

GeoCoder: Solving Geometry Problems by Generating Modular Code through Vision-Language Models

Aditya Sharma^{1,2}, Aman Dalmia^{1,3}, Mehran Kazemi⁵,
Amal Zouaq^{1,2}, Christopher J. Pal^{1,2,4}

¹Mila, ²Polytechnique Montréal, ³Université de Montréal,

⁴Canada CIFAR AI Chair, ⁵Google DeepMind

Correspondence: aditya.sharma@mila.quebec

Abstract

Geometry problem-solving demands advanced reasoning abilities to process multimodal inputs and employ mathematical knowledge effectively. Vision-language models (VLMs) have made significant progress in various multimodal tasks. Yet, they still struggle with geometry problems and are significantly limited by their inability to perform mathematical operations not seen during pre-training, such as calculating the cosine of an arbitrary angle, and by difficulties in correctly applying relevant geometry formulas. To overcome these challenges, we present GeoCoder, which leverages modular code-finetuning to generate and execute code using a predefined geometry function library. By executing the code, we achieve accurate and deterministic calculations, contrasting the stochastic nature of autoregressive token prediction, while the function library minimizes errors in formula usage. We also propose a multimodal retrieval-augmented variant of GeoCoder, named RAG-GeoCoder, which incorporates a non-parametric memory module for retrieving functions from the geometry library, thereby reducing reliance on parametric memory. Our modular code-finetuning approach enhances the geometric reasoning capabilities of VLMs, yielding an average improvement of over 16% across various question complexities on the GeomVerse dataset compared to other fine-tuning methods.

1 Introduction

Geometry problem solving requires interpreting a figure, comprehending a question, and applying the appropriate geometry formulas to derive the answer. This presents an ideal test-bed for evaluating multimodal models' visual and mathematical reasoning capabilities, with practical relevance in fields such as education, architecture, and engineering.

Various vision-language models (VLMs) like GPT-4V (Achiam et al., 2023), GPT-4o (OpenAI, 2024), Gemini (Team et al., 2024), PaLI (Chen

et al., 2022b) and LLaVA (Liu et al., 2024) have been introduced and benchmarked on tasks involving geometric reasoning (Gao et al., 2023; Kazemi et al., 2023). Despite this progress, the inherent nature of auto-regressive language model training, which focuses on next-token prediction, presents challenges in modeling mathematical reasoning that requires precise calculations and the correct application of formulas (Kazemi et al., 2023; Liao et al., 2024). While language models can handle computations they were exposed to during training, such as $\cos 45$, they tend to struggle with unseen calculations, for instance, $\cos 18$. The Chain-of-Thought (CoT) approach aims to improve reasoning in language models by eliciting a step-by-step natural language reasoning along with the final answer (Wei et al., 2022) but suffers from the same problems of imprecise computations and wrong formula usage (Gou et al., 2023; Kazemi et al., 2023). This problem is especially aggravated when the geometry problems require multi-hop mathematical reasoning over figures with multiple complex shapes like in the GeomVerse dataset (Kazemi et al., 2023), as seen in figure 1.

Mathematical code generation, where a model generates code that is then executed to obtain the solution to a mathematical problem, addresses the issue of inaccurate calculations (Chen et al.; Wang et al., 2023; Liao et al., 2024). Code execution ensures precise and deterministic results, unlike the stochastic behavior of autoregressive next-token prediction. However, this method does not fully resolve the issue of incorrect formula application, as the language model relies on its parametric memory to recall and implement the relevant formulas.

In this study, we answer the following research question: "Can VLMs enhance their multimodal mathematical reasoning capabilities by using modular code generation and execution instead of relying on autoregressive answer generation?" We

propose GeoCoder¹, a multi-modal modular code generation approach for geometric reasoning using VLMs, which generates code that includes calls to predefined mathematical functions and mitigates this limitation by offloading the responsibility of recalling formulas to a function library. Modular code generation offers the additional advantage of improving interpretability by including templated print statements in the function outputs. Solving most geometry questions involves applying multiple functions in sequence, which produces a series of templated explanations outlining each step of the problem-solving process.

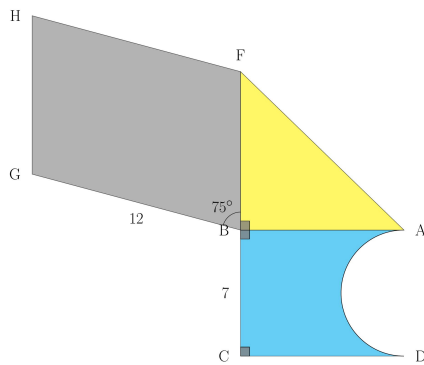


Figure 1: Sample geometry problem from the GeomVerse (Kazemi et al., 2023) dataset. **Question:** If the ABCD shape is a rectangle where a semi-circle has been removed from one side of it, the area of the BAF right triangle is 40, and the area of the BGHF parallelogram is 102, compute the area of the ABCD shape.

While recent LLMs (large language models) have demonstrated remarkable performance in Python code generation tasks (Jiang et al., 2024), VLMs still encounter difficulties, largely due to the lack of datasets pairing images and questions with code. To address this gap, we capitalize on the code generation capabilities of LLMs to construct the first multimodal code instruction-tuning dataset for solving geometry problems and subsequently finetune a VLM on this data to create GeoCoder. In addition, we introduce a retrieval-augmented version of our model, named RAG-GeoCoder, which leverages a non-parametric function memory and generates modular code that uses functions more often than GeoCoder (see Section 5.4). Given its reliance on non-parametric memory, RAG-GeoCoder has the potential to adapt to both modifications in the function library and significantly larger libraries. However, we leave further investigation of

¹Code and data available at <https://github.com/Lama-West/Geocoder>

this capability for future work.

