Curriculum Learning, and (ii) Structured Solution Creation followed by Bilingual Translations. Lastly, we showcase the implementation of curriculum learning that enabled bilingual fine-tuning on the prepared training datasets.

#### 5.3.4.5 Classification based on Complexity

We compiled a comprehensive set of mathematical problems, classifying them into three levels of complexity: easy, medium, and hard. This compilation draws from both our dataset and established benchmark datasets, including GSM8K and MATH. The following sections describe the methodology used to classify each problem based on its complexity. By including datasets like GSM8K and MATH, we aim to introduce a broader range of mathematical topics to enhance the diversity of our dataset.

**IndiMathQA:** The IndiMathQA dataset was systematically annotated by a team of five experts, categorizing 136 questions as easy, 218 as medium, and 244 as hard. To verify the consistency of these annotations, we computed the Average Fleiss' Kappa score, which yielded a value of 0.58, reflecting substantial agreement among the annotators and minimizing bias. Following this, the dataset was expanded to 7,823 questions (820 easy, 2,470 medium, and 4,533 hard) through the use of the GPT-4 API. These generated questions underwent a thorough review by the same five experts to ensure correctness and uniformity.

**GSM8K:** The GSM8K dataset consists of grade-school-level math problems, most of which are inherently simple. To ensure our classification was precise, we applied Llama 3 (405B) with specialized prompt engineering to evaluate and rank the problems based on several complexity dimensions, such as Language Understanding, Mathematical Complexity, Reasoning Complexity, Number of Variables, and Conceptual Com-

plexity. For our classification purposes, we selected the 700 least complex questions from GSM8K to form the Easy level.

**MATH:** The MATH dataset comprises competition-grade questions aimed at students in grades 8 to 12. These problems are pre-annotated by complexity, from Level 1 (simplest) to Level 5 (most difficult). In our study, we classified Level 1 questions as Easy, Levels 2 and 3 as Medium, and Levels 4 and 5 as Hard. This resulted in the distribution of 664 Easy, 3,140 Medium, and 3,994 Hard questions for our experiments.

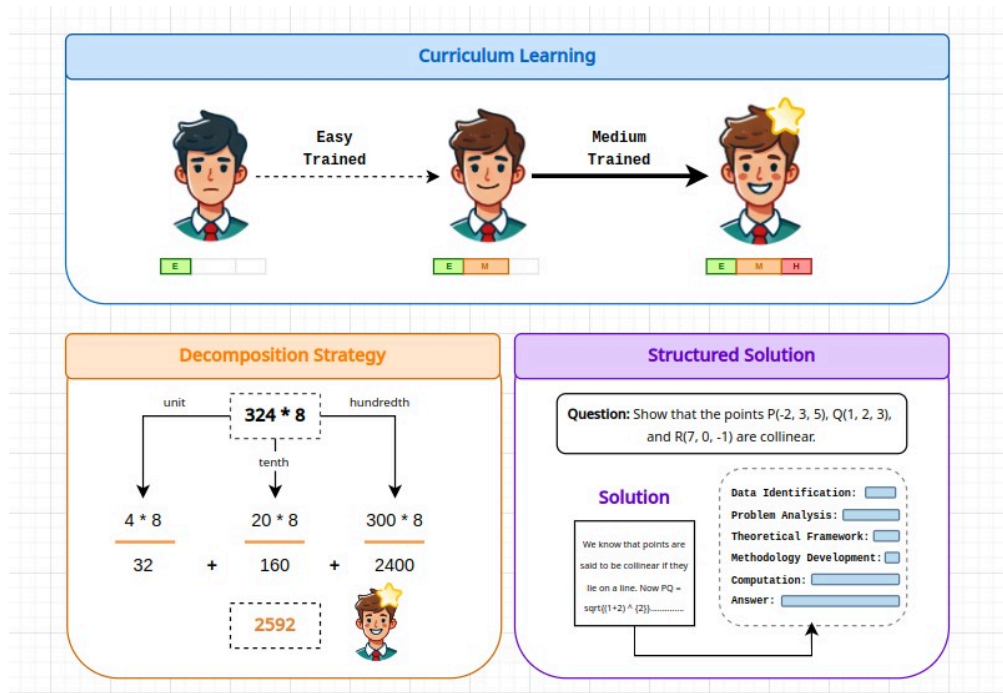


Figure 5.17: Curriculum Learning with Structured Solutions: A Comprehensive Framework to Gradually Guide Models Through Complex Mathematical Challenges.

### 5.3.4.6 Structured Solution Generation and Language Translations

LLMs often face difficulties with hallucinations when tasked with reasoning problems. Upon manual inspection of base model outputs, we observed that LLMs occasionally become overly focused on solving the problem, neglecting the essential theoretical principles required for a precise solution. This aligns with the findings of [Zheng et al. \(2023\)](#), which presents a novel prompting strategy to encourage LLMs to take a step back and ask clarifying questions to deepen their understanding of the problem context. Based on this, we hypothesized that fine-tuning LLMs on solutions that begin by

considering the underlying theoretical framework would lead to more accurate outputs. We extended this concept further by not only focusing on the theoretical aspects but also creating a detailed, structured solution format for fine-tuning. We believe this will encourage the LLM to approach reasoning tasks systematically, resulting in improved accuracy. To operationalize this, we restructured the existing solutions in our datasets into a well-organized format under the following sections: (i) Data Identification, (ii) Problem Analysis, (iii) Theoretical Framework, (iv) Methodology Development, (v) Computation, and (vi) Answer. (see Figure [5.17](#))

To ensure consistency in this structured approach, we provided a few-shot examples demonstrating how sample answers should be divided according to this format. These examples, along with a comprehensive prompt, were fed to GPT-4, which then generated structured solutions for all problems in both the training and testing datasets. A team of five human experts reviewed and corrected any errors in the structured solutions to ensure accuracy.

Our original datasets were in English. After converting the solutions into the structured format, we employed LLAMA 3 (405B) to translate both the questions and their structured solutions into Hindi. Subsequently, we merged the English and Hindi versions of GSM8K, MATH, and IndiMathQA datasets, categorizing them based on difficulty levels: Easy, Medium, and Hard. The final dataset includes 2184 Easy, 5470 Medium, and 8527 Hard problems.

#### **5.3.4.7 Curriculum Learning based Fine-Tuning**

We applied Curriculum Learning to SLLMs with the hypothesis that gradually increasing the complexity of problems during fine-tuning would mimic human learning, where mastering simpler problems helps in tackling more complex ones. Our method utilizes the Easy and Medium datasets, which are carefully designed to cover a broad spectrum of mathematical topics. Each dataset was split into 70% for training and 30% for testing, ensuring a consistent distribution across all difficulty levels: easy, medium, and hard.

The Curriculum Learning approach begins by training the SLLMs on the Easy dataset, producing a model checkpoint referred to as SFT\_Easy. This checkpoint is then further fine-tuned on the Medium dataset, yielding a final checkpoint named SFT\_Easy

+ Medium. We evaluate the performance differences between these checkpoints using the test sets from both benchmark datasets and our curated dataset.

Our hypothesis also extends to the notion that fine-tuning LLMs on a dataset containing parallel question-answer pairs in both English and Hindi could enhance the model’s proficiency in solving mathematical problems in Hindi. Since the LLM is more proficient in English, exposing it to parallel data could help it leverage its strengths in English to build stronger reasoning capabilities in Hindi.

To evaluate this hypothesis, our Curriculum Learning-based fine-tuning is carried out in two distinct ways:

1. Training the SLLMs separately on English and Hindi datasets, with the results summarized in Table [5.16](#)
2. Using Bilingual Combined Training, where the model is fine-tuned on a combined dataset of both English and Hindi question-answer pairs. The results of this bilingual training are presented in Table [5.17](#) and [5.18](#)

For the evaluation of SLLMs’ performance in Hindi math reasoning tasks, we focused exclusively on the Hindi version of the IndiMathQA and HAWP datasets. For clarity, we refer to the Hindi version of IndiMathQA as HMKB and the English version as EMKB. (Table [5.16](#))

#### 5.3.4.8 Outcomes & Insights

In this section, we begin by analyzing the effects of Curriculum Learning-based fine-tuning on Hindi and English datasets individually. Following that, we explore results from combined bilingual training. Lastly, we compare the performance of lightweight open-source models (SLLMs) against closed-source models (LLMs) across multiple languages and difficulty levels.

**Curriculum Learning - English Training:** We assess the impact of Curriculum Learning on the English dataset in improving the performance of SLLMs. Initially, the models were fine-tuned on the complete English dataset without any differentiation in problem complexity. In this baseline configuration, WizardMath-7B outperformed all others, while LLeMMA-7B recorded the lowest performance across benchmarks, including our English dataset, EMKB, as shown in Table [5.16](#). Subsequently, the models

Table 5.16: Performance Comparison of Open-Source and Closed-Source Models on English and Hindi Mathematical Benchmarks

Models	Settings	English Benchmarks						Hindi Benchmarks			
		GSM8K	MATH	PRM800K	EMKB			Enhanced	HMKB		
					Easy	Medium	Hard		HAWP	Easy	Medium
LLaMA-7B	Base	33%	22%	27%	36%	28%	21%	19%	17%	11%	8%
LLeMMA-7B	Base	14%	10%	12%	14%	12%	9%	12%	11%	8%	5%
Mistral-7B	Base	37%	23%	29%	39%	30%	24%	25%	22%	14%	10%
MAmmoTH-7B	Base	24%	14%	19%	27%	13%	11%	30%	27%	22%	18%
WizardMath-7B	Base	71%	36%	40%	64%	48%	44%	68%	61%	46%	36%
LLaMA-7B	[SFT_easy]	39%	25%	28%	40%	29%	21%	24%	22%	12%	8%
LLeMMA-7B	[SFT_easy]	21%	11%	12%	21%	13%	9%	15%	14%	9%	5%
Mistral-7B	[SFT_easy]	43%	25%	31%	45%	32%	24%	30%	27%	15%	10%
MAmmoTH-7B	[SFT_easy]	30%	16%	22%	33%	14%	13%	41%	37%	23%	18%
WizardMath-7B	[SFT_easy]	79%	37%	42%	70%	52%	44%	73%	66%	47%	37%
LLaMA-7B	[SFT_easy+medium]	42%	35%	34%	41%	36%	24%	25%	25%	20%	16%
LLeMMA-7B	[SFT_easy+medium]	21%	18%	19%	21%	18%	12%	15%	15%	11%	10%
Mistral-7B	[SFT_easy+medium]	45%	37%	34%	46%	39%	26%	31%	29%	28%	22%
MAmmoTH-7B	[SFT_easy+medium]	33%	25%	30%	34%	21%	15%	42%	40%	32%	26%
WizardMath-7B	[SFT_easy+medium]	<b>80%</b>	<b>45%</b>	<b>44%</b>	<b>73%</b>	<b>64%</b>	<b>46%</b>	<b>77%</b>	<b>69%</b>	<b>52%</b>	<b>42%</b>
<i>Bilingual Model Evaluation</i>											
OpenHathi-7B	Base	33%	19%	24%	36%	26%	20%	50%	32%	28%	24%
Airavata-7B	Base	22%	11%	15%	21%	12%	9%	12%	14%	10%	6%
Hi-NOLIN-9B	Base	31%	16%	22%	30%	21%	16%	45%	30%	26%	24%
<i>Closed Source Models</i>											
Gemini 1.0 Pro	Base	75%	39%	38%	68%	60%	43%	81%	72%	60%	48%
GPT-4	Base	91%	57%	70%	92%	90%	81%	93%	91%	83%	70%

were fine-tuned on a subset of easy problems (SFT\_easy), resulting in a 4-6% improvement on easy problems and a 6-8% increase on the GSM8K benchmark, indicating effective learning of simpler questions. However, there were modest gains of only 1-2% on more challenging benchmarks like MATH and PRM800K.

In the next phase, the models were fine-tuned on both easy and medium problems (SFT\_easy+medium). This strategy led to a consistent 6% improvement on medium problems and a 3% increase on hard problems. These results (Table 5.16) demonstrate that gradually increasing problem difficulty allows models to exceed their baseline performance.

**Curriculum Learning - Hindi Training:** When applying Curriculum Learning to the Hindi datasets, WizardMath-7B initially led, with LLeMMA-7B trailing behind on the Enhanced HAWP Benchmark. Fine-tuning on easy problems (SFT\_easy) resulted in a 3-5% improvement, though performance gains on medium and hard problems were limited. Implementing Curriculum Learning with both easy and medium problems

(SFT\_easy+medium) further enhanced performance by 2-4% on the benchmark and by 3-5% on more difficult problems (Table 5.16). This stepwise training approach improved the models' ability to handle progressively complex tasks, confirming the value of structured learning in Hindi datasets.

**Curriculum Learning - Bilingual Combined Training:** Next, we evaluated the performance of SLLMs on the complete IndiMathQA dataset, encompassing both Hindi and English versions (Tables 5.17 and 5.18). SLLMs were fine-tuned using Curriculum Learning on a combined bilingual training set. Across the board, models trained in a bilingual setting (Tables 5.17 and 5.18) outperformed those fine-tuned in a single language (Table 5.16), particularly on Hindi benchmarks. This supports our hypothesis that combined fine-tuning in English and Hindi improves a model's Hindi mathematical reasoning.

Initially, WizardMath-7B achieved the best results, while Airavata-7B scored the lowest (Base Settings: Table 5.16). Fine-tuning on easy problems (SFT\_Easy: Table 5.17) in both languages consistently improved performance by 3-5% on easy questions, showcasing the models' ability to generalize across linguistic contexts. However, improvements on medium and hard problems were minimal, illustrating the limitations of focusing primarily on easy problems.

When fine-tuned on both easy and medium problems in both languages (SFT\_Easy + Medium: Table 5.18), models exhibited substantial improvements, with medium problems improving by 11-18% and hard problems by around 2%. This emphasizes the effectiveness of Curriculum Learning in boosting problem-solving abilities through bilingual training.

### **Fine-Tuning Open-Source Models (SLLMs)**

We evaluated open-source models such as LLaMA-7B, LLeMMA-7B, Mistral-7B, MAmmoTH-7B, and WizardMath-7B when fine-tuned on the combined Hindi and English versions of IndiMathQA (HMKB and EMKB) (Tables 5.17 and 5.18). Fine-tuning across both languages yielded significant improvements in both languages compared to single-language fine-tuning. As demonstrated in Table 5.17 and 5.18, fine-tuning on easy problems from both languages resulted in a 2-3% improvement on easy problems in both Hindi and English. This improvement, while modest, exceeds the pre-trained model's performance but is less pronounced than the gains seen with single-

language fine-tuning (Table 5.16). Nevertheless, fine-tuning on easy and medium problems from both languages led to a major improvement of 11-18% accuracy, as illustrated in Table 5.17 and 5.18.

### **SLLMs (Lightweight open-source) vs LLMs (closed-source)**

WizardMath-7B emerged as the top-performing SLLM in our study. Although GPT-4 outperformed even the enhanced Wizard Math-7B (Tables 5.16, 5.17, and 5.18), through Curriculum Learning (SFT\_easy+medium) and bilingual parallel training, WizardMath-7B surpassed Gemini 1.0 Pro by approximately 5% on English datasets (Tables 5.16, 5.17, and 5.18). This highlights the efficacy of our approach in improving SLLMs' problem-solving capabilities in English. However, for Hindi datasets, while WizardMath-7B's performance was comparable to Gemini Pro, it lagged by about 3% on medium and hard problems, likely due to WizardMath's stronger proficiency in solving English-language mathematical problems compared to Hindi.

**English Models vs Bilingual Models** In this section, we compare the performance of bilingual models with other open-source models, as shown in Tables 5.16, 5.17, and 5.18. The results reveal that bilingual models demonstrate a consistently higher performance across both English and Hindi, which is an exception among open-source models, aside from WizardMath-7B. This consistent performance can be attributed to the language-agnostic nature of mathematical reasoning. However, it is noteworthy that bilingual models like OpenHathi-7B, which were not pre-trained specifically on mathematical tasks, display only marginal improvements after fine-tuning. This suggests that their learning efficiency in these tasks is somewhat constrained. On the other hand, the standout performance of WizardMath-7B emphasizes the critical role of pre-training on mathematical datasets to achieve strong cross-lingual performance in mathematical reasoning tasks.

## **5.4 Conclusion and Future Scope**

This section presented three datasets, GPSM4K, GeoVQA, and MathQuest, designed to challenge and evaluate the capabilities of modern LLMs and Multi-Modal LLMs in handling complex geometric and mathematical queries. The GPSM4K and GeoVQA datasets, both derived from grades 7-12 mathematics textbooks, include numerical-

Table 5.17: Performance of Bilingual Models on IndiMathQA Using SFT\_easy Training

Models	IndiMathQA					
	EMKB			HMKB		
	Easy	Medium	Hard	Easy	Medium	Hard
LLaMA-7B	43%	31%	22%	25%	16%	13%
Llemma-7B	20%	12%	9%	14%	10%	7%
Mistral-7B	48%	33%	24%	30%	21%	16%
Mammoth-7B	36%	15%	13%	40%	28%	23%
WizardMath-7B	<b>73%</b>	<b>64%</b>	<b>44%</b>	<b>68%</b>	<b>46%</b>	<b>38%</b>
<i>Bilingual Models</i>						
OpenHathi-7B	41%	30%	23%	36%	34%	26%
Airavata	23%	14%	11%	16%	11%	9%
Hi-NOLIN	38%	27%	25%	33%	32%	25%

Table 5.18: Performance of Bilingual Models on IndiMathQA Using SFT\_easy+medium Training

Models	IndiMathQA					
	EMKB			HMKB		
	Easy	Medium	Hard	Easy	Medium	Hard
LLaMA-7B	44%	35%	23%	29%	24%	19%
Llemma-7B	21%	14%	8%	15%	10%	9%
Mistral-7B	50%	39%	29%	33%	26%	22%
Mammoth-7B	40%	20%	18%	44%	35%	27%
WizardMath-7B	<b>75%</b>	<b>66%</b>	<b>47%</b>	<b>72%</b>	<b>57%</b>	<b>45%</b>
<i>Bilingual Models</i>						
OpenHathi-7B	43%	33%	24%	40%	37%	31%
Airavata	25%	16%	13%	18%	14%	11%
Hi-NOLIN	40%	30%	21%	39%	35%	28%

answer and theorem-proving questions aimed at enhancing model performance in geometry-based visual question answering. Our findings indicate a performance gap that underscores the need for further improvements, particularly in how models incorporate theorem-based knowledge and stepwise solution generation.