This work makes the following contributions:

1. Our work presents a modular code-finetuning framework for VLM-based geometry problem-solving utilizing a predefined function library. The VLM trained via this process is referred to as GeoCoder.
2. We propose a retrieval-augmented variant of GeoCoder, called RAG-GeoCoder, which includes a non-parametric memory to reduce the VLM’s reliance on its parametric memory for recalling geometry functions, thus increasing function usage.
3. GeoCoder and RAG-GeoCoder outperform models trained with alternate finetuning strategies on the GeomVerse dataset by more than 16% on average across question complexities and by 14.2% on the GeoQA-NO dataset (see Section 4.1.2).
4. GeoCoder and RAG-GeoCoder add interpretability to VLM-based geometry problem-solving through templated print statements in the proposed geometry function library.
5. We release a first-of-its-kind multimodal dataset for code instruction-tuning for solving geometry problems, with geometry questions and images paired with corresponding code generated by our approach. This dataset contains 35k data points from GeomVerse (Kazemi et al., 2023) and 48k from Geo170k (Gao et al., 2023).

2 Related Work

In this section, we discuss a literature review of existing approaches.

2.1 VLMs for solving geometry problems

Recent advancements have seen the development of various general-purpose VLMs (Liu et al., 2024; OpenAI, 2024; Team et al., 2024; Li et al., 2024; Han et al., 2024), but they continue to underperform on geometry tasks. Gao et al. (2023) and Zhang et al. (2024b) argue that VLMs struggle to comprehend basic geometric visual elements accurately and release two large geometry instruction-tuning datasets called Geo170k and MAVIS, respectively. Kazemi et al. (2023) release a new

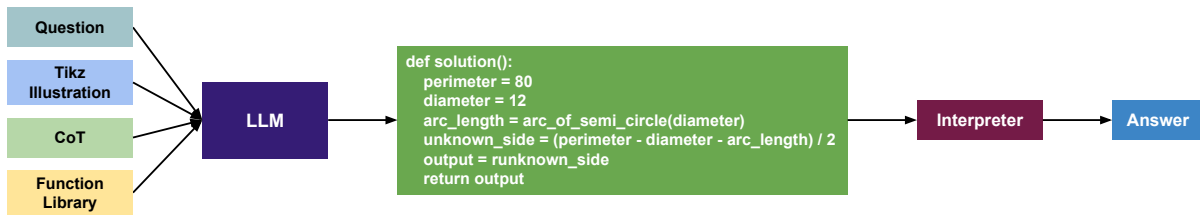


Figure 2: The first step in our methodology consists of generating modular code by employing few-shot prompting with a code generation-capable LLM, utilizing questions, TikZ image illustrations, CoT reasoning, and the predefined function library. The generations that execute to produce the correct answer are selected as the basis for our "gold" code-tuning data, as discussed in Section 3.1.

large synthetic dataset called GeomVerse for solving multi-hop geometry problems that require complex reasoning. They benchmark state-of-the-art VLMs on this dataset and show that these models are not as capable in subjects like geometry, which require long chains of reasoning. Our geometry code-tuning dataset is built using the Geo170k and GeomVerse datasets. Additionally, AlphaGeometry (Trinh et al., 2024) introduces a neuro-symbolic system that solves geometry theorem-proving problems. In contrast to our model which generates modular code through a VLM, AlphaGeometry’s generation combines a VLM with a rule-bound deduction engine. GeoGPT4V (Cai et al., 2024) is a dataset that contains 4.9k basic geometry questions and images in the style of GeoQA (Chen et al., 2021) and Geometry3k (Lu et al., 2021) generated by GPT4V + 19k questions from existing datasets like GeoQA. We use Geo170k (Gao et al., 2023) in this work, a much larger dataset augmented from GeoQA and Geometry3k, for finetuning our models to generate code.

2.2 Code generating LLMs for non-visual mathematical reasoning

Wang et al. (2023) and Liao et al. (2024) suggest a methodology for producing new datasets comprising textual math problems and their code-based solutions using large LLMs as teachers for smaller LLMs. We modify this knowledge-distillation approach to address multimodal mathematical reasoning by using a teacher LLM to generate our geometry code instruction-tuning dataset. (see Section 3.1). Additionally, we extend this approach to modular code generation by incorporating a predefined geometry function library, ensuring the model applies precise mathematical formulas instead of relying on its parametric memory. Gou et al. (2023) propose text-only mathematical reasoning agents that interweave natural language rationale genera-

tion with program-based tool use, this is in contrast to our multimodal approach where the VLM generates modular code which is executed to get the answer as well as the rationale through templated print statements.

2.3 Code generating VLMs for general VQA

While there have been approaches like ViperGPT (Surfís et al., 2023) and VizProg (Gupta and Kembhavi, 2023) for generating code for tasks like visual question answering (VQA), they do so using LLMs, ignoring the image information completely and only looking at the question. Subsequently, they add the image information while executing the generated code. In contrast, our approach builds a VLM that looks at the geometry images and incorporates this information with the question to generate code for solving geometry problems. Multimodal Self-Instruct (Zhang et al., 2024c) uses an LLM to generate code to synthesize a benchmark with abstract images and visual reasoning instructions. This paper reinforces our observation that VLMs are not good coders because of the lack of paired image-plus-question and code datasets, resulting in a lack of training to condition VLMs to generate code while looking at both the text and the image.