Through our baseline analysis using models such as GPT-4 and Gemini-pro, we observed that stepwise prompting and zero-shot, as well as few-shot inference techniques, are effective, with GPT-4 showing superiority in zero-shot inference and Gemini-pro excelling in few-shot dynamic inference. Furthermore, the introduction of MathQuest’s tailored question-answer pairs that include single or multiple mathematical operators and expressions has demonstrated significant enhancements in solution precision and

comprehensibility, which are critical for educational applications.

By making these datasets publicly available, we aim to empower researchers to develop more robust and versatile models capable of tackling intricate geometric and mathematical concepts. Future research should focus on refining these datasets to include more theorem-rich content and on developing advanced techniques for generating captions that encapsulate more detailed geometric information. Such efforts are likely to bridge the current gaps between LLMs and Vision Models and enhance the overall efficacy of models in complex problem-solving scenarios. Ultimately, our combined efforts contribute to advancing conceptual understanding and problem-solving in high school-level mathematical question-answering, offering valuable assistance to students and professionals grappling with challenging questions.

# CHAPTER 6

## Grammar Correction

### 6.1 Motivation

Effective grammar correction is pivotal to the readability and precision of educational and scientific writing, where even small lapses can obscure meaning or undermine credibility. Traditional neural-machine-translation (NMT)-based GEC systems rely on vast parallel corpora yet still falter when errors hinge on information scattered across sentences or paragraphs. Their sentence-bounded perspective cannot resolve tense drift, dangling references, or terminology that evolves over the course of a document.

Transformer-based models have advanced the field, but most remain optimised for isolated sentence edits. Widely used grammar-checking plug-ins embrace the same paradigm: they polish one line at a time, delivering quick fixes that can inadvertently fracture a narrative’s flow. **Document-level context is therefore the missing piece.** Scholarly prose builds cohesion through cross-sentence cues—tense consistency, referential clarity, disciplinary jargon—and those cues lie beyond the reach of sentence-only systems. A truly effective corrector must revise each sentence with awareness of what came before and what follows, preserving stylistic unity while eliminating discourse-level errors.

This thesis meets that need through the Dynamic Context Learner (DCL-GEC), a model that dynamically selects and incorporates relevant context as it edits. By coupling fine-grained error identification with a sliding contextual window, DCL-GEC delivers coherent, paragraph-spanning revisions without sacrificing sentence-level accuracy. In doing so, it pushes grammar correction toward the level of fluency and consistency demanded in modern educational and scientific communication.

## 6.2 Introduction

The GEC systems aim at improving text readability and comprehensibility through error correction. The traditional methods of GEC used the NMT approach as seen in [Sennrich \*et al.\* \(2016\)](#). However, this has drawbacks such as that of huge amounts of training data required, and it can not recognize errors. Language understanding is handled by the encoder, while generation is left to the decoder in an NMT-based system. However, generating language is harder than processing language, so these systems process through longer periods of time.

Significant advancements in grammar error correction (GEC) have recently been achieved using transformer-based sequence-to-sequence (seq2seq) models, as demonstrated by prominent GEC benchmarks [Bryant \*et al.\* \(2019\)](#). These transformer models, initially presented in [Vaswani \*et al.\* \(2017\)](#), are trained to enhance grammatically incorrect sentences, making GEC a practical application of seq2seq transformations.

The GECToR model takes a novel approach by reframing error correction as a classification task instead of a generation task. In this framework, each token in the input sequence is assigned a specific correction rule to generate the correct form, rather than generating an entirely new sequence. This classification-based method improves inference speed, reduces reliance on extensive training data, and enhances model interpretability. However, it is limited in addressing complex, context-sensitive errors that involve multiple tokens or intricate linguistic nuances.

Previous work in grammatical error correction (GEC) using language models like GPT often focused on correcting single, isolated sentences without considering contextual clues from surrounding text. Simple GEC models using a fixed context window can fail to capture these wider connections. In contrast, human proofreaders, as discussed in [Yuan and Bryant \(2021b\)](#), incorporate the document’s overall context to effectively identify and correct errors related to tense, verb choice, definite articles, and connectives.

We propose that incorporating contextual information directly into the model input can enhance its performance. Traditional language models struggle with maintaining coherence across longer sequences, necessitating the processing of dense information in smaller sections or chunks.

We can improve the accuracy of grammatical error detection by dividing lengthy

paragraphs into smaller, manageable segments while preserving essential context. Previous research has explored static context addition techniques, often by including one or two adjacent sentences around the target sentence. However, these static methods may not be optimal for managing context, as they can lead to information loss and failure to adapt to varying contextual needs across different cases.

We hypothesize that context plays a pivotal role in GEC tasks, which is effectively leveraged by GEC-DCL [Anand et al. \(2023e\)](#) to achieve superior performance. Our model demonstrated notable improvements, with  $F_{0.5}$  scores increasing by 77%, 19.61%, and 10.49% on the CoNLL-2014, BEA-Dev, and FCE-Test datasets, respectively. Additionally, it surpassed existing state-of-the-art models. The CoNLL-2014 dataset includes 26 sentences, BEA-Dev contains 13, and FCE-Test has 14. Our model, by effectively doubling the context size relative to CoNLL-2014, achieved higher performance, supporting our hypothesis on the importance of context in GEC.

The main contributions of this work can be summarized as follows:

1. We introduce a new model, the DCL, designed to dynamically select a relevant context from the corpus for GEC.
2. We propose a novel GEC architecture, GEC-DCL, based on a Transformer model capable of correcting errors in individual sentences and entire paragraphs within a document.
3. Our model was evaluated on standard GEC datasets, achieving SOTA results, and was benchmarked against LLaMA. Additionally, we developed synthetic datasets, XSUM and CNN, to further train and assess model performance.
4. We have created a scientific dataset: GEC S2ORC using the original S2ORC dataset [Lo et al. \(2020\)](#) and showed our model performance which is SOTA for scientific grammar error correction.

## 6.3 Experiments

### 6.3.1 Datasets

#### 6.3.1.1 CNN/DailyMail

The CNN/DailyMail dataset serves as another major resource for text summarization, widely used in NLP research. With approximately 312,000 news articles averaging 781 words each, the dataset includes both full articles and summaries. It provides a

substantial data volume for model training and evaluation, with standardized constraints on article and summary lengths and an established data split.

#### **6.3.1.2 S2ORC**

The S2ORC dataset contains an extensive repository of academic papers across fields such as computer science, medicine, and social sciences, with over 81 million abstracts included. Collected from the Semantic Scholar Corpus, this dataset provides rich metadata, including titles, author information, abstracts, publication sources, citation data, and references, making it a comprehensive resource for academic research and the exploration of scholarly trends.

#### **6.3.1.3 FCE-Test**

The FCE-Test dataset offers 1,326 writing samples from English learners of various skill levels. Alongside each writing sample, metadata such as the test-takers age, gender, and nationality is provided, with detailed annotations on aspects like grammar, coherence, and structure. These features make it a valuable tool for assessing written English proficiency among non-native speakers.

#### **6.3.1.4 XSUM**

The XSUM dataset is a widely recognized benchmark for abstractive summarization tasks, containing approximately 226,711 news articles sourced from the BBC. Each article is paired with a headline and a brief summary, with input lengths capped at 512 tokens and target summaries limited to 72 tokens. The dataset is divided into 204,045 training, 11,000 validation, and 5,503 test samples, providing a robust foundation for training and evaluating summarization models.

#### **6.3.1.5 BEA-Dev**

The BEA-Dev dataset is extensively used in grammatical error correction (GEC) evaluation, featuring 1,456 sentences and 17,434 tokens annotated for error types at three levels: morphological, syntactic, and lexical. With 11 distinct error categories, it allows for detailed analysis and correction of common language issues in texts written by

non-native English speakers.

Table 6.1: Statistics of the datasets that were used in the research

Parameters	XSum	CNN	S2ORC
Avg. no. of characters	428	925	1039
Avg. no. of words	69	138	139
Avg. no. of unique words	58	109	98
Avg. no. of punctuation marks	7	18	19
Avg. no. of sentences	3	2	6
Sum	756057	1308093	3269096
Max no. of sentences	25	26	346
Min no. of sentences	1	1	1
Standard Deviation	1.662	1.715	1.832
Train Set	183896	1139187	490000
Test Set	1000	1000	1000
Dev Set	19432	22248	9000

### 6.3.1.6 CoNLL-2014

The CoNLL-2014 Shared Task dataset is a well-regarded benchmark for tasks including Joint Named Entity Recognition (NER), Coreference Resolution (CR), and Semantic Role Labeling (SRL). It contains English news articles annotated with rich linguistic details, such as named entities, coreference chains, and semantic roles. Spanning roughly 1.2 million words across 29,669 instances, this dataset is divided into training, development, and test sets, providing a robust resource for multi-faceted linguistic analysis.

To enable our model to learn error correction from context, we constructed a new dataset by incorporating synthetic errors into pre-existing datasets that closely mirror common human grammatical errors. Specifically, we selected the CNN/DailyMail [Nal-lapati et al. \(2016\)](#) and XSUM [Narayan et al. \(2018\)](#) datasets because they provide extensive sequences of context-rich language, facilitating model adaptation to new data inputs.

We utilized the CNN/DailyMail and XSUM datasets, which feature lengthy sequences of sentences and paragraphs. These datasets were instrumental in helping the

model understand the broader context of sentences, enabling our system to identify and group related phrases. The aggregated statistics for these datasets are provided in Table 6.1

For the error generation process, we used a custom script to modify 1,000-character paragraphs extracted from the datasets. Errors were introduced, generating both grammatically correct and erroneous paragraphs.

In addition, we leveraged the S2ORC dataset (Lo et al. (2020)) to further enhance our model’s performance in handling scientific texts. This dataset allowed us to more effectively model context in a scientific setting. Using our error generation script, we created the GEC\_S2ORC-train, GEC\_S2ORC-dev, and GEC\_S2ORC-test datasets by introducing errors into the S2ORC data.

The original articles and paragraphs from the datasets mentioned above served as the grammatically correct corpus. To create the dataset of erroneous sentences, we introduced a variety of errors using custom-developed tools. These tools were designed to simulate common human errors, such as the unnecessary addition or omission of pronouns, determiners, and prepositions, as illustrated in Table 6.2.

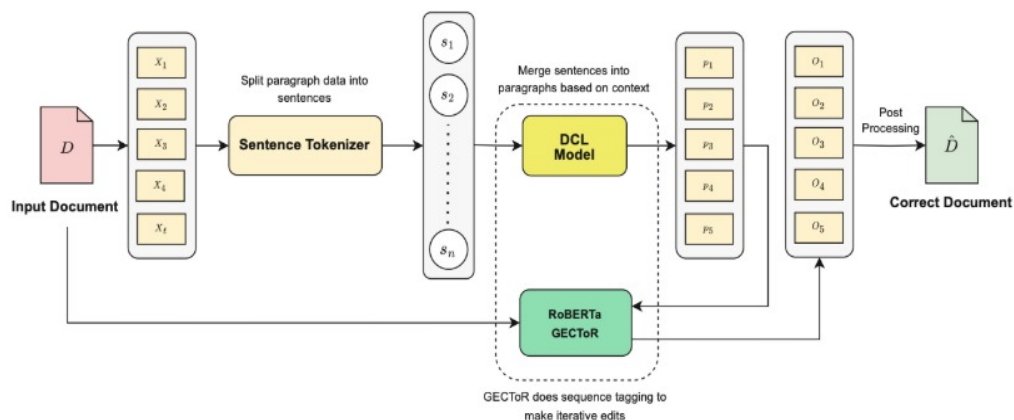


Figure 6.1: The DCL-GEC Architecture is composed of two main components: (1) a *Sentence Tokenizer*, which divides the input paragraph into individual sentences; and (2) the *DCL Model*, which reassembles these sentences into contextually coherent paragraphs. Once restructured, the *GECtoR* model applies sequence tagging to identify and implement corrections, producing an edited and grammatically accurate paragraph.

## 6.3.2 Methodology

This section outlines the requirements essential to the research goals, detailing the key components involved. It explains the foundational structure and framework that support the study’s objectives. Furthermore, the pipeline—a series of interconnected processes—is described to provide clarity on data flow across different stages, enabling a systematic and structured approach to data analysis.

### 6.3.2.1 GEC-DCL

Table 6.2: Types of Error and Their Distribution in the Error Script

Type of Error	Probability	Original	After Infection
<b>Token-level</b>	0.167		
Swapping random character	0.1	This sentence demonstrates our initial script	This sentence demonstrates our initial script
Infect word using "py-inflect"	0.1	This sentence demonstrates our initial script	This sentence demonstrates <b>demonstrate</b> our initial script
Insert Space	0.1	This sentence demonstrates our initial script	This sentence demonstrates <b>&lt;space&gt; our</b> initial script
Swap two tokens	0.1	This sentence demonstrates our initial script	This sentence <b>our demon-</b> <b>strates</b> initial script
Change case of token	0.1	This sentence demonstrates our initial script	This sentence <b>Demonstrates</b> our initial script
No Change in token	0.5	This sentence demonstrates our initial script	This sentence demonstrates our initial script
<b>Sentence Level</b>	1		
Remove Pronouns	0.167	This sentence demonstrates our initial script	<b>This</b> sentence demonstrates our initial script
Remove Determiners	0.167	This sentence demonstrates our initial script	This sentence demonstrates <b>our</b> initial script
Remove Preposition	0.167	This sentence demonstrates our initial script	This is <b>my</b> sentence demonstrates initial script
Adding 'the' before Verb and Nouns	0.167	This sentence demonstrates our initial script	This sentence demonstrates our initial <b>the</b> script
Replacing 'because' with 'Because'	0.167	Passengers were delayed because of icy roads	Passengers were delayed <b>Because</b> of icy roads

The DCL-GEC model is structured around two core components: the DCL and a Grammar Error Correction (GEC) model. The GEC model, a transformer-based sequence tagger inspired by GECToR, undergoes a two-stage training process. Initially, it is

trained on error-laden corpora, followed by training on a mixed dataset of both error-prone and clean parallel corpora. Furthermore, the transformer encoder is enhanced with synthetic data. To correct input tokens, the GEC model applies a specialized token-level transformation known as g-transformation, aligning input tokens with their target corrections.

For baseline comparison, we also used the GECToR model without the DCL component. The performance of GECToR alone is compared with that of our integrated model, particularly in the context of paragraphs, demonstrating that our approach yields superior results. In the architecture of the DCL-GEC model, the GEC model processes inputs generated by the DCL model, which comprises a set of sentences sharing a common context. This enables the detection and correction of grammatical errors while preserving the context of the sentences. The outcome is a grammatically error-free corpus, which is generated by focusing on the shared context among the sentences.

To formalize this approach, we begin by extracting paragraphs from a document  $D$  such that  $D = \{X_1, X_2, X_3, \dots, X_t\}$ , where  $t$  denotes the number of paragraphs. Each paragraph  $X \in D$  is processed by the DCL-GEC model, which further divides it into sentences represented by  $S = \{s_1, s_2, s_3, \dots, s_n\}$ . The DCL model then generates combinations of adjacent sentences from  $S$ . For example,  $s_1$  and  $s_2$  are processed together to determine if they belong to the same context. This process is repeated for pairs such as  $s_2$  and  $s_3$ ,  $s_3$  and  $s_4$ , and so forth. Sentences identified as sharing the same context are grouped into paragraphs  $P = \{p_1, p_2, p_3, \dots, p_k\}$ , where, for instance,  $p_1 = \{s_1, s_2\}$  and  $p_2 = \{s_3\}$ .

Once the paragraph set  $P$  is created for the contextually connected sentences within document sections  $X_i$ , each paragraph  $p_i$  along with its original counterpart  $X_i$  is fed into the GECToR model, which is based on RoBERTa. In this process, each  $p_i$  is processed iteratively four times, whereas the original paragraph  $X_i$  is only processed once. The GECToR model employs its sequence tagging approach during each iteration to refine the text while preserving the coherence of the document. After the DCL model constructs the contextual framework, the GEC model applies these corrections iteratively to improve comprehension and ensure the original intent is maintained. The corrected sentences are subsequently compiled into updated paragraphs  $X'$  corresponding to each original paragraph  $X$  in  $D$ . Ultimately, the revised paragraphs are consolidated to create the fully corrected document  $D$ , as demonstrated in Figure [6.1](#)

Table 6.3: Our DCL-GEC model is evaluated against two (SOTA) models on sentence variations in the FCE-Test, BEA-Dev, and CoNLL-2014 datasets. Unlike the comparison models, which rely on a fixed number of contextual sentences, our model dynamically adjusts context, leading to superior performance across these benchmarks.

Model		FCE-Test			BEA-Dev			CoNLL-2014		
		P	R	F <sub>0.5</sub>	P	R	F <sub>0.5</sub>	P	R	F <sub>0.5</sub>
Yuan et al	sent	<b>65.36</b>	<b>44.17</b>	<b>59.64</b>	62.64	<b>40.72</b>	56.55	64.57	28.65	51.62
Chollampatt et al	sent	53.91	32.81	47.77	-	-	-	64.32	35.98	55.57
DCL-GEC	sent	-	-	-	<b>68.72</b>	33.97	<b>57.04</b>	<b>76.48</b>	<b>40.66</b>	<b>65.02</b>

Table 6.4: GECToR is compared with the DCL-GEC model. We compare results for sentences (sent) and paragraphs (para) variation of FCE-Test, BEA-Dev, and CoNLL-2014, where we achieved better results on the paragraph variant and were at par for the sentence variant.

Model		FCE-Test			BEA-Dev			CoNLL-2014		
		P	R	F <sub>0.5</sub>	P	R	F <sub>0.5</sub>	P	R	F <sub>0.5</sub>
GECToR	sent	-	-	-	66.0	33.8	55.5	77.5	40.2	<b>65.3</b>
	para	<b>64.12</b>	17.91	42.30	66.71	19.24	44.67	<b>70.00</b>	06.17	22.80
DCL-GEC	sent	-	-	-	68.72	33.97	57.04	76.48	40.66	65.02
	para	50.28	<b>37.06</b>	<b>46.93</b>	65.30	<b>30.93</b>	<b>53.43</b>	56.12	<b>18.26</b>	<b>39.67</b>

### 6.3.2.2 SciBERT

SciBERT, a pre-trained language model derived from BERT, has been developed using a large corpus consisting of 1.14 million articles from Semantic Scholar. This model is specifically designed to better capture the semantics of technical writing, particularly in scientific contexts, as reflected by its application in Semantic Scholar.

In our model, we utilize SciBERT in place of RoBERTa as the encoder within GECToR, enabling the model to handle scientific data more effectively. SciBERT has a 40% difference in vocabulary compared to BERT, with a greater focus on scientific terms. To optimize its performance for grammatical error correction in scientific writing, we fine-tuned the SciBERT model on the GEC\_S2ORC-train dataset.