3 Proposed Method

Existing approaches that finetune VLMs to generate CoT for geometry problems (Gao et al., 2023; Kazemi et al., 2023) help VLMs become better geometric reasoners, as compared to finetuning VLMs to generate just the final answer. Even so, these CoT finetuned VLMs significantly suffer from wrong calculation and wrong formula usage errors, especially for geometry problems requiring long chains of reasoning (Kazemi et al., 2023), due to the autoregressive nature of VLM language

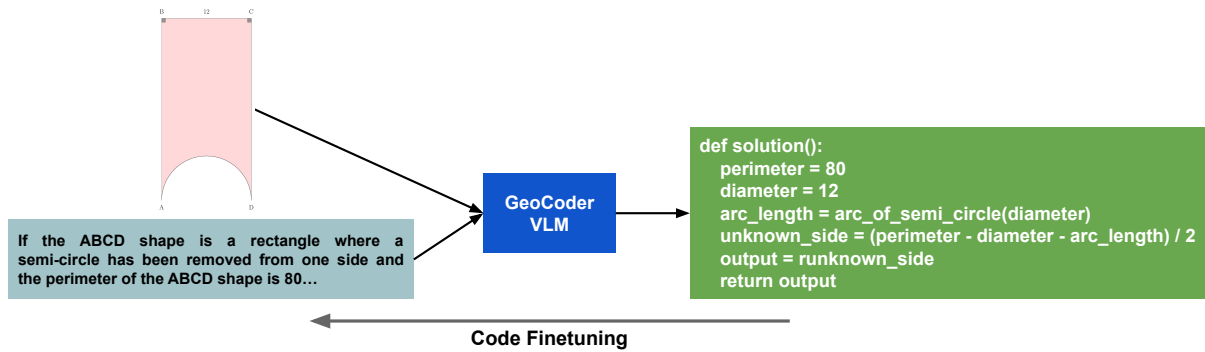


Figure 3: During modular code-finetuning, we utilize the code-tuning data produced by our teacher LLM (see Section 3.1) to finetune a significantly smaller VLM, which we refer to as GeoCoder (as discussed in Section 3.2).

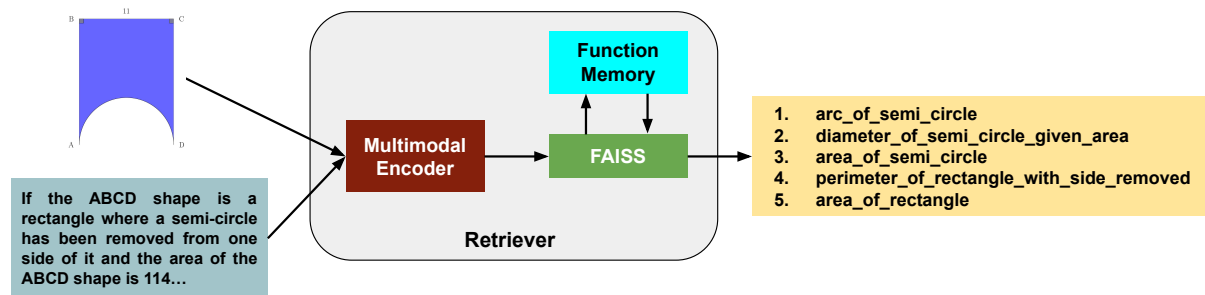


Figure 4: For each geometry problem, given the image and question text, our multimodal retriever retrieves the most similar functions from the function memory, as discussed in Section 3.3.1.

generation. To address these limitations, we propose modular code-finetuning as an alternative to CoT finetuning for training VLMs to solve geometry problems. The proposed approach, where the VLM generates Python code with calls to modules or functions from a predefined geometry library (discussed in Section 3.1), offers three key advantages: (1) code execution guarantees precise and deterministic calculations; (2) the generated code references predefined mathematical functions, reducing errors related to incorrect formula usage; and (3) using templated print statements in each function enhances the interpretability of the computational process. We propose a two-step code-finetuning strategy, capitalizing on knowledge distillation, described in the following sections.

3.1 Generating geometric code-tuning data

We first generate "gold" code-tuning data by CoT few-shot prompting an LLM skilled in code generation, using the TikZ image illustrations (Tantau, 2013). As illustrated in figure 2, we start by prompting a large LLM to generate multiple Python codes that leverage pre-defined functions to solve geometry questions in a few-shot manner (6-shot in this study). We call this model our teacher LLM. We

employ CoT prompting to guide the model generation, along with the TikZ illustration of the image (Tantau, 2013) (which is a textual representation of a geometry figure, as seen in the example in Figure 8) and a function library. We interpret the code generated by the model and use the answers it gets correct as a source of code-finetuning data for our student VLM described in section 3.2. We define a correct code as one that executes to give a floating point answer that matches the ground truth answer for that geometry question. Codes that execute but whose output does not match the ground truth answer are rejected. We also developed a Python function library that includes all formulas present in the GeomVerse dataset, as well as a selection of commonly applied geometry formulas. Along with code for calculating answers using geometry formulas, the functions in the proposed library also include templated print statements. The final library contains 47 geometry functions, which are listed in the Appendix.

3.2 GeoCoder

We use the modular code-finetuning data generated by our teacher LLM (discussed in Section 3.1) to finetune a much smaller VLM called the student

VLM. As shown in figure 3, the student VLM is finetuned to generate code directly given the image and textual question. We call the resulting code-finetuned student VLM, GeoCoder. At inference time, GeoCoder takes the image and question as input and generates Python code which is interpreted to give the final answer. In contrast to preceding models that generate code using an LLM and add the image information subsequently (Surís et al., 2023; Gupta and Kembhavi, 2023), our methodology is the first one to finetune a VLM specifically to generate code while looking at the image.