### 6.3.3 GECToR-RoBERTa & Model Training

We employ the pre-trained GECToR-RoBERTa model as our baseline encoder, which utilizes a three-stage training process on the National University of Singapore Corpus of Learner English (NUCLE) [Dahlmeier et al. \(2013\)](#), Lang-8 Corpus of Learner English (Lang-8) [Tajiri et al. \(2012\)](#), the FCE dataset [Yannakoudakis et al. \(2011\)](#), and the Write & Improve + LOCNESS Corpus [Bryant et al. \(2019\)](#). To further enhance our model, we fine-tuned it using synthetic data from the XSUM and CNN datasets to leverage the

Table 6.5: Comparison of Random/fixed split vs DCL on CoNLL-2014 having paragraphs.

Method	Precision	Recall	F <sub>0.5</sub>
Prev 2 sent	33.38	14.90	26.74
Random context	32.72	13.65	25.57
DCL based split	<b>56.12</b>	<b>18.26</b>	<b>39.67</b>

Table 6.6: Gector vs DCL-GEC on XSUM Dataset

Model	Precision	Recall	F <sub>0.5</sub>
GECToR	69.43	59.36	67.15
DCL-GEC	<b>72.25</b>	<b>60.4</b>	<b>69.5</b>

rich contextual information present in these corpora. We conducted evaluations of our model using the M<sup>2</sup> Scorer [Dahlmeier and Ng \(2012\)](#) on the FCE test [Yannakoudakis et al. \(2011\)](#), CoNLL 2014 test [Ng et al. \(2014\)](#), and BEA-dev [Bryant et al. \(2019\)](#) datasets. Additionally, we used the ERRANT Scorer [Bryant et al. \(2017\)](#) for evaluating the XSUM-test and CNN-test sets.

### 6.3.4 Training

The optimization of our models was carried out using the Adam Optimizer with default parameters, adhering to the hyperparameter guidelines outlined in [Omelianchuk et al. \(2020\)](#), and incorporating a label smoothing factor of 0.1.

We conducted training on the GECToR RoBERTa model using the XSUM and CNN datasets, while GECToR SciBERT was trained on the GEC S2ORC dataset. For both models, a batch size of 20 and a maximum sequence length of 100 tokens were utilized over the course of 15 epochs, with a default learning rate of 1e-5. Additionally, the DCL model underwent training on the FCE and W&I+LOCNESS datasets, as well as the S2ORC dataset to enhance its capability for specialized scientific grammar correction. This training employed a batch size of 128, leveraging a pre-trained BERT model over 5 epochs, with a learning rate adjusted to 5e-6.

Table 6.7: GECToR vs DCL-GEC w/ SciBERT as Transformer on S2ORC Dataset

Model	Precision	Recall	F <sub>0.5</sub>
GECToR	<b>65.22</b>	29.31	52.38
DCL-GEC	62.33	<b>37.45</b>	<b>55.02</b>

Table 6.8: GECToR vs DCL-GEC on CNN Dataset

Model	Precision	Recall	F 0.5
GECToR	<b>80.55</b>	48.60	71.19
DCL-GEC	78.02	<b>55.91</b>	<b>72.30</b>

Table 6.9: This table presents a comparison of error type-specific performance between GECToR and DCL-GEC on the FCE-Test dataset, evaluated using the F<sub>0.5</sub> score. The last column highlights the performance difference between our proposed model and GECToR, demonstrating a significant improvement of 24% with the DCL-GEC model.

Error Type	GECToR			DCL-GEC			% Diff. F <sub>0.5</sub>
	P	R	F <sub>0.5</sub>	P	R	F <sub>0.5</sub>	
M:ADV	42.86	01.92	08.15	38.46	03.25	12.14	+48.95
M:NOUN	47.06	04.42	16.06	50.00	05.00	17.86	+11.21
M:NOUN:POSS	73.33	27.50	55.00	72.22	32.50	58.04	+5.5
M:VERB	50.00	10.51	28.54	54.88	16.30	37.25	+30.52
M:VERB:FORM	66.67	16.90	41.96	78.12	35.21	62.81	+49.69
M:VERB:TENSE	52.78	16.67	36.82	60.87	24.45	46.90	+27.38
M:ADJ	77.78	10.14	33.33	66.67	08.70	28.57	-14.28
R:ADJ:FORM	60.00	13.64	35.71	100.00	22.73	59.52	+50.00
R:NOUN	59.30	10.22	30.25	68.66	13.80	38.25	+26.45
R:NOUN:POSS	58.33	15.91	38.04	56.25	20.93	42.06	+10.57
R:VERB	54.58	08.25	25.71	62.20	09.97	30.37	+18.12
R:VERB:FORM	67.50	27.39	52.21	70.14	44.04	62.71	+20.11
R:VERB:TENSE	54.37	19.63	40.16	62.41	29.20	50.84	+26.59
R:ADJ	65.45	11.92	34.48	70.49	14.38	39.59	+14.82
U:ADV	22.41	6.50	15.05	25.30	10.55	19.77	+31.36
U:NOUN	21.21	04.27	11.82	40.00	04.88	16.39	+38.66
U:NOUN:POSS	47.06	38.10	44.94	57.89	52.38	56.70	+26.17
U:VERB	40.48	11.11	26.48	42.86	13.73	30.09	+13.63
U:VERB:FORM	68.75	21.57	47.83	82.35	27.45	58.82	+22.98
U:VERB:TENSE	49.37	22.54	39.88	55.74	39.31	51.44	+28.99
U:ADJ	18.18	03.23	09.43	20.00	04.84	12.13	+25.04
						<b>Total Diff</b>	<b>+24%</b>

### 6.3.5 Results

The current SOTA model, as discussed in [Yuan and Bryant \(2021b\)](#), uses a fixed context setup, incorporating two preceding sentences for CoNLL 2014 and one for the FCE-Test and BEA-dev datasets. Table [6.3](#) shows that our model, which employs dynamic context learning, surpasses this fixed-context method on both BEA-dev and CoNLL 2014.

In Table [6.5](#) we compare the performance of our dynamic approach using random context averaged across multiple seeds with the two-sentence fixed context. Our ap-

proach achieves a deeper understanding of the impact of previous sentences, leading to an approximate improvement of 13.4  $F_{0.5}$  points over the static two-sentence context approach of Yuan and Bryant (2021b). Further, as shown in Table 6.4, our model achieves improvements of 8.76, 4.63, and 16.87  $F_{0.5}$  points on paragraph-level versions of BEA-dev, FCE-Test, and CoNLL-2014, respectively, when compared to GECToR.

Table 6.9 reveals that our DCL model yields notable improvements in error types that benefit from a broader context, such as tense and verb usage. For the FCE-Test dataset, we observed an average enhancement of 24% across the three primary error types: missing (M), replacement (R), and unnecessary (U) errors.

Moreover, Table 6.4 contrasts the performance of GECToR and DCL-GEC in both sentence- and paragraph-level contexts. The results confirm that DCL enhances paragraph-level error corrections while preserving high accuracy on sentence-level corrections.

We also evaluated our approach on the XSUM and CNN datasets, which feature paragraphs up to 25 sentences long. Our SOTA results for scientific documents, summarized in Table 6.6, underscore the method’s efficacy in handling lengthy, context-rich text. Our SOTA results for scientific documents, summarized in Table 6.6 and Table 6.7, underscore the method’s efficacy in handling lengthy, context-rich text. The performance comparison with GECToR on the CNN dataset is detailed in Table 6.8.

### 6.3.6 Error Analysis

To further verify that incorporating broader contextual information improves error correction, we conducted an error analysis. Our findings, summarized in Table 6.9, indicate notable gains in correcting VERB: FORM (+49.69%), VERB TENSE (+27.38%), and NOUN (+38.66%) errors. The model shows the highest improvement margins for errors involving agreement, coreference, and tense within sentences.

## 6.4 Conclusion

This research demonstrates that dynamically detected contextual information outperforms static context approaches for Grammar Error Correction (GEC). By including relevant context with each input sentence, we achieve more accurate grammatical corrections. Our approach is suitable for document-level correction, allowing comprehen-

sive revisions of extended texts in one pass. Unlike conventional GEC methods, which often address errors within individual phrases without considering cross-sentence context, our model appends relevant preceding phrases, enabling context-aware grammar correction. The efficacy of our approach has been validated across various domains, including scientific and general datasets, where preserving extended context is essential. The model thus serves as a complete, end-to-end solution for document-level error correction.

# CHAPTER 7

## Citation and Text Generation

### 7.1 Motivation

In scientific writing, generating accurate citations is crucial to upholding academic integrity, ensuring proper attribution, and maintaining clarity in complex scholarly discourse. However, manual citation practices are not only time-consuming but also prone to inconsistency and errors, especially in multi-reference scenarios. Existing automated approaches often lack the capacity to capture the nuanced relationships between cited documents and the context within which they are referenced. This limitation is particularly challenging in domains where intricate conceptual links between works are fundamental to understanding.

This thesis introduces a multi-source citation text generation (M-CTG) framework to address these needs. By integrating knowledge graphs, keyphrase embeddings, and fine-tuned language models, our approach enhances the model's ability to understand and generate contextually relevant, high-quality citations. This research contributes to the fields of natural language generation and scientific communication by advancing AI capabilities in domain-specific tasks, ultimately promoting more precise, accessible, and coherent citation practices that align with the evolving needs of academic writing and scientific knowledge dissemination.

### 7.2 Introduction

The generation of text within the scientific domain, particularly for citation purposes, presents a unique challenge due to the need for a deep understanding of both the input text and the associated domain-specific knowledge. CTG is a specialized task in NLP that involves creating accurate and contextually relevant citation text, which links a source document to its referenced documents. This task necessitates the use of ML models to accurately summarize the connections between citing and cited documents.

It is essential to employ appropriate terminology and structure to express these relationships in a clear and concise manner. Recent advancements in transformer-based models have greatly enhanced the field of text generation, resulting in notable improvements in tasks such as CTG.

CTG has numerous practical applications, particularly in scientific writing and academic research. For instance, it can be integrated into writing assistants to help researchers generate citations that are not only accurate but also contextually relevant, thereby improving the quality of scientific literature. Additionally, CTG can be employed in educational settings to teach students the proper techniques for citing sources, and in literature review systems to assist researchers in quickly identifying key relationships between papers. Moreover, CTG models can help streamline the process of citation generation, ensuring consistency, accuracy, and adherence to citation styles, while also reducing the risk of plagiarism by promoting proper source attribution.

Despite its potential, CTG remains a challenging task, especially when dealing with the lengthy and complex nature of scientific texts. Early approaches to CTG primarily focused on generating single-sentence citations from the abstracts of papers. However, real-world scenarios often involve generating multi-reference citations within a single sentence or paragraph. To address these challenges, recent studies have proposed the use of LLMs and multi-source architectures that incorporate various types of embeddings, such as graph embeddings and keyphrase embeddings, to enhance the performance of CTG models.

In this thesis, we propose a novel multi-source citation text generation (M-CTG) framework that leverages multiple types of inputs, including graph embeddings, keyphrase embeddings, and abstracts of papers, to generate citation text. We have fine-tuned three LLMs—LLaMA, Alpaca, and Vicuna—specifically for the CTG task. Additionally, we introduce the CTG-S2ORC dataset, a new resource derived from the S2ORC dataset, which contains metadata and citation information for computer science papers. Our approach incorporates knowledge graphs created from the metadata of research papers to improve the contextual understanding of the relationships between citing and cited documents.

Our experiments demonstrate that incorporating these multi-source embeddings and knowledge graphs significantly improves the performance of CTG models. We achieve substantial gains in evaluation metrics such as METEOR and ROUGE-1, particularly

when using the Alpaca model on the CTG-S2ORC dataset. Furthermore, we explore various prompting techniques in LLMs, including Zero-Shot CoT, Few-Shot CoT, and Auto-CoT, and show that our M-CTG framework consistently outperforms baseline models in generating high-quality citation text.

In summary, our contributions include:

1. The development of a multi-source architecture (M-CTG) that integrates customized graph and keyphrase embeddings for improved CTG performance.
2. The creation of the CTG-S2ORC dataset, a valuable resource for CTG research.
3. A significant improvement in CTG performance, as evidenced by increased ROUGE-1 and METEOR scores.
4. The implementation of various prompting techniques in LLMs to further enhance the quality of generated citation text.

## 7.3 Experiment

### 7.3.1 KG-CTG: Citation Generation with Knowledge Graph-Enhanced LLMs

#### 7.3.1.1 Methodology

In this section, we outline the objective of generating citations, which involves creating citation text that takes into account both the source publication and the cited paper [Anand et al. \(2023c\)](#). This task leverages the advancements in LLMs, which have demonstrated significant effectiveness in various text generation tasks [Zhao et al. \(2023\)](#).

1. **Model Fine-Tuning:** We initiate our process by fine-tuning three selected LLMs: LLaMA [Touvron et al. \(2023b\)](#), Alpaca [Taori et al. \(2023\)](#), and Vicuna [Chiang et al. \(2023\)](#) specifically for citation generation. The fine-tuning procedure involves training these models on a subset of the S2ORC dataset that is tailored for citation generation, followed by an evaluation of their performance based on the quality and relevance of the generated citation text.
2. **Incorporation of Knowledge Graphs:** To enhance the model’s understanding of the context and relationships among research papers, we incorporate a knowledge graph [Koncel-Kedziorski et al. \(2019\)](#) sourced from existing literature into the input prompt. This knowledge graph provides structured insights into key concepts, entities, and their connections. Additionally, we investigate the benefits of combining knowledge graphs with large language models (LLMs) [Pan et al.](#)

(2023) and how this integration contributes to improved performance in citation generation tasks.

By employing this methodology, we seek to illustrate the efficacy of fine-tuned large language models (LLMs) in generating precise and contextually relevant citation text. Furthermore, we examine how the incorporation of knowledge graphs enhances the models' ability to understand intricate relationships and contextual subtleties between the source documents and the cited references.

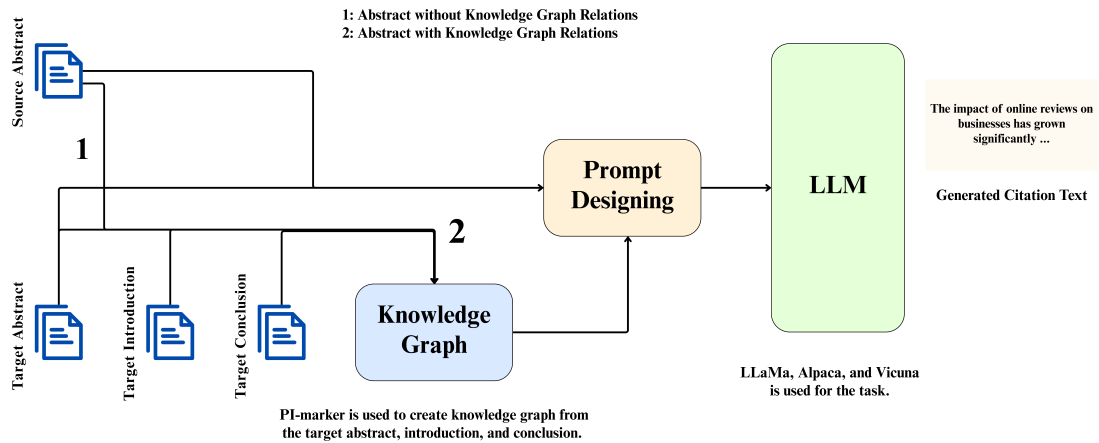


Figure 7.1: Workflow diagram illustrating the experimental setup, including two workflows: **1.** Prompt creation without knowledge graph relations, and **2.** Relation extraction from abstracts followed by prompt creation.

### 7.3.1.2 Large Language Models

In this section, we describe the LLMs employed in our research for generating citation texts. The models considered include LLaMA [Touvron et al. \(2023b\)](#), Alpaca [Taori et al. \(2023\)](#), and Vicuna [Chiang et al. \(2023\)](#). LLaMA features a transformer-based architecture with several parameter sizes, including 7B, 13B, 33B, and 65B. For this study, we focused on the LLaMA-7B variant. Alpaca, a fine-tuned version of LLaMA, has been trained on 52,000 instruction examples from OpenAI's text-DaVinci-003 model, which enhances its ability to generate instructional content effectively. Vicuna, another fine-tuned adaptation of LLaMA, has been trained on 70,000 user-contributed conversations collected from ShareGPT.com, enabling it to produce text that captures the conversational style and context characteristic of user interactions.

To assess the effectiveness of these models in CTG, we fine-tuned each one using our specific CTG dataset. This evaluation helped us identify the relative strengths and

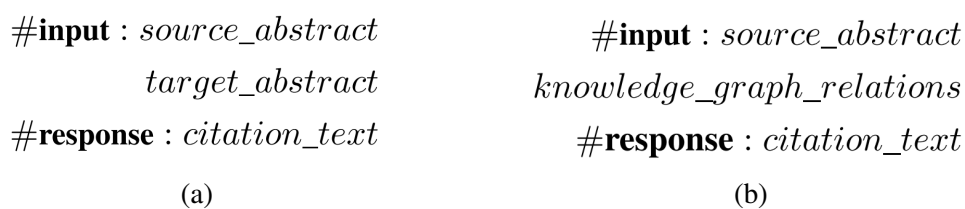


Figure 7.2: Structures of Prompts used for the Large Language Models.

limitations of each LLM in producing accurate and contextually appropriate citation text.

### 7.3.1.3 Knowledge Graphs and Prompting

A knowledge graph is a structured representation of information that organizes entities and their relationships, facilitating advanced data analysis and inference [Ji et al. (2022)]. In this thesis, we constructed knowledge graphs for both the source and target abstracts using PL-Marker, a SOTA tool [Ye et al. (2022)]. PL-Marker utilizes a novel packed levitated marker method that implements a packing strategy centered on the subject and its contextual surroundings to generate pair representations. This approach produces a knowledge graph that effectively encapsulates the relationships and context among different entities present in the abstracts of research papers. Figure 7.2 shows an example of knowledge graph generated using PL-Marker.

### 7.3.1.4 CTG Dataset

For this study, we utilized the S2ORC (Semantic Scholar Research Corpus) [Lo et al. (2019)], which encompasses a comprehensive collection of around 81.1 million academic papers in English spanning various disciplines. This corpus provides extensive details, including abstracts, full texts, bibliographic references, and metadata for each document.

To concentrate on the Computer Science field, we filtered the corpus to include only those papers categorized under "Field of Study" as Computer Science. This filtering process reduced the dataset from 81.1 million papers to roughly 6.0 million that are directly related to computer science. However, during the data-cleaning phase, we discovered that many papers in this subset either lacked valid abstracts or body text or contained irrelevant or empty content. Consequently, these were eliminated, resulting

in a final dataset comprising approximately 100,000 samples.