3.3 RAG-GeoCoder

Additionally, we propose RAG-GeoCoder which is a retrieval-augmented version of GeoCoder based on the popular retrieval-augmented generation (RAG) paradigm (Lewis et al., 2020). RAG-GeoCoder first employs a multimodal retriever (discussed in Section 3.3.1) to retrieve relevant functions from the function library, which are then provided as input, along with the question and image, to the student VLM during both code-finetuning and inference. This approach allows the student VLM to use a non-parametric memory to select the most appropriate functions, rather than relying on its parametric memory, thus increasing function usage (see Section 5.4).

3.3.1 Multimodal Retriever

Given an image and accompanying question text from a geometry problem, our multimodal retriever identifies the most similar functions from the function memory. As illustrated in Figure 4, the multimodal encoder first encodes the image and text into an embedding, and then FAISS (Douze et al., 2024; Johnson et al., 2019) is used to retrieve a set of functions from a non-parametric memory whose embedding is closest to the input embedding.

Multimodal Encoder. In this work, we utilize the VISTA model (Zhou et al., 2024) as our multimodal encoder, as it embeds text, image, and image-text data into a shared vector space, facilitating dense retrieval across different modalities.

Function Memory. We employ a non-parametric function memory structured as a dictionary, where multimodal encoder-generated embeddings of function names, arguments, and descriptions serve as keys, with the corresponding function names as values. This module stores a dictionary with embeddings and names for all functions in our library.

Dataset	Train	Test
GeomVerse	10k	3k
Geo170k-NO	110k	✗
GeoQA-NO	✗	754
MathVerse (testmini)	✗	3.94k

Table 1: Statistics of datasets used in this work, as discussed in Section 4.1.

3.3.2 Finetuning and inference

RAG-GeoCoder student VLM receives the image, question text and a set of retrieved functions (details in Section 4.3) and learns to generate modular code using the code-finetuning data generated by our teacher LLM (described in Section 3.1). At inference time, we give the image, question and retrieved functions to RAG-GeoCoder, which generates modular code which is executed to get the final answer.

4 Experimental Setup

4.1 Datasets

We evaluate our proposed models on three datasets: GeomVerse (Kazemi et al., 2023), GeoQA (Chen et al., 2021), and MathVerse testmini (Zhang et al., 2024a). Additionally, we use the Geo170k (Gao et al., 2023) geometry instruction-tuning dataset and the train set of GeomVerse to create code-tuning data for finetuning GeoCoder, as described in Section 3.1. An overview of the dataset statistics is outlined in Table 1.

4.1.1 Training Datasets

GeomVerse is a synthetically generated dataset of geometry questions designed to require multi-step mathematical reasoning across both text and images. The dataset categorizes questions into three levels of complexity: depth 1, depth 2, and depth 3. As seen in Figure 7, depth 1 questions address a single geometric shape, while higher-depth questions involve multiple shapes, requiring a more complex reasoning process with multiple formula application steps. Furthermore, all questions in GeomVerse are designed to necessitate models to look at the image to find the answer, as some critical information like relevant angles and side lengths are only mentioned in the image and not in the textual question. The dataset contains 10k questions each in the train and validation sets. For our experiments on GeomVerse, we use the train set of GeomVerse

to create a code-finetuning dataset by generating 5 Python codes per geometry problem (using the process detailed in Section 3.1), with 35k unique image-question and code pairs encompassing 8.5k unique geometry problems.

Geo170k is an instruction-tuning dataset divided into two components: alignment and QA-tuning, with no test set. The alignment portion of the dataset contains 60k geometry image-caption pairs, while QA-tuning includes 110k geometry question-answering problems derived by augmenting Geometry3k (Lu et al., 2021) and GeoQA+ (Cao and Xiao, 2022). In Geo170k, each question is accompanied by four answer options for the model to select from, whereas in GeomVerse, the model must directly predict the final answer. Even though the final answer prediction task is much harder, we argue it is more realistic than selecting from options, as real-world geometry problems rarely provide answer choices. Therefore, we create and utilize a 'no-option' version of QA-tuning in this study, where the task is to predict the final answer. We call this version Geo170k-NO, which contains 110k CoT-annotated questions, to fine-tune our models. As there are no TikZ illustrations available in Geo170k, we use our teacher LLM to generate single code solutions for problems in Geo170k-NO given only the CoT, question and function library using the method mentioned in Section 3.1. This creates a code-finetuning dataset with 48k unique image-question and code pairs.

4.1.2 Evaluation Datasets

GeomVerse: We evaluate our models on the test set of GeomVerse which contains 1k questions from each depth, resulting in 3k test questions overall.

GeoQA includes geometry questions derived from Chinese middle school exams, with each problem annotated with the associated CoT (as seen in Table 8). The test set in GeoQA contains 754 problems. Since the original dataset is in Chinese, we follow Gao et al. (2023); Liang et al. (2023) and utilize the English version from Chen et al. (2022a) to ensure consistency in language across datasets. As with Geo170k, each question in GeoQA is provided with four answer options for the model to choose from. In this work, we evaluate using GeoQA-NO, a version of the dataset where models predict the final answer without multiple-choice options.

MathVerse is a visual math benchmark crafted to evaluate multi-modal mathematical reasoning skills of VLMs and spans three areas: plane geometry,

solid geometry, and functions. We evaluate on the plane geometry subset of the testmini set of MathVerse. As with GeoQA-NO, we convert MathVerse to a "no-options" setting and evaluate on questions where the model has to give the final floating point answer without seeing any options.

4.2 Metrics

Relaxed Accuracy: Following the methodology of Kazemi et al. (2023), Masry et al. (2022) and Methani et al. (2020), we assess model performance based on relaxed accuracy, where a prediction is deemed correct if it is within three percent of the ground truth label. This adjustment accounts for slight floating point computational differences arising from rounding during each calculation step.