Table 7.1: Statistics of the dataset extracted from the S2ORC corpus.

Statistic	CTG-S2ORC	Train	Validation	Test
# citations	100,000	79,588	9,944	9,946
# unique papers	38,530	34,147	8,076	8,070
<b>CITATIONS</b>				
Avg # characters	171.3	171.34	171.47	170.87
Max # characters	3,420	2,986	1,840	2,170
<b>SOURCE ABSTRACTS</b>				
Avg # characters	1,225.89	1,185.74	1,189.3	1,190.75
Max # characters	56,771	8,945	7,858	8,024
<b>TARGET ABSTRACTS</b>				
Avg # characters	1,065.64	1,002.27	1,001.54	993
Max # characters	93,551	8,059	7,736	6,647

The "body text" of each paper comprises several paragraphs structured into sections, including Introduction, Methodology, and Conclusion, among others. Our analysis concentrated on detecting the presence of cite spans within these sections. Cite spans serve as dictionaries that offer citation details for the referenced papers within each paragraph. We extracted sentences containing these citations from the body text, omitting any citations that referenced multiple papers in a single sentence. Additional statistical information regarding the dataset is summarized in Table 7.4.

### 7.3.1.5 Experimental settings, Evaluation metrics, and Model comparisons

Here, we detail our experimental framework, the metrics used for evaluation, and the comparative performance analysis of the models. We fine-tuned and assessed three LLMs for text generation using our citation generation (CTG) dataset. The results of this analysis are summarized in Table 7.2. A comparison of LLaMA (Touvron *et al.* (2023b)), Alpaca (Taori *et al.* (2023)), and Vicuna (Chiang *et al.* (2023)) reveals that Vicuna outperforms the other models.

Subsequently, we improved the models by integrating knowledge graphs derived from both the source and target papers using PL-Marker (Ye *et al.* (2022)). This integration led to a notable enhancement in the effectiveness and quality of the generated citation text. Specifically, Alpaca demonstrated significant improvement, with a 33.14% increase in the METEOR score and a 36.98% rise in the Rouge-1 score. The outcomes from this configuration are presented in Table 7.3. These findings suggest that the integration of knowledge graphs is essential for enhancing the performance of LLMs in text-generation tasks.

#### Experimental Settings:

---

---

**source\_abstract**

Recovering missing values plays a significant role in time series tasks in practical applications. How to replace the missing data and build the dependency relations from the incomplete sample set is still a challenge. The previous research has found that residual network (ResNet) helps to form a deep network and cope with degradation problem by shortcut connection. Gated recurrent unit (GRU) can improve network model and reduce training parameters by update gate which takes the place of forgetting gate and output gate in long short-term memory (LSTM). Inspired by this finding, we observe that shortcut connection and mean of global revealed information can model the relationship among missing items, the previous and overall revealed information. Hence, we design an imputation network with decay factor for shortcut connection and mean of the global revealed information in GRU, called decay residual mean imputation GRU (DRMI-GRU). We introduce a decay residual mean unit (DRMU), which takes full advantage of the previous and global revealed information to model incomplete time series; and the decay factor is applied to balance the previous long-term dependencies and all non-missing values in the sample set. In addition, a mask unit is designed to check the missing data existing or not. An extensive body of empirical comparisons with other existing imputation algorithms over real-world data and public dataset with different ratio of missing data verifies the performance of our model.

**target\_abstract**

In this work we propose a novel interpretation of residual networks showing that they can be seen as a collection of many paths of differing length. Moreover, residual networks seem to enable very deep networks by leveraging only the short paths during training. To support this observation, we rewrite residual networks as an explicit collection of paths. Unlike traditional models, paths through residual networks vary in length. Further, a lesion study reveals that these paths show ensemble-like behavior in the sense that they do not strongly depend on each other. Finally, and most surprising, most paths are shorter than one might expect, and only the short paths are needed during training, as longer paths do not contribute any gradient. For example, most of the gradient in a residual network with 110 layers comes from paths that are only 10-34 layers deep. Our results reveal one of the key characteristics that seem to enable the training of very deep networks: Residual networks avoid the vanishing gradient problem by introducing short paths which can carry gradient throughout the extent of very deep networks.

---

**target\_citation**

**REF** propose an enlightening analysis of the Residual Networks (ResNets), and suppose that ResNet **REF** can be looked upon as a set of shallow networks and helps to form a deep network and copes with degradation problem by shortcut connection.

---

**predicted**

Recovering missing values plays a significant role in time series tasks in practical applications. How to replace the missing data and build the dependency relations from the incomplete sample set is still a challenge. The previous research has found that residual network (ResNet) helps to form a deep network and cope with degradation problem by shortcut connection. Gated recurrent unit (GRU) can improve network model and reduce training parameters by update gate.

---

---

Figure 7.3: Example of Generated Citation text from **Vicuna** without knowledge graph relations

In this study, the CTG dataset was partitioned into 79,588 training samples, 9,946 testing samples, and 9,944 validation samples. To fine-tune the LLMs, we implemented QLoRA [Dettmers et al. \(2023\)](#), a method designed to minimize GPU resource consumption. By facilitating backpropagation through a frozen, 4-bit quantized pre-trained language model into LoR), QLoRA optimizes memory usage. For optimization, we utilized the AdamW optimizer [Kingma and Ba \(2014a\)](#) alongside a Linear Scheduler. The learning rate was established at  $3e-4$ , incorporating 100 warmup steps to gradually ad-

---

---

**source\_abstract**

Recovering missing values plays a significant role in time series tasks in practical applications. How to replace the missing data and build the dependency relations from the incomplete sample set is still a challenge. The previous research has found that residual network (ResNet) helps to form a deep network and cope with degradation problem by shortcut connection. Gated recurrent unit (GRU) can improve network model and reduce training parameters by update gate which takes the place of forgetting gate and output gate in long short-term memory (LSTM). Inspired by this finding, we observe that shortcut connection and mean of global revealed information can model the relationship among missing items, the previous and overall revealed information. Hence, we design an imputation network with decay factor for shortcut connection and mean of the global revealed information in GRU, called decay residual mean imputation GRU (DRMI-GRU). We introduce a decay residual mean unit (DRMU), which takes full advantage of the previous and global revealed information to model incomplete time series; and the decay factor is applied to balance the previous long-term dependencies and all non-missing values in the sample set. In addition, a mask unit is designed to check the missing data existing or not. An extensive body of empirical comparisons with other existing imputation algorithms over real-world data and public dataset with different ratio of missing data verifies the performance of our model.

**relations**

decay factor **USED-FOR** shortcut connection . update gate **USED-FOR** Gated recurrent unit ( GRU ) . identity skip-connections **USED-FOR** residual layers . Gated recurrent unit ( GRU ) **USED-FOR** network model . public dataset **EVALUATE-FOR** imputation algorithms . unraveled view **USED-FOR** residual networks . Gated recurrent unit ( GRU ) **USED-FOR** training parameters . update gate **USED-FOR** training parameters . AlexNet **HYPONYM-OF** architectures . residual network ( ResNet ) **USED-FOR** deep network . ensemble-like behavior **FEATURE-OF** they . VGG **HYPONYM-OF** architecture . Residual networks **USED-FOR** vanishing gradient problem . forgetting gate **CONJUNCTION** output gate . missing data **USED-FOR** dependency relations . biological systems **HYPONYM-OF** architectures . imputation network **USED-FOR** shortcut connection . real-world data **USED-FOR** imputation algorithms . gradient **PART-OF** residual network . ensemble-like behavior **FEATURE-OF** paths . architecture **USED-FOR** computer vision systems . AlexNet **CONJUNCTION** biological systems . redundancy **FEATURE-OF** they . architectures **USED-FOR** complex concepts . mask unit **USED-FOR** missing data . decay factor **FEATURE-OF** imputation network . forgetting gate **USED-FOR** update gate . residual networks **USED-FOR** gradient flow . incomplete sample set **USED-FOR** dependency relations . residual networks **USED-FOR** vanishing gradient problem . decay residual mean unit ( DRMU ) **USED-FOR** incomplete time series . shortcut connection **USED-FOR** degradation problem . decay residual mean imputation GRU ( DRMI-GRU ) **HYPONYM-OF** imputation network . Recovering missing values **USED-FOR** time series tasks . residual network ( ResNet ) **USED-FOR** degradation problem . they **USED-FOR** deep networks . output gate **PART-OF** long short-term memory ( LSTM ) . decay factor **USED-FOR** long-term dependencies .

---

**target\_citation**

**REF** propose an enlightening analysis of the Residual Networks (ResNets), and suppose that ResNet **REF** can be looked upon as a set of shallow networks and helps to form a deep network and copes with degradation problem by shortcut connection.

---

**predicted**

We propose a novel approach to address the issue of replacing missing data and building dependency relations from an incomplete sample set in time series tasks using residual networks (ResNets) and gated recurrent units (GRUs). Our proposed method, called Decay Residual Mean Imputation GRU (DRMI-GRU), incorporates a decay residual mean unit (DRMU) and a mask unit to handle missing data. The DRMU models the incomplete time series by.

---

---

Figure 7.4: Example of Generated Citation text from **Alpaca** with knowledge graph relations

just the learning rate. This methodology allowed us to effectively train the LLMs on the CTG dataset and evaluate their performance using the testing samples.

$$k_i = \frac{1}{2} \left( Q_X \left( \frac{i}{2^n + 1} \right) + Q_X \left( \frac{i + 1}{2^n + 1} \right) \right) \quad (7.1)$$

Here,  $Q_X(\cdot)$  denotes the quantile function of the standard normal distribution  $N(0, 1)$ . In our experiments, we set  $n = 4$  to implement 4-bit quantization.

Table 7.2: Results of the CTG Task Using LLMs (Without Knowledge Graphs)

Model	METEOR	Rouge-1	Rouge-2	Rouge-L
LLaMA	12.83	11.26	1.36	9.59
Alpaca	10.53	9.22	1.21	7.81
Vicuna	<b>14.15</b>	<b>12.88</b>	<b>1.52</b>	<b>10.94</b>

**Evaluation Metrics:** To evaluate the performance of the text generation and summarization tasks, we utilized widely recognized metrics, namely METEOR, ROUGE-N, and ROUGE-L. ROUGE-L focuses on evaluating the longest common subsequence between the generated text and the reference text, thereby measuring structural similarity. In comparison, ROUGE-N quantifies the overlap of n-grams between the generated and reference texts. Furthermore, METEOR enriches the evaluation by considering word similarity and implementing stemming, thereby offering additional insights in conjunction with ROUGE-N.

Table 7.3: Results of CTG Task using LLMs (With Knowledge Graphs)

Model	METEOR	Rouge-1	Rouge-2	Rouge-L
LLaMA	11.61	10.61	0.99	9.01
Alpaca	<b>14.02</b>	12.63	<b>1.54</b>	10.71
Vicuna	13.80	<b>12.87</b>	1.48	<b>10.96</b>

### 7.3.1.6 Evaluation

This study demonstrates the successful use of LLMs for generating citation text within scientific literature. We highlight the significance of employing knowledge graphs constructed from both source and target papers, as these graphs enhance the understanding of relationships and provide structured contextual information among the documents. Our results show that the Alpaca LLM surpasses both LLaMA and Vicuna in CTG, as evidenced by the analysis of outcomes and the overall quality of the produced content. Figures 7.3 and 7.4 illustrate citation examples generated by the top-performing model during inference, showcasing the high-quality results obtained. These findings

emphasize the advantages of integrating LLMs with knowledge graphs to improve the accuracy and contextual relevance of citation text in academic publications.

## 7.3.2 Context-Enhanced Language Models for Generating Multi-paper Citations

### 7.3.2.1 Methodology

In this study [Anand et al. \(2023f\)](#), we focused on fine-tuning three LLMs: LLaMA [Touvron et al. \(2023b\)](#), Alpaca [Taori et al. \(2023\)](#), and Vicuna [Chiang et al. \(2023\)](#), specifically for the task of generating multi-citation text. We evaluated these models using performance metrics such as METEOR, ROUGE-1, ROUGE-2, and ROUGE-L, which served as our baseline metrics. To enhance citation generation, we extracted relationships from both the source and target documents and integrated them into the prompting process. Our experimental results clearly show that incorporating knowledge graph relationships into the prompting significantly improves the performance of the generated citation texts compared to the baseline models.

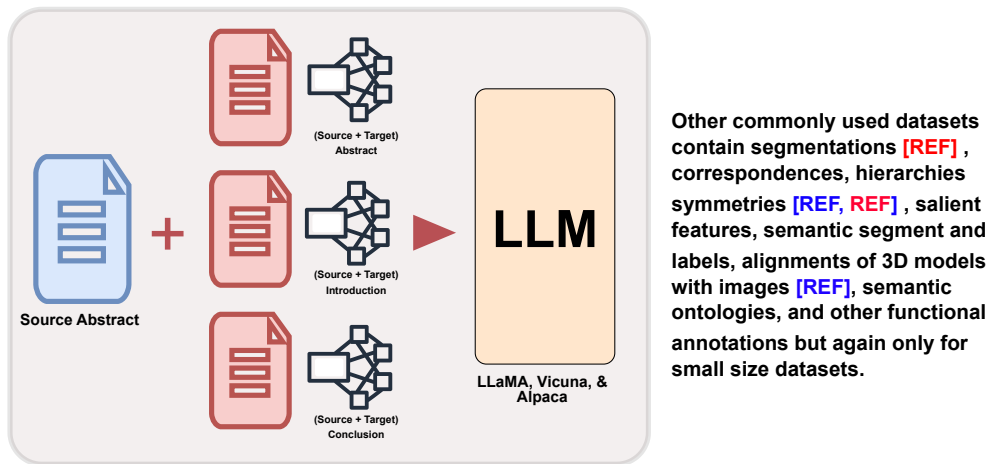


Figure 7.5: Multi-Sentence CTG

### 7.3.2.2 Problem Formulation and Notations

The problem addressed in this work can be articulated as follows: given a citing document’s abstract  $A$  and a collection of abstracts, introductions, and conclusions from related documents  $B = \{b_1, b_2, \dots, b_n\}$ , the objective is to create a multi-sentence para-

graph summarizing all cited documents  $b_i$  in the context of the citing abstract  $A$ . To achieve this, we curated a dataset from the benchmark S2ORC [Lo et al. \(2019\)](#), making modifications to better align with our task. For each citing abstract, we included two to three additional cited papers, ensuring that the output consisted of multiple sentences referencing several documents for each pair  $(A, B)$ . The process is illustrated in Fig. [7.5](#). We initiate the process by taking the source abstract and combining it with the abstract, introduction, and conclusion of the target paper. Utilizing PL-Marker [Ye et al. \(2022\)](#), we extract knowledge graph relationships, which are subsequently utilized as input in prompts for the LLM, as depicted in Fig. [7.7b](#).

### 7.3.2.3 Dataset

For the task of generating multi-sentence citation text, we developed a new dataset called MCG-S2ORC, using data from the S2ORC [Lo et al. \(2019\)](#). S2ORC (Semantic Scholar Open Research Corpus) is a comprehensive repository comprising 81.1 million English-language academic papers across various fields. From this extensive dataset, we concentrated on papers categorized within the "Computer Science" domain, which encompasses approximately 6.0 million papers. Each sample in this subset contains the fields "source paper id," "source abstract," and "body text," with sections such as Introduction and Methodology being particularly relevant.

We processed the Computer Science dataset to extract citation details in JSON format. The resulting dataset contains samples that represent individual citation instances, with each sample structured as key-value pairs of relevant information. The "source paper id" field assigns a unique identifier to the source paper, while the "source abstract" field provides a concise summary of the paper's content. Citation information is found in the "citation texts" field, structured as an array where each element corresponds to a specific citation instance. Within each citation, the "citation text" field contains the actual citation extracted from the source material. Additional metadata is included in the "citation meta" field, which details elements such as citation number, the section of the paper referenced, and information about the cited works (including their titles, abstracts, introductions, and conclusions). To ensure the dataset's suitability for generating multi-reference citation text, we specifically filtered for citations that referenced multiple papers within a single sentence, implementing necessary data transformations.

The final version of our dataset consists of 17,210 samples of multi-reference citation texts. Table 7.4 provides a detailed overview of the statistics for the MCG-S2ORC dataset.

Table 7.4: Statistics of the dataset created from the S2ORC corpus.

Statistic	CTG-S2ORC	Train	Validation	Test
# citations	17,210	13,779	1,716	1,715
# unique papers	17,210	13,779	1,716	1,715
<b>CITATIONS</b>				
Avg # characters	227.29	227.40	230.25	223.37
Max # characters	2,416	2,416	1,862	1,061
<b>SOURCE ABSTRACTS</b>				
Avg # characters	1,122.95	1,120.73	1,111.55	1,152.23
Max # characters	5,516	5,516	4,343	3,642
<b>TARGET ABSTRACTS</b>				
Avg # characters	998.48	997.87	999.35	1,002.56
Max # characters	93,551	93,551	8,674	4,924
Avg # of Targets per sample	2	2	2	2

### 7.3.2.4 Large Language Models

We fine-tuned three LLMs to generate multi-sentence citation text. The specifications of these models are presented in Fig. 7.6 with further details outlined below.

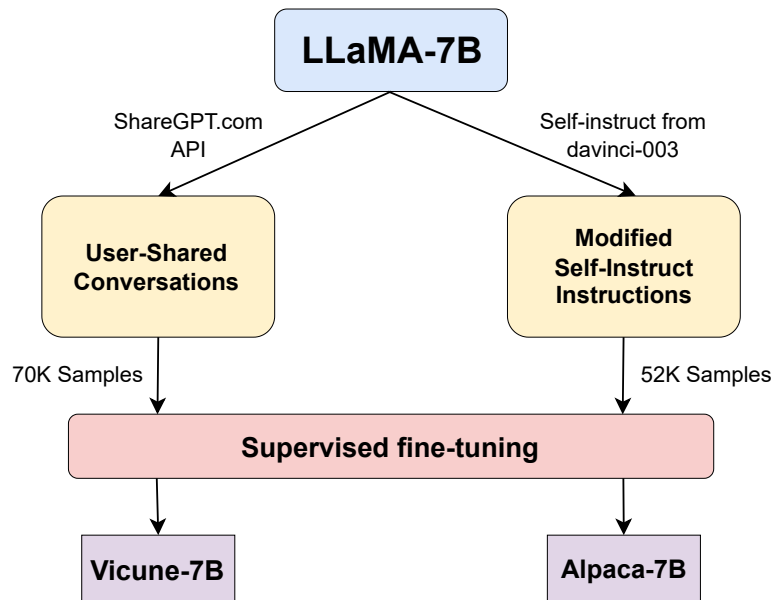


Figure 7.6: LLaMA, Vicuna & Alpaca

LLaMA is a transformer-based model available in four different configurations: 7B, 13B, 33B, and 65B parameters. It is trained solely on publicly available datasets, encompassing roughly 1.4 trillion tokens across 20 different languages (Touvron *et al.*)

<pre> <b>### Instruction:</b>     Generate the citation text. <b>### Input:</b>     {data_point["source_abstract"]}     {data_point["Target1_abstract"]}     {data_point["Target2_abstract"]} ... <b>### Response:</b>     {data_point["citation_text"]} </pre>	<pre> <b>### Instruction:</b>     Generate the citation text. <b>### Input:</b>     {data_point["source_abstract"]}     {relations(source_abstract + target_abstracts)}     {relations(source_abstract + target_conclusions)}     {relations(source_abstract + target_introductions)} ... <b>### Response:</b>     {data_point["citation_text"]} </pre>
(a)	(b)

Figure 7.7: Prompt structures for LLMs. (a) without KG relationships; (b) with KG relationships.