4.3 Our Models

Teacher LLM. We use Llama 3.1 70B Instruct (Dubey et al., 2024) as our teacher LLM in this study because it is open-source and has strong code generation performance.

Student VLM. We use LLaVA 1.5 7B as the student VLM for code-finetuning GeoCoder and RAG-GeoCoder in this study, as discussed in Section 3. We generate 4 codes using beam search and select the first one that executes as the model's answer in our experiments.

Multimodal Retrieval. During training, we provide the VLM with k (here k= 7) retrieved functions, in addition to the "gold" functions extracted from the corresponding "gold" code in the fine-tuning dataset. These "gold" functions are supplied during training to encourage the model to consistently select functions from the input list. During inference, RAG-GeoCoder is supplied with 10 retrieved functions, where recall is 76% (see Figure 6). These choices are intended to optimize the recall of relevant functions while keeping the context length of the VLM manageable.

4.4 Baselines

In this study, we specifically compare the advantages of code-finetuning over CoT-finetuning. On the GeomVerse dataset, we report results from experiments with three popular VLMs namely LLaVA 1.5 (Liu et al., 2024), PaLI (Chen et al., 2022b) and GPT4V (Achiam et al., 2023)². We consider five settings: zero-shot prompting (GPT4V),

²The results on GPT4V are from Kazemi et al. (2023) and were obtained on a subset of randomly selected 10 examples per depth, and the correctness was determined manually.

Model	Type	Depth 1	Depth 2	Depth 3
Human	Human	80%	65%	55%
PaLI 55B	CoT few-shot	0.1%	0.3%	0.2%
GPT4V	CoT few-shot	70%	0%	0%
PaLI 5B	Finetuned	22.8%	15.6%	14.5%
PaLI 5B	CoT-finetuned	69.5%	46.8%	25.8%
LLaVA 1.5 7B	CoT-finetuned	82%	64%	42%
GeoCoder 7B (ours)	Code-finetuned	95.0%	77.5%	60.1%
RAG-GeoCoder 7B (ours)	Code-finetuned	95.7%	80.1%	58.1%

Table 2: Relaxed accuracy (described in Section 4.2) on the GeomVerse dataset shows that our best-performing code-finetuned model, RAG-GeoCoder, surpasses the corresponding CoT-finetuned LLaVA 1.5 by a margin of 13.7% on depth 1 and 16.1% on depth 2, while GeoCoder outperforms CoT-finetuned LLaVA 1.5 by 18.1% on depth 3, as discussed in Section 5.1.

	Type	Overall	Vision Intensive	Vision Dominant	Text Lite	Text Dominant	Vision Only
G-LLaVA 7B	CoT Finetuned	14.95	17.14	11.83	16.78	21.07	7.88
GeoCoder 7B (ours)	Code Finetuned	25.18	27.86	20.07	31.07	35.71	11.11

Table 3: Code-finetuned GeoCoder outperforms the corresponding CoT-finetuned G-LLaVA by 10.23% overall on the plane geometry subset of the MathVerse testmini dataset, as discussed in Section 5.1.

Model	Type	Relaxed Accuracy
G-LLaVA 7B	CoT-finetuned	28.0%
GeoCoder 7B	Code-finetuned	42.3%

Table 4: Code-finetuned GeoCoder outperforms the corresponding CoT-finetuned G-LLaVA by 14.3% on the GeoQA-NO dataset, as discussed in Section 5.1.

few-shot (4-shot) prompting with CoT (PaLI 55B), finetuning to predict the label directly (PaLI 5B), finetuning to predict CoT (PaLI 5B, which is the state-of-the-art on GeomVerse (Kazemi et al., 2023) and LLaVA 1.5 7B) and our proposed code-finetuning (GeoCoder 7B, RAG-GeoCoder 7B), described in Section 3. To show the difficulty of the GeomVerse dataset, we also report human accuracy on a small subset of the test split of the dataset from Kazemi et al. (2023). In our experiments on GeoQA-NO and MathVerse, we compare code-finetuning with CoT-finetuning using the G-LLaVA 7B VLM (Gao et al., 2023). The original G-LLaVA model is LLaVA 1.5 CoT finetuned on their proposed Geo170k dataset, where every question has 4 answer options. To work with our ‘no-options’ setting, we CoT finetune LLaVA 1.5 on the Geo170k-NO dataset (discussed in Section 4.1.1) and call this finetuned model G-LLaVA.

5 Results and discussion

5.1 Quantitative Results

Results on GeomVerse dataset: As seen in Table 2, code-finetuning demonstrates significantly superior performance to CoT finetuned baselines on GeomVerse across depths, with or without retrieval-based augmentation. RAG-GeoCoder, outperforms CoT-finetuned PaLI 5B (the state-of-the-art on this dataset) by 26.2% on depth 1 and 33.3% on depth 2, while GeoCoder does so by 36.3% on depth 3. Code-finetuned RAG-GeoCoder exceeds the performance of the corresponding CoT-finetuned LLaVA 1.5 by 13.7% on depth 1 and 16.1% on depth 2, while GeoCoder outperforms it by 18.1% on depth 3. The most significant improvement occurs at depths 2 and 3, which shows that modular code-finetuning enables language models to handle problems with long chains of mathematical reasoning and formula usage steps better than other finetuning approaches.

Results on GeoQA-NO dataset: Unlike the GeoQA dataset which has four answer choices for every geometry problem, in GeoQA-NO there are no options available and the task is to directly predict the final answer (as discussed in Section 4.1). As seen in Figure 4, G-LLaVA (Gao et al., 2023) (LLaVA 1.5 7B CoT finetuned on Geo170k-

	Correct Output	Wrong Output	Error Count	Math-domain Errors	Name Errors	Syntax Errors	Zero Div. Errors	Other Errors
GeoCoder								
Depth 1	950	38	12	0	2	9	1	0
Depth 2	775	211	14	0	0	13	0	1
Depth 3	601	365	34	2	9	21	0	2
RAG-GeoCoder								
Depth 1	957	43	0	0	0	0	0	0
Depth 2	801	178	21	0	17	2	0	2
Depth 3	581	369	50	3	34	2	1	10

Table 5: Error analysis of GeoCoder and RAG-GeoCoder on the test set of GeomVerse with 1000 examples per depth, as discussed in Section 5.2.

Model	D1	D2	D3
GeoCoder	71.7%	72.6%	70.6%
RAG-GeoCoder	90.3%	85.2%	90.0%

Table 6: Function usage analysis. RAG-GeoCoder uses functions for 17% more questions on average across depths as compared to GeoCoder on the test set of GeomVerse, as discussed in Section 5.4.

NO) gets 28% relaxed accuracy on the test set of GeoQA-NO. In contrast, GeoCoder gets 42.3% after code-finetuning on the gold code-tuning data made (discussed in Section 3.1) using Geo170k-NO. Although this performance far exceeds (by 14.3%) that of the CoT-finetuned model, it remains relatively low. This indicates that GeoQA-NO is more challenging than GeoQA, as models tend to underperform when the task shifts from predicting options to predicting final answers. In contrast to GeomVerse, where the figures consist of combinations of relatively simple shapes and the CoTs are detailed (see Table 7), GeoQA-NO and Geo170k-NO involve figures that require more sophisticated image understanding. Additionally, the CoTs are brief and often lack clear explanations of the solution (see Table 8). This also results in open code and low function usage in the solutions produced by our teacher LLM. Hence, we do not evaluate RAG-GeoCoder in our experiments with GeoQA-NO.

Results on MathVerse dataset: As seen in Table 3, we also evaluate GeoCoder on the plane geometry subset of the MathVerse testmini dataset under the no-options setting (as discussed in Section 4.1.2). GeoCoder outperforms the comparable CoT finetuned G-LLaVA model (trained on Geo170k-NO) across all question categories and by more than 10% overall.

5.2 Error Analysis

In this study, 'wrong outputs' refer to cases where the generated code runs but yields an incorrect answer, whereas 'error count' refers to instances where the code fails to execute and throws an exception. These exceptions are classified into five categories: math-domain errors (input values outside the function's domain), name errors (calling a function not present in the library), syntax errors (incorrect Python syntax), zero-division errors, and other uncategorized exceptions. As seen in Table 5, on the GeomVerse dataset, RAG-GeoCoder reduces the amount of syntax errors compared to GeoCoder, but the number of name errors increases. RAG-GeoCoder introduces more exceptions in depth 2 and 3 problems compared to GeoCoder. The primary cause of this increase is the rise in name errors for higher-depth problems. We attribute this to the relatively low recall rate of approximately 76% during multimodal retrieval at test time (see Figure 6). In depth 2 and 3 problems, where multiple functions are needed within a single solution, missing functions in the prompt lead the model to select incorrect ones, unlike GeoCoder which relies on its parametric memory.

5.3 Interpretability

In addition to geometric formula calculations, the functions in our proposed function library integrate templated print statements, making the computation process more interpretable. As seen in the example in Figure 5, the manually created templates in our functions are filled in automatically with the values from the arguments of the function and the computed results, thus introducing deterministic interpretability. This contrasts with the stochastic interpretability provided by CoT solutions where

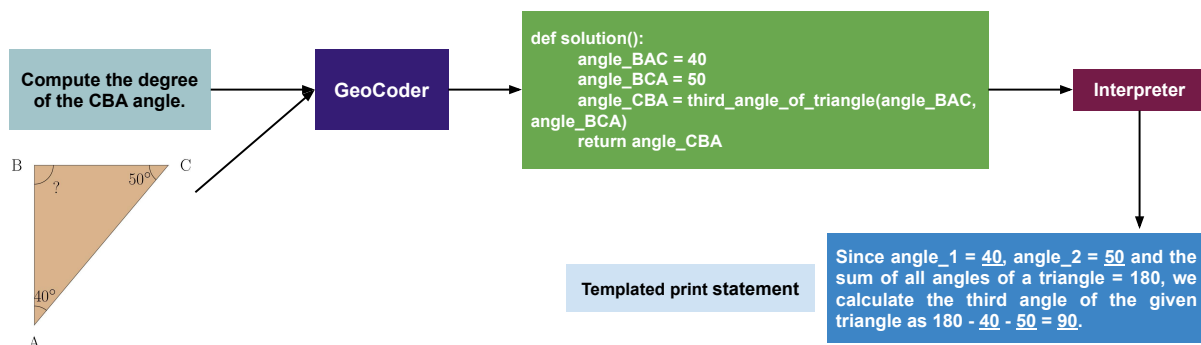


Figure 5: Modular functions add interpretability, as discussed in Section 5.3. In this example, the underlined values are filled in the template by the `'third_angle_of_triangle'` function.

the autoregressive language model can hallucinate the rationale behind the computation and still arrive at the correct result.