(2023b). In contrast, Alpaca Taori *et al.* (2023) is a language model fine-tuned through supervised learning based on the LLaMA 7B model, incorporating 52K instruction-following examples generated by OpenAI’s text-davinci-003 model Touvron *et al.* (2023b); Katar *et al.* (2022). Vicuna, another variant, is fine-tuned on a LLaMA base model using approximately 70K user-shared conversations sourced from ShareGPT.com via open APIs Chiang *et al.* (2023). It is important to note that Vicuna has limitations in reasoning abilities, mathematical understanding, self-identification, and fact-checking, and it is not specifically designed to mitigate bias, toxicity, or enhance safety Chiang *et al.* (2023).

The prompts used for fine-tuning the baseline LLMs are illustrated in Fig. 7.7a, where each data point corresponds to an individual sample from our dataset.

### 7.3.2.5 Experimental Settings and Evaluation Metrics

We partitioned the complete dataset, MCG-S2ORC, which comprises 17,210 samples, into training, testing, and validation sets. The training set includes 13,000 samples, while both the testing and validation sets contain 1,000 samples each. After preparing and preprocessing the dataset, we fine-tuned three LLMs on MCG\_S2ORC to generate citation text, as described earlier. The prompt structure for the LLMs, shown in Fig. 7.7a, was tokenized and provided to the models during fine-tuning, allowing them to adjust their weights accordingly. According to the results summarized in Table 7.5, Vicuna Chiang *et al.* (2023) outperformed LLaMA and Alpaca in generating citation text from our dataset. These outcomes establish our baseline for subsequent experiments.

We subsequently carried out further experiments utilizing knowledge graphs derived from the abstract, introduction, and conclusion sections of the target paper as prompts.

This approach aimed to enhance the capture of relationships and coherence within the text, ultimately striving to produce more meaningful citations. The relations within the knowledge graph were extracted using PL-Marker [Ye et al. \(2022\)](#). The results of this experiment, presented in Table [7.6](#), indicate an improvement in the performance of all models compared to the baseline results.

For the fine-tuning of the LLMs, we implemented QLora [Dettmers et al. \(2023\)](#), a technique designed for memory efficiency. QLora increases memory efficiency by backpropagating gradients through a frozen, 4-bit quantized pre-trained language model, which facilitates the generation of LoRA.

$$k_i = \frac{1}{2} \left( Q_X \left( \frac{i}{2^n + 1} \right) + Q_X \left( \frac{i + 1}{2^n + 1} \right) \right) \quad (7.2)$$

In this equation,  $Q_x(\cdot)$  represents the quantile function of the standard normal distribution  $N(0, 1)$ . For our experiments, we used  $n = 4$ , corresponding to 4-bit quantization.

$$\begin{aligned} m_t &= \beta m_{t-1} + \eta \nabla J(w_t) \\ v_t &= \gamma v_{t-1} + (1 - \gamma) \nabla J(w_t)^2 \end{aligned} \quad (7.3)$$

For optimization, we employed the **AdamW** optimizer [Kingma and Ba \(2014a\)](#) in conjunction with a Linear Scheduler. The learning rate was set at 3e-4, with 100 warmup steps included to incrementally adjust the learning rate.

$$\hat{m}_t = \frac{m_t}{1 - \beta^t} \quad \hat{v}_t = \frac{v_t}{1 - \gamma^t} \quad (7.4)$$

The bias correction is represented in Equation [7.4](#), and the final weight update for the Adam optimizer is as follows:

$$w_{t+1} = w_t - \frac{\eta}{\sqrt{\hat{v}_t + \epsilon}} \hat{m}_t \quad (7.5)$$

Here,  $\epsilon$  serves as a small error term to ensure the denominator never becomes zero.

### 7.3.2.6 Results and Discussion

The results obtained from fine-tuning the LLMs for the multi-reference CTG task are summarized in Table [7.5](#). The findings clearly indicate that Vicuna [Chiang et al. \(2023\)](#)

---

---

**source\_abstract**

In order to exploit remote sensing data operationally for precision agriculture applications, efficient and automated methods are required for the accurate detection of vegetation, crops and different crop varieties. To this end, we have designed, developed and evaluated an object-based classification framework towards the detection of vineyards, the vine canopy extraction and the vine variety discrimination from very high resolution multispectral data ... The combined analysis of the experimental results with the spectral signatures from the in situ reflectance data indicated that certain vine varieties (e.g., Merlot) presented distinct spectral patterns across the VNIR spectrum.

**target\_1\_abstract**

Image classification is one of the most basic operations of digital image processing. The present review focuses on the strengths and weaknesses of traditional pixel-based classification (PBC) and the advances of object-oriented classification (OOC) algorithms employed for the extraction of information from remotely sensed satellite imageries. The state-of-the-art classifiers are reviewed for their potential usage in urban remote sensing (RS), with a special focus on cryospheric applications. Generally, classifiers for information extraction can be divided into three catalogues: 1) based on the type of learning (supervised and unsupervised), 2) based on assumptions on data distribution (parametric and non-parametric) and, 3) based on the number of outputs for each spatial unit (hard and soft) ... The present review would be beneficial for developing new classifiers in the cryospheric environment for better understanding of spatial-temporal changes over long time scales.

**target\_2\_abstract**

Tree species diversity is a key parameter to describe forest ecosystems. It is, for example, important for issues such as wildlife habitat modeling and close-to-nature forest management. We examined the suitability of 8-band WorldView-2 satellite data for the identification of 10 tree species in a temperate forest in Austria. We performed a Random Forest (RF) classification (object-based and pixel-based) using spectra of manually delineated sunlit regions of tree crowns ... Beside the impact on overall classification accuracy, the importance of the spectral bands was evaluated with two measures provided by RF. An in-depth analysis of the RF output was carried out to evaluate the impact of reference data quality and the resulting reliability of final class assignments. Finally, an extensive literature review on tree species classification comprising about 20 studies is presented.

---

**target**

Motivated by the successful application of object-based image analysis in very high resolution satellite data **REF REF, REF, REF, REF, REF, REF, REF** the developed methodology was based on OBIA principles, while it was designed to address vineyard detection, vine canopy extraction and vine variety discrimination.

---

**predicted**

The citation text for the source "Vegetation monitoring in vineyards using a hybrid feature extraction/classification approach" is: **REF, REF, REF**. Vegetation monitoring in vineyards using a hybrid feature extraction/classification approach. IEEE Transactions on Geoscience and Remote Sensing, 50(2)

---

---

Figure 7.8: Example of Generated Citation text from the best Model (**Vicuna**) without knowledge graph relations

outperforms the other models across all assessed metrics. An example of the citation text produced by Vicuna, the top-performing fine-tuned model, as shown in Figure [7.8](#)

Table [7.6](#) displays the evaluation outcomes for models that utilize prompts incorporating knowledge graph relations extracted from the abstract, introduction, and conclusion sections of both source and target papers. The results indicate that Vicuna continues to excel in citation generation performance compared to baseline models that do not leverage knowledge graph relations. This enhancement is largely due to the inte-

---

---

**source\_abstract**

In order to exploit remote sensing data operationally for precision agriculture applications, efficient and automated methods are required for the accurate detection of vegetation, crops and different crop varieties. To this end, we have designed, developed and evaluated an object-based classification framework towards the detection of vineyards, the vine canopy extraction and the vine variety discrimination from very high resolution multispectral data ... The combined analysis of the experimental results with the spectral signatures from the in situ reflectance data indicated that certain vine varieties (e.g., Merlot) presented distinct spectral patterns across the VNIR spectrum.

**relations**

detection of vegetation **CONJUNCTION** crops . remote sensing data **USED-FOR** precision agriculture applications . rules **CONJUNCTION** segmentation scales . spectral , spatial and textural features **CONJUNCTION** rules . object-based image analysis **USED-FOR** parameters . multitemporal WorldView-2 satellite data **USED-FOR** methodology . portable spectroradiometer **USED-FOR** in situ canopy reflectance observations . approach **USED-FOR** vineyards . completeness and correctness detection rates **EVALUATE-FOR** approach . ... . object-based classification framework **USED-FOR** detection of vineyards . vine canopy extraction **CONJUNCTION** vine variety discrimination . object-based classification framework **USED-FOR** vine variety discrimination . object-based classification framework **USED-FOR** vine variety discrimination . detection of vineyards **CONJUNCTION** vine canopy extraction . object-based classification framework **USED-FOR** vine canopy extraction . high resolution multispectral data **USED-FOR** object-based classification framework . high resolution multispectral data **USED-FOR** vine variety discrimination . segmentation scales **CONJUNCTION** parameters . rules **CONJUNCTION** segmentation scales . spectral , spatial and textural features **CONJUNCTION** rules . object-based image analysis **USED-FOR** parameters . multitemporal WorldView-2 satellite data **USED-FOR** methodology . detection of vineyards **CONJUNCTION** vine canopy extraction . portable spectroradiometer **USED-FOR** in situ canopy reflectance observations . approach **USED-FOR** vineyards . completeness and correctness detection rates **EVALUATE-FOR** approach . overall accuracy ( OA ) rates **EVALUATE-FOR** vine canopy extraction methodology . spectral patterns **FEATURE-OF** VNIR spectrum . in situ reflectance data **USED-FOR** spectral signatures . Merlot **HYPONYM-OF** vine varieties . Tree species diversity **USED-FOR** forest ecosystems . wildlife habitat modeling **CONJUNCTION** close-to-nature forest management . close-to-nature forest management **HYPONYM-OF** issues . object-based classification framework **USED-FOR** vine canopy extraction . wildlife habitat modeling **HYPONYM-OF** issues . object-based approach **COMPARE** pixel-based approach . Scots pine **CONJUNCTION** European beech . Norway spruce **CONJUNCTION** Scots pine . European beech **CONJUNCTION** English oak . classification accuracy **EVALUATE-FOR** spectral bands . high resolution multispectral data **USED-FOR** object-based classification framework . high resolution multispectral data **USED-FOR** vine variety discrimination . segmentation scales **CONJUNCTION** parameters.

---

**target**

Motivated by the successful application of object-based image analysis in very high resolution satellite data **REF REF, REF, REF, REF, REF, REF** the developed methodology was based on OBIA principles, while it was designed to address vineyard detection, vine canopy extraction and vine variety discrimination.

---

**predicted**

The paper presents an object-based classification framework for the detection of vineyards, the vine canopy extraction, and the vine variety discrimination from very high resolution multispectral data. The authors propose a novel set of spectral, spatial, and textural features, as well as rules, segmentation scales, and parameters based on object-based image analysis. The validation of the developed methodology was carried out on multitemporal WorldView-2 satellite.

---

---

Figure 7.9: Example of Generated Citation text from the best Model (**Vicuna**) with knowledge graph relations

gration of knowledge graphs, which facilitates a deeper contextual understanding and improves the coherence and relevance of the generated text. As a result, the outputs produced by the model are more contextually rich and of higher quality, contributing to

Table 7.5: Results of Fine-Tuned LLM

Model	METEOR	Rouge-1	Rouge-2	Rouge-L
LLaMA	11.73	10.74	1.21	9.15
Alpaca	9.74	9.04	1.33	7.78
Vicuna	<b>12.56</b>	<b>12.02</b>	<b>1.44</b>	<b>10.24</b>

its overall superior performance. An example of the citation generated during inference is presented in Figure 7.9

Table 7.6: Results of Fine-Tuned Model + Knowledge Graph as Prompt

Model	METEOR	Rouge-1	Rouge-2	Rouge-L
LLaMA	11.46	10.79	1.23	9.14
Alpaca	<b>13.39</b>	12.42	<b>1.74</b>	10.59
Vicuna	13.18	<b>12.65</b>	1.49	<b>10.80</b>

### 7.3.3 Advances in CTG: Leveraging Multi-Source Seq2Seq Models and LLMs

#### 7.3.3.1 Methodology

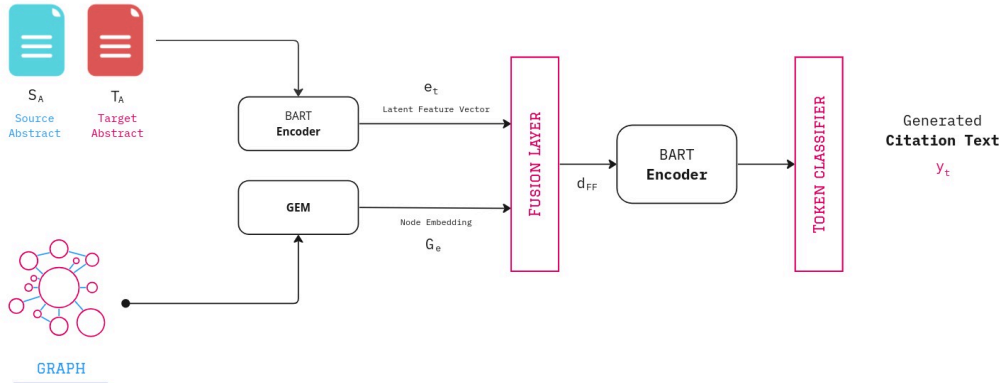


Figure 7.10: Our proposed M-CTG. BART encoder takes source and cited (target) abstracts for text encoding. Graph and style embedding passed in MLP. BART Encoder output combined with MLP output in the Fusion Layer.

Our proposed framework [Anand et al. \(2024f\)](#), M-CTG, is illustrated in Figure 7.10. It processes the source and cited (target) abstracts as inputs and employs a BART [Lewis et al. \(2020\)](#) encoder to derive a latent feature vector denoted as  $e_t$ . Additionally, a Graph Embedding Model (GEM) is employed to generate node embeddings (NE) for a heterogeneous graph representing the authorship relationships. These graph embed-

dings are integrated with the textual representations, forming a succinct semantic representation of the contextual information from both the source and target papers. This composite representation is subsequently merged with the latent feature vector  $e_t$  via a fusion layer. The resulting fused output is then fed into the BART decoder, with the final layer functioning as a token classifier for predicting individual tokens.

In the following sections, we first formalize our problem statement and then provide a detailed explanation of the individual components within the M-CTG framework.

### 7.3.3.2 Problem Statement and Notation

In this section, we approach the CTG problem by modelling the probability of generating the citation text  $C$ , which is denoted as  $Y_t$  in Figure 7.10. The CTG is conditioned on three key elements: the source paper abstract  $S_A$ , the cited paper abstract  $T_A$ , and the Graph embedding  $G_e$ . The corresponding conditional probability is formulated as  $P(C|S_A, T_A, G_e)$ . The main objective is to maximize the probability  $P$  of generating the citation text  $C$  by learning an effective Graph embedding  $G_e$  based on both  $S_A$  and  $T_A$ . Our proposed Graph embedding mechanism is designed to guide the model, enhancing the quality of the generated citation texts. During inference, the model generates citation text  $C$  that articulates the relationship between the source and cited papers, thus providing a coherent summary of their connection.

### 7.3.3.3 Encoder

In our model architecture, we employ a pre-trained BART model Lewis *et al.* (2020) as the encoder. This model is a multi-layered bi-directional transformer, known for its effectiveness in handling various NLP tasks. The encoder processes three types of input: the source abstract ( $S_A$ ), the cited abstract ( $T_A$ ), and the metadata ( $N$ ) related to the papers. The metadata includes four key attributes: author ID, paper ID, year, and venue, for both the source and cited papers. To ensure that the model can distinguish and learn the relevance of these attributes during the text generation phase, we introduce four distinct sets of special separator tokens, represented by  $\langle . \rangle$  and  $\langle / . \rangle$ . The details of these separator token sets are presented in Table 7.3.3.3.

In Table 7.3.3.3, each token serves to demarcate the start and end of an attribute value. We employ these separator tokens ( $\langle . \rangle, \langle / . \rangle$ ) to combine the source ab-

Table 7.7: Metadata used with abstracts of the papers

Token	Description
$\langle author \rangle, \langle /author \rangle$	To specify author id attribute
$\langle paper \rangle, \langle /paper \rangle$	To specify paper id attribute
$\langle year \rangle, \langle /year \rangle$	To specify year attribute
$\langle venue \rangle, \langle /venue \rangle$	To specify venue attribute

abstract  $S_A = s_1^A, s_2^A, \dots, s_n^A$  with the cited (target) abstract  $T_A = t_1^A, t_2^A, \dots, t_n^A$  and the metadata attributes  $N_i = x_1, x_2, \dots, x_n$ . Here,  $e_t$  represents the output of the encoder. The tokens  $\langle s\_abstract \rangle$  and  $\langle t\_abstract \rangle$  are used to indicate the beginning of the source and cited abstracts, respectively. The metadata attributes encode the relationships between key concepts shared across the source and cited papers, assisting the encoder in capturing the context critical for the CTG task. Each metadata attribute corresponds to a distinct aspect of the research paper’s content.

$$e_t = BART_{Encoder}(\langle . \rangle', N_i, \langle /. \rangle', \langle s\_abstract \rangle', S_A, \langle t\_abstract \rangle', T_A) \quad (7.6)$$

### 7.3.3.4 Graph Embedding Model (GEM)

We generate graph embeddings ( $G_e$ ) using GEM. These embeddings represent node embeddings (NE) produced by GEM, which are then combined with the author’s latent vector in the fusion layer. Below, we explain the detailed functionality of GEM.

### 7.3.3.5 GEM

GEM operates on a heterogeneous graph as input and extracts node embeddings (NE) from it. These NE encode the relationships between papers as low-dimensional vectors. By projecting nodes and their neighboring nodes into a latent space, GEM captures higher-level relationships, such as shared research concepts. NE is highly effective in representing both similarities and dissimilarities between papers, offering a computationally efficient alternative to token-level attention between paper pairs. It employs random walk graph traversal to preserve the graph’s structure and the features of research papers during the aggregation process [Perozzi \*et al.\* \(2014\)](#).

**Heterogeneous Graph Generation:** We generate three distinct heterogeneous graphs.

Each of these graphs incorporates various combinations of metadata to identify and evaluate the relationships that most effectively contribute to generating high-quality citation texts. The structures of these graphs are described below:

1. **Graph 1:** In this graph, the nodes represent papers, authors, venues, and years. An edge is created between paper A and paper B if A cites B. Additionally, an edge is created between paper A and author X if X is the author of A, between paper A and venue Y if Y is the venue of A, and between paper A and year Z if Z is the publication year of A.
2. **Graph 2:** In this graph, the nodes consist of papers, authors, and venues. An edge is created between paper A and paper B if A cites B. Similarly, an edge is created between paper A and author X if X is the author of A, and between paper A and venue Y if Y is the venue of A.
3. **Graph 3:** In this graph, the nodes represent papers and authors. An edge is created between paper A and paper B if A cites B. Likewise, an edge is created between paper A and author X if X is the author of A.

To determine the most relevant metadata for citation generation, we constructed these three heterogeneous graphs. Graph 1 integrates papers, authors, venues, and years to explore their interconnections. Graph 2 includes papers, authors, and venues, while Graph 3 focuses solely on papers and authors. Node embeddings (NE) for all three graphs are created using two GEM models, namely Nod2Vec [Grover and Leskovec \(2016\)](#) and DeepWalk [Perozzi \*et al.\* \(2014\)](#).

**Nod2Vec** is a scalable and efficient algorithm designed to generate feature representations for individual nodes, which can also be extended to represent edges (i.e., pairs of nodes). It employs a biased random walk mechanism that explores the graph neighborhoods in both breadth-first search (BFS) and depth-first search (DFS) fashion.

**DeepWalk** is another scalable algorithm that applies deep learning principles to analyze graphs. DeepWalk executes a random walk over the graph to generate a sequence of nodes, which is subsequently processed by a Word2Vec model to produce embeddings. The resulting structural representations capture the graph's underlying regularities through short random walks.

### 7.3.3.6 Fusion Layer

We utilize the fusion layer [Çağlar Gülçehre \*et al.\* \(2015\)](#) to strengthen the interaction between text and graph features. The fusion layer processes the author latent vector

$e_t$ , which encapsulates the text features, along with the output of the MLP layer. It concatenates  $e_t$  with  $G_e$ , producing the following result:

$$d_{FF} = [e_t : G_e] \quad (7.7)$$

By merging the latent feature vector with the graph embedding, the decoder is able to generate citation text more effectively, drawing on all available information in a balanced manner. This approach ensures that the decoder gives equal importance to both the latent text features and the graph embeddings.

### 7.3.3.7 Decoder

The decoder architecture utilized in our model is a multi-layered, uni-directional autoregressive transformer designed to generate citation text. The output of the fusion layer, denoted as  $(d_{FF})$ , is provided as input to this decoder. The model predicts the next token based on the distribution  $Y_t$ , which is formulated as follows:

$$Y_t = \text{softmax}(TC(BART_{dec}(d_{FF}))) \quad (7.8)$$

In this equation,  $d_{FF}$  refers to the output from the fusion layer,  $TC$  represents a linear layer responsible for token classification, and  $BART_{dec}$  is the BART decoder. The BART decoder is fine-tuned using our dataset to generate citation text in an autoregressive manner, meaning that tokens generated at each time step are conditioned on the tokens from previous time steps.

### 7.3.3.8 Large Language Models

We carried out a series of experiments using the Large Language Model (LLM) for text generation and performed a comparative analysis. The LLM utilized in this study was Vicuna (7B and 13B) [Chiang et al. \(2023\)](#). To assess the model’s performance in CTG tasks, we fine-tuned the models using our CTG-S2ORC dataset. Vicuna, a supervised fine-tuned version of LLaMA, was trained on 70K user-shared conversations sourced from ShareGPT.com.

### 7.3.3.9 Zero-Shot CoT

In the Zero-Shot Chain of Thought (CoT) method, the model is prompted to generate rationales (or explanations) in a step-by-step format. These rationales help the model generate citation text. The prompt consists of four primary components: the instruction; the source abstract from the citing paper; the combination of abstract, introduction, and conclusion from the cited paper; and finally, the phrase **Let's think step-by-step**.

### 7.3.3.10 Few-Shot CoT

In the Few-Shot CoT method for CTG, the model is provided with a few examples, each containing abstracts from both the citing and cited papers. The cited paper's example also includes the introduction and conclusion sections. These examples are accompanied by the generated citation text and a step-by-step rationale. For example, two pairs of examples  $(x_1, y'_1)$  and  $(x_2, y'_2)$  illustrate how the model identifies key topics and supporting details from the papers, structuring these findings into standard citation texts. The model then applies this reasoning process to new tasks, guiding it to generate citation texts for new inputs in a similar manner.

### 7.3.3.11 Auto-CoT

Automatic Chain-of-Thought (Auto-CoT) [Zhang et al. \(2022\)](#) is a technique that automates the process of generating demonstrations with questions and associated reasoning chains. This method is beneficial because it minimizes the need for manual creation of examples with rationale chains, while also facilitating the creation of relevant examples where each consists of a question followed by a logical reasoning chain that leads to the answer. The Auto-CoT process is composed of two key stages:

*Question Clustering* - First, a vector representation of each question is generated using Sentence-BERT. These contextualized vectors are averaged to create a fixed-size representation of each question, capturing its semantic content. The KMeans algorithm is then applied to partition the vectors into  $k$  clusters of semantically related questions, ensuring consistency by using a random seed.

*Demonstration Sampling* - Based on their distance to the centroid, representative

questions are chosen from each cluster. The reasoning chain for each question is then produced using the Zero-Shot CoT approach, ensuring adherence to text criteria like a maximum length and the inclusion of a period at the end. These demonstrations, which include the question, reasoning chain, and answer, are subsequently incorporated into the prompt for effective model performance.

### 7.3.3.12 Dataset for CTG

The S2ORC dataset [Lo et al. \(2020\)](#) serves as the foundational resource for the CTG task, encompassing a total of 81.1 million English-language academic papers spanning multiple disciplines. This dataset offers comprehensive metadata, resolved bibliographic references, and abstracts, along with the structured full text for 8.1 million open-access documents. Moreover, it includes annotated inline mentions of citations, figures, and tables, all of which are linked to their corresponding references. In order to address generalizability challenges and facilitate effective evaluation of CTG tasks, we curated a specialized dataset from S2ORC, named CTG-S2ORC, which was synthesized with this specific objective in mind.

### 7.3.3.13 Dataset Build

From the S2ORC dataset [Lo et al. \(2020\)](#), we begin by selecting papers whose "Field of Study" is classified under "Computer Science." Currently, the computer science domain includes a total of 6.0M papers, of which up to 580k have full-text PDF parses available. Each PDF parse is divided into sections such as "abstract" and "body\_text." For the purposes of our dataset, we focus solely on the citations found in the "body\_text."

The "body\_text" consists of multiple paragraphs, often organized into sections like Introduction and Methodology. To process this text, we use NLTK's Sentence Tokenizer, which enables us to break the body text into individual sentences. Each section in the body\_text may include "cite\_spans," which are dictionaries containing details about citations to other papers within that paragraph. For example, a cite\_span might appear as [start: 121, end: 124, text: [12], bibref: BIBREF23], providing information about the position of the citation in the text. We analyze each section of the body\_text to detect cite\_spans and extract the corresponding citation sentences.

For our dataset, we assume that each citation is expressed in a single sentence. In

cases where a paper is cited multiple times within the same paragraph, we only consider the first citation sentence. Furthermore, we disregard citations that reference more than one paper within a single sentence.

Each paper is assigned a unique "paper\_id," which allows us to track its citations. A paper's "inbound\_citations" refer to the list of paper\_ids that cite the current paper, while its "outbound\_citations" refer to the list of paper\_ids cited by the current paper. To ensure a denser citation network for analysis, we restrict our evaluation to publications with more than 30 "inbound\_citations" and "outbound\_citations."

#### **7.3.3.14 Experimental settings, Evaluation metrics, and Model comparisons.**

In this section, we describe the experimental setup, evaluation criteria, and model comparisons. We fine-tuned and tested several text generation models using our dataset. Upon comparing Pegasus [Zhang et al. \(2020a\)](#), T5 [Raffel et al. \(2020\)](#), BART, and LLMs (LLaMA-7B/13B, Vicuna-7B/13B), we found that Vicuna-13B demonstrated superior performance in Auto-CoT, which led to its selection as our baseline model.

#### **7.3.3.15 Experimental Settings:**

In this research, we partitioned the CTG-S2ORC dataset into 89,000 samples for training and 20,000 samples for testing. The model architecture was built upon pre-trained BART layers for both the encoder and decoder components. We optimized the training process by minimizing the negative log-likelihood between the generated citation text and the corresponding target citation text. The AdamW optimizer [Kingma and Ba \(2014b\)](#) was employed, along with a Linear Scheduler, setting the learning rate to  $1 \times 10^{-5}$  and utilizing 5,000 warmup steps. All experiments were executed on a single A100 GPU. During inference with the LLMs, we applied a top-k value of 40, a temperature setting of 0.8, and a top-p value of 0.90. To measure the similarity between the generated citations and the original text, we implemented two categories of scoring metrics.

##### **Standard Metrics:**

For the standard assessment of text generation and summarization, we utilized widely recognized metrics, including BLEU [Papineni et al. \(2002b\)](#), METEOR [Michael Denkowski \(2014\)](#), and ROUGE-N and ROUGE-L [Lin \(2004b\)](#). ROUGE-N evaluates the overlap

of n-grams between the generated and reference texts, whereas ROUGE-L emphasizes the longest common subsequence (LCS) shared by both texts. METEOR enhances this analysis by integrating word similarity through stemming, providing additional insights alongside ROUGE-N. Furthermore, we conducted a human evaluation to determine the quality of the generated citations based on factors such as readability, contextual relevance, and coherence.

**Custom Metrics:** We developed a custom metric that blends 60% Jaccard Similarity and 40% Cosine Similarity to assess the accuracy of the generated citation text. The heavier emphasis on Jaccard Similarity ensures a focus on the correct identification of crucial technical terms, which is essential for tasks that require high precision in term usage rather than general semantic similarity. This custom approach involves segmenting the text to align with the length of the reference citation and averaging the similarity scores across these segments to minimize length bias. Jaccard Similarity measures the overlap of key terms, processed using TF-IDF while omitting common English words, thus highlighting essential concepts. In contrast, Cosine Similarity evaluates the semantic closeness between the generated and reference texts based on their TF-IDF vectors, providing a more comprehensive assessment of content and context.

The choice of a 60% weight for Jaccard Similarity and 40% for Cosine Similarity reflects the priority of identifying technical terms with precision (Jaccard Similarity) over semantic coherence (Cosine Similarity) when evaluating CTG. We experimented with various weightings and determined that this combination achieved the best balance. This strategy helps overcome the limitations of conventional metrics such as BLEU, METEOR, ROUGE, and BERTScore, which may place too much emphasis on either literal matching or overall coherence, thereby failing to capture the level of precision needed for accurate citation of technical content.

### 7.3.3.16 Knowledge Graph Citation Text Scores

Table [7.8](#) illustrates the comparison of scores for various models in generating citation text using two different inputs: Knowledge Graph (KG) and Knowledge Graph combined with a Summary (KG + Summary). It is evident from the results that all models demonstrate improved performance across all metrics when utilizing the (KG + Summary) input, compared to using KG alone. Specifically, the increase in BLEU, ME-

Table 7.8: Comparison of model performance with and without Summary in Knowledge Graphs (KG)

Model	BLEU-1	METEOR	ROUGE-1	BERT F1	Custom Score
<b>KG</b>					
LLAMA_7B	2.17	9.70	3.91	44.63	3.18
LLAMA_13B	2.19	9.81	3.94	44.65	3.07
VICUNA_7B	2.08	9.24	3.86	44.61	3.11
VICUNA_13B	2.30	10.20	4.09	44.65	3.48
<b>KG + Summary</b>					
LLAMA_7B	2.85	11.26	4.80	40.06	3.64
LLAMA_13B	3.12	12.14	5.28	40.05	3.65
VICUNA_7B	2.96	11.46	5.12	39.81	3.64
<b>VICUNA_13B</b>	<b>3.40</b>	<b>12.81</b>	<b>5.57</b>	<b>40.18</b>	<b>3.68</b>

TEOR, and ROUGE scores suggests that the inclusion of summaries leads to a greater overlap of key terms with the reference text. However, there is a slight reduction in the BERTscore, indicating a marginal decrease in capturing semantic context. Furthermore, the Custom Score shows improvement across all models, highlighting that the addition of summaries aids in generating text that better aligns with the reference text, particularly in terms of Jaccard and Cosine Similarity.

### 7.3.3.17 LLM’s Citation Text Scores

Table 7.9: Comparison of model performance with various methods

Model	BLEU-1	METEOR	ROUGE-1	BERT F1	Custom Score
<b>vicuna-7b-v1.5-16k</b>					
Auto-CoT 1s	10.97	11.90	12.38	54.50	10.26
Auto-CoT 2s	10.54	12.71	13.04	54.70	10.77
Auto-CoT 3s	12.52	14.59	13.33	54.74	11.32
Few Shot 1s	7.92	10.80	9.29	50.72	7.41
Few Shot 2s	8.54	11.18	9.73	51.38	8.01
Few Shot 3s	10.18	11.69	10.68	53.19	9.53
Zero-Shot-CoT	1.16	6.77	2.41	35.69	0.75
<b>vicuna-13b-v1.5-16k</b>					
Auto-CoT 1s	15.63	18.87	13.50	57.27	9.97
Auto-CoT 2s	18.18	21.01	17.09	58.93	13.04
<b>Auto-CoT 3s</b>	<b>19.71</b>	<b>21.77</b>	<b>18.17</b>	<b>59.64</b>	<b>18.76</b>
Few Shot 1s	14.36	16.01	13.60	44.95	13.41
Few Shot 2s	16.14	19.57	14.05	42.20	15.79
Few Shot 3s	18.94	21.22	15.22	47.72	19.04
Zero-Shot-CoT	1.38	8.26	2.53	39.38	0.89

We conducted a comprehensive analysis of various prompting techniques on LLMs, namely Vicuna 7B and Vicuna 13B, using our CTG-S2ORC dataset shown in Table 7.9. We have fine-tuned these models on the CTG-S2ORC dataset and calculated the scores.

We observe that Auto-CoT (1 shot, 2 shot, and 3 shot) performs better than its respective counterparts in Few Shot. This is primarily because Auto-CoT automates the construction of demonstrations, which includes generating reasoning chains using

Zero-Shot CoT. Secondly, Auto-CoT also emphasizes the diversity of demonstration questions by clustering and selecting representative questions from each cluster. This approach ensures that the model is exposed to a variety of reasoning patterns and scenarios, enhancing its generalization capabilities.

We also observed that citation texts generated using Zero-shot CoT underperformed compared to those produced by Knowledge Graph (KG) methods, both with and without summaries, in Vicuna 7B and 13B models. This outcome highlights Zero-shot CoT’s limitations in CTG, where KG methods demonstrate superiority due to their structured representation of relationships and contexts. Conversely, Few-shot CoT exhibited superior performance over KG methods in the same models. This advantage stems from Few-shot CoT’s ability to utilize example-based learning, effectively guiding the model to produce citation texts that are contextually richer and more nuanced, thus aligning more closely with the evaluative metrics.

### 7.3.3.18 Comparison with other models

We conducted a comparative analysis between our best-performing model and the model proposed by [Luu \*et al.\* \(2020\)](#). To ensure a fair comparison, we curated an additional smaller dataset from S2ORC, which is distinct from CTG-S2ORC. This dataset comprised 79K examples used for fine-tuning and 19.7K examples for testing both models. The results of this comparison, presented in [Table 7.10](#), demonstrate that our model surpasses [Luu \*et al.\* \(2020\)](#)’s model across all evaluation metrics. Specifically, our model achieves an improvement of 18.09% in ROUGE-1 and a significant gain of 46.28% in the METEOR score.

Table 7.10: Model Comparison

Model	METEOR	Rouge-1	Rouge-2	Rouge-L
Luu’s model (SOTA)	10.75	12.46	9.6	13.87
<b>M-CTG (Our Model)</b>	<b>15.73</b>	<b>14.71</b>	<b>19.57</b>	<b>15.57</b>

### 7.3.3.19 Evaluation

In our study, the proposed model M-CTG demonstrates superior performance over the baseline BART across all selected metrics for text generation in the research domain. This underscores the significance of the graph embedding we introduced, which ef-

fectively enables our model to capture the intricate relationship between the source and cited papers. Furthermore, we have shown that employing the Vicuna-13B LLM for CTG produces higher quality outputs when compared to both other LLMs and the M-CTG model itself. The citation example generated by Vicuna during the inference process is shown in Fig. 7.9, clearly illustrating the outstanding quality of the citation text generated by this model.

## 7.4 Conclusion

In this study, we tackled the challenge of generating coherent and contextually relevant citation texts in scientific papers, a critical aspect of academic writing. To advance this field, we curated a specialized dataset, MCG-S2ORC, derived from the S2ORC corpus, and introduced the CTG-S2ORC dataset specifically designed for citation generation research. We fine-tuned several LLMs, including LLaMA, Alpaca, and Vicuna, to perform the citation generation task. Among these, Vicuna consistently outperformed the others, demonstrating superior capability in generating multi-sentence citations.

To further enhance the accuracy and coherence of the generated citations, we integrated knowledge graphs into the prompting mechanism of the models. By extracting and incorporating entity relations from the abstracts, introductions, and conclusions of both source and target papers using the PL-Marker, we significantly improved the models' understanding of the intricate relationships within the scientific texts. This integration not only bolstered the models' performance but also underscored the potential of knowledge graphs as valuable tools in CTG.

Our experiments also revealed that advanced prompting techniques, such as Zero-Shot CoT, Few-Shot CoT, and Auto-CoT, markedly improved the citation generation capabilities of the models, particularly when using the Vicuna 13B model. These techniques enabled the models to produce more contextually accurate and plausible citation texts, outperforming baseline methods.

Looking ahead, future research could focus on developing unique embeddings for individual authors, which could refine the models' ability to capture detailed, author-specific information. Additionally, incorporating advanced reasoning techniques, such as CoT prompting, holds promise for further enhancing the depth, precision, and over-

all quality of generated citation texts. This study highlights the significant potential of combining LLMs with knowledge graphs and sophisticated prompting strategies to advance the field of CTG.

# CHAPTER 8

## Summary and Future Directions

### 8.1 Introduction

This thesis explores the application of advanced AI models, particularly LLMs and multimodal reasoning systems, across a variety of domain-specific tasks, with a focus on education, geometry problem-solving, GEC, CTG, and controlled text synthesis. Current general-purpose models face significant limitations in these areas, particularly when addressing tasks that require deep reasoning or multimodal interpretation. This research investigates methods to augment SOTA models, proposing novel algorithms and creating specialized datasets to overcome these challenges. The goal is to improve AI's capacity for domain-specific reasoning, enhance performance in real-world tasks, and contribute to future developments in AI-driven education and multimodal processing.

### 8.2 Summary

This thesis presents several key contributions across a range of AI applications, focusing on the augmentation of LLMs and multimodal models for domain-specific tasks. The research spans multiple domains, from physics education and student engagement analysis to geometry problem-solving, grammatical error correction, and citation text generation (CTG). Below is a summary of the key contributions made by each work included in this thesis.

- 1. Physics Education and Question Answering (PhyQA, RetriPhy, and MM-PhyQA Systems):** This research developed the largest dataset of high school physics questions, designed to fine-tune LLMs for physics education. The dataset enables more accurate AI performance in solving complex physics problems, particularly in text-based question answering. Future work will focus on extending these models to handle multimodal elements, such as diagrams and graphs. Techniques like Chain of Thought (CoT) and Tree of Thought (ToT) prompting show promise for further enhancing model performance in complex reasoning tasks, as demonstrated by the PhyQA and RetriPhy systems. These developments lay the groundwork for improving AI's capabilities in STEM education.

2. **Student Engagement in Online Education (ECLIPSE Dataset and CG-ViT Model):** The ECLIPSE dataset is a significant contribution to the field of student engagement detection in online education, providing large-scale data on student behavior during virtual lectures. The CG-ViT model, introduced as part of this work, integrates video lecture content and student behavior to enhance the detection of low-engagement states. The NeuralGaze model and GLAMOR-Net also demonstrated significant improvements in engagement detection accuracy, outperforming previous benchmarks. This research advances the ability of AI systems to deliver personalized and responsive learning experiences, particularly in virtual environments.
3. **Multimodal Numerical Reasoning, Geometry Problem-Solving, and Bilingual Curriculum Learning (GPSM4K, GeoVQA, MathQuest, and Indi Math QA):** This work introduced four datasets—GPSM4K, GeoVQA, MathQuest, and IndiMathQA—that push the limits of LLMs and multimodal systems in geometry, numerical reasoning, and bilingual mathematical problem-solving. By focusing on geometry-based visual question answering, this research highlighted the need for better integration of theorem-based knowledge and stepwise solution generation. We investigated curriculum learning for bilingual fine-tuning on complex mathematical datasets, comparing performance across lightweight open-source models and closed-source models in both English and Hindi. The WizardMath-7B model emerged as a top performer on the IndiMathQA dataset, demonstrating enhanced accuracy through structured solution formats and bilingual training approaches. These contributions provide a valuable resource for further developments in multilingual and multimodal AI systems for mathematics education.
4. **GEC with Context-Aware Models:** This research tackled a key limitation in GEC models: their inability to consider extended context when correcting grammatical errors. By incorporating dynamic context into GEC tasks, the proposed model significantly improved error detection in longer sequences of text. This context-aware approach is particularly valuable in document-level corrections, where maintaining coherence across multiple sentences is critical. The model's success across both scientific and general-purpose datasets demonstrates its adaptability and potential for wide-scale use in real-world applications.
5. **CTG with Knowledge Graph Integration:** The citation generation task was addressed by curating specialized datasets and integrating knowledge graphs into the prompting mechanisms of LLMs. This approach, tested on models such as Vicuna and Alpaca, significantly improved the quality and coherence of generated citations in scientific texts. By leveraging entity relations extracted from source and target papers, this work showcased how knowledge graphs can enhance the performance of citation generation models. Future work will explore embeddings for individual authors to further refine model accuracy in capturing detailed author-specific information.

## Project GitHub Links

- **SciPhyRAG:** <https://github.com/arnav10goel/SciPhy-RAG>
- **MM-PhyQA:** <https://github.com/midas-research/Optimizing-Multimoda>

- **Mathify:** <https://github.com/midas-research/mathify>
- **TC-OCR:** <https://huggingface.co/spaces/mohit15/mmasia>
- **Randomlaynet:** <https://github.com/midas-research/randomlaynet>
- **KG-CTG:** <https://github.com/midas-research/M-CTG>
- **Multilingual Math Reasoning:** <https://github.com/midas-research/Multilingual-Mathematical-Reasoning>
- **M-CTG:** <https://github.com/midas-research/M-CTG>

### 8.3 Future Work

The research presented in this thesis opens up several exciting directions for future exploration. In the domain of physics education, extending LLMs to handle multimodal problems involving diagrams and graphs will be crucial for tackling the complexity of real-world physics questions. Techniques like Chain of Thought prompting, combined with RLHF, hold great promise for improving model alignment and performance in educational contexts.

For online education, future work will focus on personalizing engagement detection models to individual learners. Tailoring AI systems to student-specific behaviors could significantly enhance learning outcomes, particularly in virtual and hybrid learning environments. The introduction of real-time interventions based on engagement levels could further improve educational experiences.

In geometry problem-solving and numerical reasoning, future efforts will aim to develop models that can better integrate theorem-based knowledge and perform stepwise reasoning. Expanding the datasets to include more advanced mathematical concepts, as well as developing models that combine visual and textual reasoning more effectively, will be key to advancing this field.

In grammatical error correction, future research will explore more sophisticated ways to incorporate extended context into correction tasks, particularly in scientific writing, where preserving meaning across longer passages is essential. The goal will be to create models capable of handling document-level corrections with minimal human intervention.

Finally, For CTG, future work could focus on developing unique embeddings for individual authors, allowing models to capture nuanced, author-specific writing styles and preferences. Additionally, incorporating advanced reasoning techniques, such as Chain of Thought prompting, could further improve the depth and accuracy of generated citations.

Together, these future directions represent a continued effort to push the boundaries of AI in domain-specific tasks, advancing the fields of education, NLP, and multimodal reasoning. The contributions of this thesis lay the groundwork for future innovation and collaboration, with the potential to significantly impact AI's role in real-world applications.

## REFERENCES

1. **Anand, A., K. Addala, K. Baghel, A. Goel, M. Hira, R. Gupta, and R. R. Shah**, Revolutionizing High School Physics Education: A Novel Dataset. *In International Conference on Big Data Analytics*. Springer, 2023a.
2. **Anand, A., A. Goel, M. Hira, S. Buldeo, J. Kumar, A. Verma, R. Gupta, and R. R. Shah**, SciPhyRAG-Retrieval Augmentation to Improve LLMs on Physics Q &A. *In International Conference on Big Data Analytics*. Springer, 2023b.
3. **Anand, A., M. Gupta, K. Prasad, U. Goel, N. Lal, A. Verma, and R. R. Shah**, KG-CTG: Citation Generation Through Knowledge Graph-Guided Large Language Models. *In International Conference on Big Data Analytics*. Springer, 2023c.
4. **Anand, A., M. Gupta, K. Prasad, N. Singla, S. Sanjeev, J. Kumar, A. R. Shivam, and R. R. Shah** (2023d). Mathify: Evaluating Large Language Models on Mathematical Problem Solving Tasks.
5. **Anand, A., A. Jairath, N. Lal, S. Bangar, J. Sikka, V. Astha, R. Shah, and S. Satoh**, *GEC-DCL: Grammatical Error Correction Model with Dynamic Context Learning for Paragraphs and Scholarly Papers*. 2023e. ISBN 978-3-031-49600-4, 95–110.
6. **Anand, A., R. Jaiswal, A. Dharmadhikari, A. Marathe, H. Papat, H. Mital, A. R. Nair, K. Prasad, S. Kumar, A. Verma, et al.**, Geovqa: A comprehensive multimodal geometry dataset for secondary education. *In 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*. IEEE, 2024a.
7. **Anand, A., R. Jaiswal, A. Dharmadhikari, A. Marathe, H. P. Papat, H. Mital, K. Prasad, R. R. Shah, and R. Zimmermann** (2024b). Improving multimodal llms ability in geometry problem solving, reasoning, and multistep scoring. *arXiv preprint arXiv:2412.00846*.
8. **Anand, A., J. Kapuriya, A. Singh, J. Saraf, N. Lal, A. Verma, R. Gupta, and R. Shah**, MM-PhyQA: Multimodal Physics Question-Answering with Multi-image

- CoT Prompting. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2024c.
9. **Anand, A., J. Kapuriya, A. Singh, J. Saraf, N. Lal, A. Verma, R. Gupta, and R. Shah**, Mm-phyqa: Multimodal physics question-answering with multi-image cot prompting. *In Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2024d.
  10. **Anand, A., A. Mittal, L. Dhawan, M. Ramesh, J. Krishnamurthy, N. Lal, R. Jaiswal, P. Bhuyan, Himani, V. Astha, R. Shah, R. Zimmermann, and S. Satoh**, *Unveiling Learner Dynamics: The ECLIPSE Dataset and NeuralGaze Framework for Prolonged Engagement Assessment in Online Learning*. 2024e. ISBN 9781643685489.
  11. **Anand, A., A. Nair, K. Prasad, V. Narayan, N. Lal, D. Mahata, Y. Singla, and R. Shah**, Advances in citation text generation: Leveraging multi-source seq2seq models and large language models. 2024f.
  12. **Anand, A., K. Prasad, U. Goel, M. Gupta, N. Lal, V. Astha, and R. Shah**, *Context-Enhanced Language Models for Generating Multi-paper Citations*. 2023f. ISBN 978-3-031-49600-4, 80–94.
  13. **Anand, A., K. Prasad, C. Kirtani, A. R. Nair, M. K. Nema, R. Jaiswal, and R. R. Shah**, Multilingual mathematical reasoning: Advancing open-source llms in hindi and english. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39. 2025.
  14. **Andoni, A., P. Indyk, and I. Razenshteyn** (2018). Approximate nearest neighbor search in high dimensions.
  15. **Arora, D. and H. G. Singh** (2023). Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models. *arXiv preprint*.
  16. **Arora, D., H. G. Singh, and Mausam** (2023). Have LLMs Advanced Enough? A Challenging Problem Solving Benchmark For Large Language Models.
  17. **Awasthi, A., S. Sarawagi, R. Goyal, S. Ghosh, and V. Piratla**, Parallel iterative edit models for local sequence transduction. *In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

18. **Balachandran, V., A. Pagnoni, J. Y. Lee, D. Rajagopal, J. Carbonell, and Y. Tsvetkov** (2020). StructSum: Summarization via structured representations. *arXiv preprint arXiv:2003.00576*.
19. **Baltrusaitis, T., A. Zadeh, Y. C. Lim, and L.-P. Morency**, Openface 2.0: Facial behavior analysis toolkit. *In 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*. IEEE, 2018.
20. **Banerjee, S. and A. Lavie**, Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *In Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. 2005.
21. **Bays, C.**, Introduction to cellular automata and conway’s game of life. *In Game of Life Cellular Automata*. Springer, 2010, 1–7.
22. **Beltagy, I., M. E. Peters, and A. Cohan** (2020). Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.
23. **Bertasius, G., H. Wang, and L. Torresani**, Is Space-Time Attention All You Need for Video Understanding? *In Proceedings of the International Conference on Machine Learning (ICML)*. 2021.
24. **Bhagavatula, C., S. Feldman, R. Power, and W. Ammar** (2018). Content-based citation recommendation. *arXiv preprint arXiv:1802.08301*.
25. **Bradbury, N. A.** (2016). Attention span during lectures: 8 seconds, 10 minutes, or more?
26. **Brown, T., B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei** (2020). Language models are few-shot learners.
27. **Bryant, C., M. Felice, Ø. E. Andersen, and T. Briscoe**, The BEA-2019 shared task on grammatical error correction. *In Proceedings of the Fourteenth Workshop on Innovative*

- Use of NLP for Building Educational Applications*. Association for Computational Linguistics, Florence, Italy, 2019. URL <https://aclanthology.org/W19-4406>
28. **Bryant, C., M. Felice, and T. Briscoe**, Automatic annotation and evaluation of error types for grammatical error correction. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Vancouver, Canada, 2017. URL <https://aclanthology.org/P17-1074>
  29. **Bryant, S.** (2023). Assessing GPT-4’s Role as a Co-Collaborator in Scientific Research: A Case Study Analyzing Einstein’s Special Theory of Relativity.
  30. **Chen, D., C. Lee, Y. Lu, D. Rosati, and Z. Yu** (2023). Mixture of soft prompts for controllable data generation.
  31. **Chen, H., H. Takamura, and H. Nakayama** (2021). Scixgen: A scientific paper dataset for context-aware text generation. *arXiv preprint arXiv:2110.10774*.
  32. **Chen, J., J. Tang, J. Qin, X. Liang, L. Liu, E. P. Xing, and L. Lin** (2022). GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning.
  33. **Chen, J. and H. Zhuge** (2014). Summarization of scientific documents by detecting common facts in citations. *Future Generation Computer Systems*, **32**, 246–252.
  34. **Chen, J. and H. Zhuge** (2019). Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*, **31**(3), e4261.
  35. **Chiang, W., Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. Gonzalez, I. Stoica, and E. Xing** (2023). Vicuna: An open-source chatbot impressing gpt-4 with 90 <https://lmsys.org/blog/2023-03-30-vicuna/>.
  36. **Chollampatt, S., W. Wang, and H. T. Ng**, Cross-sentence grammatical error correction. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019.
  37. **Chopard, B. and M. Droz** (1998). Cellular automata. *Modelling of Physical*, 6–13.

38. **Christiano, P. F., J. Leike, T. Brown, M. Martic, S. Legg, and D. Amodei** (2017). Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, **30**.
39. **Cobbe, K., V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman** (2021). Training verifiers to solve math word problems. *ArXiv Preprint ArXiv:2110.14168*.
40. **Cobbe, K., V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman** (2023). Training Verifiers to Solve Math Word Problems.
41. **Cohan, A., W. Ammar, M. van Zuylen, and F. Cady**, Structural scaffolds for citation intent classification in scientific publications. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2019. URL <https://aclanthology.org/N19-1361>
42. **Dahlmeier, D. and H. T. Ng**, Better evaluation for grammatical error correction. *In Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Montréal, Canada, 2012. URL <https://aclanthology.org/N12-1067>.
43. **Dahlmeier, D., H. T. Ng, and S. M. Wu**, Building a large annotated corpus of learner english: The nus corpus of learner english. *In Proceedings of the eighth workshop on innovative use of NLP for building educational applications*. 2013.
44. **Deng, J., W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei**, Imagenet: A large-scale hierarchical image database. *In 2009 IEEE Conference on Computer Vision and Pattern Recognition*. 2009.
45. **Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer** (2023). QLoRA: Efficient Finetuning of Quantized LLMs.
46. **Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova**, Bert: Pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Confer-*

- ence of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 2019a. URL <https://aclanthology.org/N19-1423>
47. **Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova** (2019b). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.
  48. **Dhall, A., G. Sharma, R. Goecke, and T. Gedeon**, Emotiw 2020: Driver gaze, group emotion, student engagement and physiological signal based challenges. *In Proceedings of the 2020 International Conference on Multimodal Interaction*. 2020.
  49. **Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al.** (2020a). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
  50. **Dosovitskiy, A., L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al.** (2020b). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.
  51. **Feng, K., Y. Zhao, Y. Liu, T. Yang, C. Zhao, J. Sous, and A. Cohan** (2025). Physics: Benchmarking foundation models on university-level physics problem solving. *arXiv preprint arXiv:2503.21821*.
  52. **Gao, J., R. Pi, J. Zhang, J. Ye, W. Zhong, Y. Wang, L. Hong, J. Han, H. Xu, Z. Li, et al.** (2023). G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.
  53. **Ge, Y., L. Dinh, X. Liu, J. Su, Z. Lu, A. Wang, and J. Diesner**, BACO: A Background Knowledge- and Content-Based Framework for Citing Sentence Generation. *In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 2021. URL <https://aclanthology.org/2021.acl-long.116>.

54. **Goel, A., M. Hira, A. Anand, S. Bangar, and R. R. Shah** (2023). Advancements in scientific controllable text generation methods. URL <https://arxiv.org/abs/2307.05538>.
55. **Goldberg, P., Ö. Sümer, K. Stürmer, W. Wagner, R. Göllner, P. Gerjets, E. Kasneci, and U. Trautwein** (2021). Attentive or not? toward a machine learning approach to assessing students' visible engagement in classroom instruction. *Educational Psychology Review*, **33**, 27–49.
56. **Grover, A. and J. Leskovec**, node2vec: Scalable feature learning for networks. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 2016. URL <https://doi.org/10.1145/2939672.2939754>.
57. **Gu, J., K. Cho, and V. O. Li** (2017). Trainable greedy decoding for neural machine translation. *arXiv preprint arXiv:1702.02429*.
58. **Gu, N., E. Ash, and R. Hahnloser**, MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2022. URL <https://aclanthology.org/2022.acl-long.450>.
59. **Gu, N. and R. H. R. Hahnloser** (2022). Controllable Citation Text Generation.
60. **Gupta, A., A. D’Cunha, K. Awasthi, and V. Balasubramanian** (2016). Daisee: Towards user engagement recognition in the wild. *arXiv preprint arXiv:1609.01885*.
61. **Hoang, C. D. V. and M.-Y. Kan**, Towards automated related work summarization. In *Proceedings of the 23rd Int’l Conf. on Computational Linguistics: Posters, COLING ’10*. Association for Computational Linguistics, USA, 2010.
62. **Holtzman, A., J. Buys, L. Du, M. Forbes, and Y. Choi** (2019). The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*.
63. **Hu, E., Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen** (2021). Lora: Low-rank adaptation of large language models.

64. **Hu, Y.** and **X. Wan**, Automatic generation of related work sections in scientific papers: An optimization approach. In **A. Moschitti, B. Pang,** and **W. Daelemans** (eds.), *Proceedings of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 2014. URL <https://aclanthology.org/D14-1170>
65. **Ji, S., S. Pan, E. Cambria, P. Marttinen,** and **P. S. Yu** (2022). A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, **33**(2), 494–514.
66. **Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. I. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al.** (2023). Mistral 7B. *arXiv preprint arXiv:2310.06825*.
67. **Junczys-Dowmunt, M.** and **R. Grundkiewicz**, Phrase-based machine translation is state-of-the-art for automatic grammatical error correction. In **J. Su, K. Duh,** and **X. Carreras** (eds.), *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, 2016. URL <https://aclanthology.org/D16-1161/>
68. **Jung, S.-Y., T.-H. Lin, C.-H. Liao, S.-M. Yuan,** and **C.-T. Sun** (2022). Intent-Controllable Citation Text Generation. *Mathematics*, **10**, 1763.
69. **Karpukhin, V., B. Oğuz, S. Min, P. Lewis, L. Wu, S. Edunov, D. Chen,** and **Wen-Yih** (2020). Dense passage retrieval for open-domain question answering.
70. **Katar, O., D. Ozkan, GPT, O. Yildirim,** and **U. R. Acharya** (2022). Evaluation of GPT-3 AI language model in research paper writing.
71. **Kaur, A., A. Mustafa, L. Mehta,** and **A. Dhall**, Prediction and localization of student engagement in the wild. In *2018 Digital Image Computing: Techniques and Applications (DICTA)*. IEEE, 2018.
72. **Kingma, D. P.** and **J. Ba** (2014a). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
73. **Kingma, D. P.** and **J. Ba** (2014b). Adam: A Method for Stochastic Optimization. *CoRR*, **abs/1412.6980**, 1–15. URL <https://api.semanticscholar.org/CorpusID:6628106>.

74. **Kobayashi, M., M. Mita, and M. Komachi** (2024). Revisiting meta-evaluation for grammatical error correction. *Transactions of the Association for Computational Linguistics*, **12**, 837–855. ISSN 2307-387X. URL [https://doi.org/10.1162/tacl\\_a\\_00676](https://doi.org/10.1162/tacl_a_00676)
75. **Kollepara, N., S. Chatakonda, and P. Kumar** (2021). Scimat: Science and mathematics dataset.
76. **Koncel-Kedziorski, R., D. Bekal, Y. Luan, M. Lapata, and H. Hajishirzi** (2019). Text generation from knowledge graphs with graph transformers. *arXiv preprint arXiv:1904.02342*.
77. **Koncel-Kedziorski, R., S. Roy, A. Amini, N. Kushman, and H. Hajishirzi**, MAWPS: A math word problem repository. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, 2016. URL <https://aclanthology.org/N16-1136>
78. **Kumar, D., S. Madan, P. Singh, A. Dhall, and B. Raman**, Towards engagement prediction: A cross-modality dual-pipeline approach using visual and audio features. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM '24*. Association for Computing Machinery, New York, NY, USA, 2024. ISBN 9798400706868. URL <https://doi.org/10.1145/3664647.3688986>.
79. **Kumar, V., R. Maheshwary, and V. Pudi** (2022). Practice makes a solver perfect: Data augmentation for math word problem solvers.
80. **Le, N., K. Nguyen, A. Nguyen, and B. Le** (2022). Global-local attention for emotion recognition. *Neural Computing and Applications*, **34**(24), 21625–21639.
81. **Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer**, BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2020. URL <https://aclanthology.org/2020.acl-main.703>.

82. **Li, J., D. Li, S. Savarese, and S. Hoi**, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *In International conference on machine learning*. PMLR, 2023.
83. **Li, J., D. Li, C. Xiong, and S. Hoi**, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *In International conference on machine learning*. PMLR, 2022.
84. **Li, J., Y. Lu, Y. Yang, D. Li, L. Chen, and L. Carin** (2021). Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*.
85. **Li, W. and H. Wang**, Detection-correction structure via general language model for grammatical error correction. *In L.-W. Ku, A. Martins, and V. Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 2024. URL <https://aclanthology.org/2024.acl-long.96>
86. **Li, W., X. Xiao, Y. Lyu, and Y. Wang**, Improving neural abstractive document summarization with explicit information selection modeling. *In E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii (eds.), Proceedings of the 2018 Conf. on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2018. URL <https://aclanthology.org/D18-1205>.
87. **Li, X., Y.-H. Lee, and J. Ouyang**, Cited text spans for scientific citation text generation. *In T. Ghosal, A. Singh, A. Waard, P. Mayr, A. Naik, O. Weller, Y. Lee, S. Shen, and Y. Qin (eds.), Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*. Association for Computational Linguistics, Bangkok, Thailand, 2024. URL <https://aclanthology.org/2024.sdp-1.9>.
88. **Liang, Z., K. Guo, G. Liu, T. Guo, Y. Zhou, T. Yang, J. Jiao, R. Pi, J. Zhang, and X. Zhang**, SceMQA: A scientific college entrance level multimodal question answering benchmark. *In L.-W. Ku, A. Martins, and V. Srikumar (eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Bangkok, Thailand, 2024. URL <https://aclanthology.org/2024.acl-short.11/>.
89. **Lin, C.**, Rouge: A package for automatic evaluation of summaries. *In Text Summarization Branches Out*. 2004a.

90. **Lin, C.-Y.**, ROUGE: A Package for Automatic Evaluation of Summaries. *In Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 2004b. URL <https://aclanthology.org/W04-1013>.
91. **Liu, H., C. Li, Q. Wu, and Y. J. Lee** (2024). Visual instruction tuning. *Advances in Neural Information Processing Systems*, **36**.
92. **Liu, X., Y. Zheng, Z. Du, M. Ding, Y. Qian, Z. Yang, and J. Tang** (2021). GPT understands, too. *arXiv preprint arXiv:2103.10385*.
93. **Liu, Y. and M. Lapata** (2018). Learning structured text representations. *Transactions of the Association for Computational Linguistics*, **6**(2), 63–75.
94. **Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov** (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
95. **Lo, K., L. L. Wang, M. Neumann, R. Kinney, and D. Weld**, S2orc: The semantic scholar open research corpus. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2020. URL <https://aclanthology.org/2020.acl-main.447>.
96. **Lo, K., L. L. Wang, M. Neumann, R. Kinney, and D. S. Weld** (2019). S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.
97. **Lu, P., R. Gong, S. Jiang, L. Qiu, S. Huang, X. Liang, and S.-C. Zhu** (2021). InterGPS: Interpretable geometry problem solving with formal language and symbolic reasoning. *arXiv preprint arXiv:2105.04165*.
98. **Luo, H., Q. Sun, C. Xu, P. Zhao, J. Lou, C. Tao, X. Geng, Q. Lin, S. Chen, and D. Zhang** (2023). Wizardmath: Empowering mathematical reasoning for large language models via reinforced evol-instruct.
99. **Luu, K., X. Wu, R. Koncel-Kedziorski, K. Lo, I. Cachola, and N. A. Smith** (2020). Explaining relationships between scientific documents. *arXiv preprint arXiv:2002.00317*.
100. **Malmi, E., S. Krause, S. Rothe, D. Mirylenka, and A. Severyn**, Encode, tag, realize: High-precision text editing. *In Proceedings of the 2019 Conference on Empirical*

*Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2019.

101. **Mandlecha, P., S. K. Chatakonda, N. Kollepara, and P. Kumar** (2023). Hybrid Tokenization and Datasets for Solving Mathematics and Science Problems Using Transformers.
102. **Miao, S., C. Liang, and K. Su** (2021). A diverse corpus for evaluating and developing english math word problem solvers.
103. **Michael Denkowski, A. L.** (2014). Meteor Universal: Language Specific Translation Evaluation for Any Target Language.
104. **Mishra, S., A. Mitra, N. Varshney, B. Sachdeva, P. Clark, C. Baral, and A. Kalyan,** NumGLUE: A suite of fundamental yet challenging mathematical reasoning tasks. *In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Dublin, Ireland, 2022. URL <https://aclanthology.org/2022.acl-long.246>.
105. **Mordvintsev, A., E. Randazzo, E. Niklasson, and M. Levin** (2020). Growing neural cellular automata. *Distill*, 5(2), e23.
106. **Nallapati, R., B. Zhou, C. Gulcehre, B. Xiang, et al.** (2016). Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.
107. **Narayan, S., S. B. Cohen, and M. Lapata** (2018). Don't give me the details, just the summary! topic-aware convolutional neural networks for extreme summarization. *arXiv preprint arXiv:1808.08745*.
108. **Ng, H. T., S. M. Wu, T. Briscoe, C. Hadiwinoto, R. H. Susanto, and C. Bryant,** The conll-2014 shared task on grammatical error correction. *In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task*. Association for Computational Linguistics, 2014. URL <http://www.aclweb.org/anthology/W14-1701>.
109. **Omelianchuk, K., V. Atrasevych, A. Chernodub, and O. Skurzshanskyi,** Gector—grammatical error correction: Tag, not rewrite. *In Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. 2020.

110. **Pan, S., L. Luo, Y. Wang, C. Chen, J. Wang, and X. Wu** (2023). Unifying large language models and knowledge graphs: A roadmap.
111. **Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu**, Bleu: a Method for Automatic Evaluation of Machine Translation. *In Annual Meeting of the Association for Computational Linguistics*. 2002a.
112. **Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu**, BLEU: a Method for Automatic Evaluation of Machine Translation. *In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002b. URL <https://aclanthology.org/P02-1040>.
113. **Perozzi, B., R. Al-Rfou, and S. Skiena**, Deepwalk: online learning of social representations. *In Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Association for Computing Machinery, New York, NY, USA, 2014. URL <https://doi.org/10.1145/2623330.2623732>.
114. **Radford, A., J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al.**, Learning transferable visual models from natural language supervision. *In International Conference on Machine Learning*. PMLR, 2021.
115. **Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, et al.** (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, **21**(140), 1–67.
116. **Ranzato, M., S. Chopra, M. Auli, and W. Zaremba** (2015). Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.
117. **Romanov, A., A. Rumshisky, A. Rogers, and D. Donahue** (2018). Adversarial decomposition of text representation. *arXiv preprint arXiv:1808.09042*.
118. **Rozovskaya, A. and D. Roth**, Grammatical error correction: Machine translation and classifiers. *In K. Erk and N. A. Smith* (eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, 2016. URL <https://aclanthology.org/P16-1208/>

119. **Saadany, H.** and **C. Orăsan**, Bleu, meteor, bertscore: Evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text. *In Proceedings of the Translation and Interpreting Technology Online Conference TRITON 2021*. 2021.
120. **Sathayanarayana, S., R. Kumar Satzoda, A. Carini, M. Lee, L. Salamanca, J. Reilly, D. Forster, M. Bartlett, and G. Littlewort**, Towards automated understanding of student-tutor interactions using visual deictic gestures. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 2014.
121. **Saxton, D., E. Grefenstette, F. Hill, and P. Kohli** (2019). Analysing mathematical reasoning abilities of neural models.
122. **See, A., P. J. Liu, and C. D. Manning** (2017). Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.
123. **Sennrich, R., B. Haddow, and A. Birch** (2016). Edinburgh neural machine translation systems for wmt 16. *arXiv preprint arXiv:1606.02891*.
124. **Shin, T., Y. Razeghi, R. L. Logan IV, E. Wallace, and S. Singh** (2020). Autoprompt: Eliciting knowledge from language models with automatically generated prompts. *arXiv preprint arXiv:2010.15980*.
125. **Singh, M., X. Hoque, D. Zeng, Y. Wang, K. Ikeda, and A. Dhall** (2023). Do I Have Your Attention: A Large Scale Engagement Prediction Dataset and Baselines. *arXiv preprint arXiv:2302.00431*.
126. **Tajiri, T., M. Komachi, and Y. Matsumoto**, Tense and aspect error correction for ESL learners using global context. *In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Jeju Island, Korea, 2012. URL <https://aclanthology.org/P12-2039>.
127. **Tan, M. and Q. Le**, Efficientnet: Rethinking model scaling for convolutional neural networks. *In International conference on machine learning*. PMLR, 2019.
128. **Taori, R., I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto** (2023). Stanford alpaca: An instruction-following llama model. [https://github.com/tatsu-lab/stanford\\_alpaca](https://github.com/tatsu-lab/stanford_alpaca)

129. **Taylor, R., M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic** (2022). Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.
130. **Team, G., R. Anil, S. Borgeaud, Y. Wu, J.-B. Alayrac, J. Yu, R. Soricut, J. Schalkwyk, A. M. Dai, A. Hauth, et al.** (2023). Gemini: a family of highly capable multi-modal models. *arXiv preprint arXiv:2312.11805*.
131. **Touvron, H., T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample** (2023a). Llama: Open and efficient foundation language models.
132. **Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample** (2023b). Llama: Open and efficient foundation language models.
133. **Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Kaiser, and I. Polosukhin** (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.
134. **Wang, F., M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang**, Residual attention network for image classification. *In Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.
135. **Wang, J., Z. Yang, X. Hu, L. Li, K. Lin, Z. Gan, Z. Liu, C. Liu, and L. Wang** (2022). Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.
136. **Wang, Y., Y. Kordi, S. Mishra, A. Liu, N. A. Smith, D. Khashabi, and H. Hajishirzi** (2023). Self-Instruct: Aligning Language Models with Self-Generated Instructions.
137. **Wei, J., X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, and D. Zhou** (2022). Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.
138. **Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou** (2023). Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.

139. **Welbl, J., N. F. Liu, and M. Gardner** (2017). Crowdsourcing Multiple Choice Science Questions.
140. **Welleck, S., I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston** (2019). Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.
141. **Whitehill, J., Z. Serpell, Y.-C. Lin, A. Foster, and J. R. Movellan** (2014). The faces of engagement: Automatic recognition of student engagement from facial expressions. *IEEE Transactions on Affective Computing*, **5**(1), 86–98.
142. **Wu, J.-Y., A. T.-W. Shieh, S.-J. Hsu, and Y.-N. Chen** (2021). Towards generating citation sentences for multiple references with intent control. *arXiv preprint arXiv:2112.01332*.
143. **Xing, X., X. Fan, and X. Wan**, Automatic generation of citation texts in scholarly papers: A pilot study. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020a.
144. **Xing, X., X. Fan, and X. Wan**, Automatic generation of citation texts in scholarly papers: A pilot study. *In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 2020b. URL <https://aclanthology.org/2020.acl-main.550>.
145. **Yang, Z., Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le** (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, **32**.
146. **Yannakoudakis, H., T. Briscoe, and B. Medlock**, A new dataset and method for automatically grading ESOL texts. *In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Portland, Oregon, USA, 2011. URL <https://aclanthology.org/P11-1019>
147. **Yao, S., D. Yu, J. Zhao, I. Shafran, T. Griffiths, Y. Cao, and K. Narasimhan** (2023). Tree of thoughts: Deliberate problem solving with large language models.
148. **Yasunaga, M., J. Kasai, R. Zhang, A. R. Fabbri, I. Li, D. Friedman, and D. R. Radev**, Scisummnet: A large annotated corpus and content-impact models for scientific

- paper summarization with citation networks. *In Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33. 2019.
149. **Ye, D., Y. Lin, P. Li, and M. Sun**, Packed levitated marker for entity and relation extraction. *In S. Muresan, P. Nakov, and A. Villavicencio (eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022.* Association for Computational Linguistics, 2022. URL <https://aclanthology.org/2022.acl-long.337>.
  150. **Yuan, Z. and C. Bryant**, Document-level grammatical error correction. *In J. Burstein, A. Horbach, E. Kochmar, R. Laarmann-Quante, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis, and T. Zesch (eds.), Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications.* Association for Computational Linguistics, Online, 2021a. URL <https://aclanthology.org/2021.bea-1.8/>.
  151. **Yuan, Z. and C. Bryant**, Document-level grammatical error correction. *In Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications.* 2021b.
  152. **Yuan, Z., H. Yuan, C. Tan, W. Wang, and S. Huang (2023)**. How well do large language models perform in arithmetic tasks?
  153. **Yue, X., X. Qu, G. Zhang, Y. Fu, W. Huang, H. Sun, Y. Su, and W. Chen (2023)**. Mammoth: Building math generalist models through hybrid instruction tuning. *arXiv preprint arXiv:2309.05653*.
  154. **Zaletelj, J. and A. Košir (2017)**. Predicting students' attention in the classroom from kinect facial and body features. *EURASIP journal on image and video processing*, 2017(1), 1–12.
  155. **Zhang, J., Z.-Z. Li, M.-L. Zhang, F. Yin, C.-L. Liu, and Y. Moshfeghi**, GeoEval: Benchmark for evaluating LLMs and multi-modal models on geometry problem-solving. *In L.-W. Ku, A. Martins, and V. Srikumar (eds.), Findings of the Association for Computational Linguistics: ACL 2024.* Association for Computational Linguistics, Bangkok, Thailand, 2024. URL <https://aclanthology.org/2024.findings-acl.73/>

156. **Zhang, J., Y. Zhao, M. Saleh, and P. J. Liu**, PEGASUS: pre-training with extracted gap-sentences for abstractive summarization. *In Proceedings of the 37th International Conference on Machine Learning, ICML'20*. JMLR.org, JMLR.org, 2020a.
157. **Zhang, M.-L., F. Yin, and C.-L. Liu** (2023). A Multi-Modal Neural Geometric Solver with Textual Clauses Parsed from Diagram. *arXiv preprint arXiv:2302.11097*.
158. **Zhang, T., V. Kishore, F. Wu, K. Weinberger, and Y. Artzi** (2020b). Bertscore: Evaluating text generation with bert.
159. **Zhang, Z., A. Zhang, M. Li, and A. Smola** (2022). Automatic chain of thought prompting in large language models. *arXiv preprint arXiv:2210.03493, abs/2210.03493*.
160. **Zhao, T. Z., E. Wallace, S. Feng, D. Klein, and S. Singh** (2021). Calibrate Before Use: Improving Few-Shot Performance of Language Models.
161. **Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen** (2023). A survey of large language models.
162. **Zheng, H. S., S. Mishra, X. Chen, H.-T. Cheng, E. H. Chi, Q. V. Le, and D. Zhou** (2023). Take a step back: Evoking reasoning via abstraction in large language models. *arXiv preprint arXiv:2310.06117*.
163. **Zheng, L., W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al.** (2024). Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, **36**.
164. **Zhu, Y., X. Wang, J. Chen, S. Qiao, Y. Ou, Y. Yao, S. Deng, H. Chen, and N. Zhang** (2023). Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities.
165. **Ziegler, D. M., N. Stiennon, J. Wu, T. B. Brown, A. Radford, D. Amodei, P. Christiano, and G. Irving** (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.
166. **Zong, Z., B. Ma, D. Shen, G. Song, H. Shao, D. Jiang, H. Li, and Y. Liu**, MoVA: Adapting mixture of vision experts to multimodal context. *In The Thirty-eighth An-*

*nual Conference on Neural Information Processing Systems*. 2024. URL <https://openreview.net/forum?id=uHs6RJFDsg>

167. **Çaglar Gülçehre, O. Firat, K. Xu, K. Cho, L. Barrault, H.-C. Lin, F. Bougares, H. Schwenk, and Y. Bengio** (2015). On using monolingual corpora in neural machine translation. *ArXiv*, **abs/1503.03535**, 1–9. URL <https://api.semanticscholar.org/CorpusID:15352384>.