5.4 Function Usage Analysis

Functions add interpretability (see Section 5.3) and ensure correct geometry formula usage. To understand how well our models use functions, we compare the number of times they use functions to solve a given geometry problem. Specifically, we find that out of 1000 questions per depth in the GeomVerse test set, RAG-GeoCoder uses functions to answer 903 problems in depth 1, 852 in depth 2 and 900 in depth 3. In contrast, GeoCoder uses functions for only 717 problems in depth 1, 726 in depth 2 and 706 in depth 3. As seen in Table 6, RAG-GeoCoder demonstrates a 17% higher function usage on average across different depths, compared to GeoCoder on the GeomVerse dataset. This shows the effectiveness of RAG-GeoCoder at generating modular, interpretable code for solving geometry problems.

6 Conclusion

In this work, we present GeoCoder, a novel modular framework for code-finetuning aimed at solving geometry problems using VLMs. Our experiments demonstrate that code-finetuning outperforms CoT-finetuning for training autoregressive VLMs, as the deterministic calculations provided by code execution overcome the stochastic nature of VLM-based next-token prediction. To apply correct geometry formulas, mitigate errors from incorrect usage, and increase interpretability, GeoCoder uses a predefined function library that encodes widely used geometry formulas. In addition, we present RAG-GeoCoder, a retrieval-augmented variant that uses a non-parametric memory module to retrieve func-

tions from the geometry function library, avoiding dependence on the model’s parametric memory. The code-finetuned models we propose consistently outperform CoT-finetuned models across question complexity and different datasets.

7 Limitations

Although GeoCoder is designed to utilize functions from our predefined geometry library, it is not strictly required to do so while solving problems, as there is no enforced constraint on function usage. In the prompt provided to the teacher LLM, we encourage the use of functions from the library, but the model is not restricted to generating code solely from the listed functions. While the majority of the "gold" code produced by the teacher LLM includes library functions, some instances deviate from this, marking a limitation in our method. As a result, on the GeomVerse test set, RAG-GeoCoder does not use any functions for 9.7% problems in depth 1, 14.8% problems in depth 2 and 10% problems in depth 3 (see Table 6). Another limitation of our work is that the proposed geometric function library may not cover all the formulas required for solving every geometry question. However, our approach is designed to accommodate larger libraries and external function APIs to remedy this. We leave that exploration for future work. Moreover, since the vision modules in VLMs are typically pre-trained on real-world imagery, they struggle with interpreting geometric figures. The images in GeomVerse and GeoQA-NO differ significantly, requiring us to finetune our models on their respective training datasets in this study. This limitation may be addressed as future VLMs are pre-trained on a more diverse set of geometric figures. Furthermore, our approach for generating code-tuning datasets relies on and is limited by the need for existing geometry

datasets with questions and images. To address this, approaches like GeoGPT4V (Cai et al., 2024) can be used in conjunction with our approach for generating synthetic geometry code-tuning datasets.

Acknowledgements

We thank NSERC and Samsung for supporting this work and CIFAR for their support under the Canada CIFAR AI Chair program. We would also like to thank Aishwarya Agrawal and the class of IFT 6765 for their valuable feedback during the early development of this work.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Shihao Cai, Keqin Bao, Hangyu Guo, Jizhi Zhang, Jun Song, and Bo Zheng. 2024. Geogpt4v: Towards geometric multi-modal large language models with geometric image generation. *arXiv preprint arXiv:2406.11503*.
- Jie Cao and Jing Xiao. 2022. [An augmented benchmark dataset for geometric question answering through dual parallel text encoding](#). In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1511–1520, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.
- Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022a. [UniGeo: Unifying geometry logical reasoning via reformulating mathematical expression](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Jiaqi Chen, Jianheng Tang, Jinghui Qin, Xiaodan Liang, Lingbo Liu, Eric Xing, and Liang Lin. 2021. [GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 513–523, Online. Association for Computational Linguistics.
- Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*.
- Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022b. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvassy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. [The faiss library](#).
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wan-jun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, et al. 2023. G-llava: Solving geometric problem with multi-modal large language model. *arXiv preprint arXiv:2312.11370*.
- Zhibin Gou, Zhihong Shao, Yeyun Gong, Yujie Yang, Minlie Huang, Nan Duan, Weizhu Chen, et al. 2023. Tora: A tool-integrated reasoning agent for mathematical problem solving. *arXiv preprint arXiv:2309.17452*.
- Tanmay Gupta and Aniruddha Kembhavi. 2023. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14953–14962.
- Xiaotian Han, Yiren Jian, Xuefeng Hu, Haogeng Liu, Yiqi Wang, Qihang Fan, Yuang Ai, Huaibo Huang, Ran He, Zhenheng Yang, et al. 2024. [Infimmwebmath-40b: Advancing multimodal pre-training for enhanced mathematical reasoning](#). *arXiv preprint arXiv:2409.12568*.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- Juyong Jiang, Fan Wang, Jiasi Shen, Sungju Kim, and Sunghun Kim. 2024. A survey on large language models for code generation. *arXiv preprint arXiv:2406.00515*.
- Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.
- Mehran Kazemi, Hamidreza Alvari, Ankit Anand, Jialin Wu, Xi Chen, and Radu Soricut. 2023. Geomverse: A systematic evaluation of large models for geometric reasoning. *arXiv preprint arXiv:2312.12241*.

- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.
- Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.
- Zhenwen Liang, Tianyu Yang, Jipeng Zhang, and Xiangliang Zhang. 2023. **UniMath: A foundational and multimodal mathematical reasoner**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7126–7133, Singapore. Association for Computational Linguistics.
- Minpeng Liao, Wei Luo, Chengxi Li, Jing Wu, and Kai Fan. 2024. Mario: Math reasoning with code interpreter output—a reproducible pipeline. *arXiv preprint arXiv:2401.08190*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.
- Pan Lu, Ran Gong, Shibiao Jiang, Liang Qiu, Siyuan Huang, Xiaodan Liang, and Song-Chun Zhu. 2021. **Inter-GPS: Interpretable geometry problem solving with formal language and symbolic reasoning**. 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)*, pages 6774–6786, Online. Association for Computational Linguistics.
- Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. **ChartQA: A benchmark for question answering about charts with visual and logical reasoning**. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279, Dublin, Ireland. Association for Computational Linguistics.
- Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.
- OpenAI. 2024. Introducing gpt-4o: our fastest and most affordable flagship model. <https://platform.openai.com/docs/guides/vision>. Accessed: 2024-05-26.
- Dídac Surís, Sachit Menon, and Carl Vondrick. 2023. Vipergpt: Visual inference via python execution for reasoning. *arXiv preprint arXiv:2303.08128*.
- Till Tantau. 2013. *The TikZ and PGF Packages*.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, and Anja Hauth et al. 2024. **Gemini: A family of highly capable multi-modal models**.
- Trieu H Trinh, Yuhuai Wu, Quoc V Le, He He, and Thang Luong. 2024. Solving olympiad geometry without human demonstrations. *Nature*, 625(7995):476–482.
- Ke Wang, Houxing Ren, Aojun Zhou, Zimu Lu, Sichun Luo, Weikang Shi, Renrui Zhang, Linqi Song, Mingjie Zhan, and Hongsheng Li. 2023. **Mathcoder: Seamless code integration in llms for enhanced mathematical reasoning**. *arXiv preprint arXiv:2310.03731*.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. 2024a. **Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems?** In *European Conference on Computer Vision*, pages 169–186. Springer.
- Renrui Zhang, Xinyu Wei, Dongzhi Jiang, Yichi Zhang, Ziyu Guo, Chengzhuo Tong, Jiaming Liu, Aojun Zhou, Bin Wei, Shanghang Zhang, et al. 2024b. **Mavis: Mathematical visual instruction tuning**. *arXiv preprint arXiv:2407.08739*.
- Wenqi Zhang, Zhenglin Cheng, Yuanyu He, Mengna Wang, Yongliang Shen, Zeqi Tan, Guiyang Hou, Mingqian He, Yanna Ma, Weiming Lu, et al. 2024c. **Multimodal self-instruct: Synthetic abstract image and visual reasoning instruction using language model**. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19228–19252.
- Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. **VISTA: Visualized text embedding for universal multi-modal retrieval**. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Geometry Functions Library

Here is the list of functions in our proposed function library along with their descriptions:

1. `radius_of_sector_given_arc_length(arc_length, angle)` # Calculates the radius of a sector given the angle and the arc length.

2. `side_of_parallelogram_given_area(area, side, angle)` # Calculates the second side of a parallelogram given the area, the first side and the sine of the angle (in degrees) between them.
3. `height_of_triangle_given_area(area, base)` # Calculates the height of a triangle given the area and the base.
4. `base_of_triangle_given_area(area, height)` # Calculates the base of a triangle given the area and the height.
5. `arc_length_of_sector(angle, radius)` # Calculates arc length of a sector given the angle and the radius.
6. `length_of_third_side(side_1, side_2, angle)` # Calculates the length of the third side of a triangle given two sides and the angle between them in degrees, according to the law of cosines.
7. `height_of_equilateral_triangle(base)` # Calculates the height of an equilateral triangle given the base, according to the sine rule.
8. `diagonal_of_square(side)` # Calculates diagonal of a square given side.
9. `hypotenuse_of_right_triangle(side_1, side_2)` # Calculates hypotenuse of a right triangle given two sides, according to Pythagorean theorem.
10. `side_of_right_triangle_given_side(side, hypotenuse)` # Calculates the second side of a right triangle given the first side and hypotenuse, according to Pythagorean theorem.
11. `diameter_of_semi_circle_given_perimeter(perimeter)` # Calculates the diameter of a semi-circle given the perimeter/circumference.
12. `side_of_right_triangle_given_angle(opposite_angle, hypotenuse)` # Calculates the length of the side of a right triangle given opposite angle (in degrees) and hypotenuse.
13. `side_of_equilateral_triangle(height)` # Calculates the side of an equilateral triangle given the height, according to the sine rule.
14. `side_of_square_given_area(area)` # Calculates the side of a square given the area.
15. `arc_of_semi_circle(diameter)` # Calculates the arc length or arc circumference of a semi-circle given the diameter.
16. `diagonal_of_rectangle(side_1, side_2)` # Calculates the diagonal of a rectangle given two sides, according to Pythagorean theorem.
17. `side_of_rectangle_given_diagonal(side, diagonal)` # Calculates the second side of a rectangle given the first side and diagonal, according to the Pythagorean theorem.
18. `diameter_of_semi_circle_given_area(area)` # Calculates the diameter of a semi-circle given the area.
19. `area_of_rectangle(side_1, side_2)` # Calculates area of a rectangle given two sides.
20. `area_of_square_given_side(side)` # Calculates the area of a square given a side.
21. `area_of_square_given_diagonal(diagonal)` # Calculates the area of a square given the diagonal.
22. `area_of_triangle_given_base_height(base, height)` # Calculates the area of a triangle given the base and height.
23. `area_of_right_triangle(side_1, side_2)` # Calculates the area of a right triangle given two perpendicular sides.
24. `area_of_semi_circle(diameter)` # Calculates the area of a semi-circle given the diameter.
25. `area_of_circle(diameter)` # Calculates the area of a circle given the diameter.
26. `area_of_parallelogram(side_1, side_2, angle)` # Calculates the area of a parallelogram given two sides and the sine of the angle (in degrees) between them.
27. `area_of_triangle_given_semi_perimeter(semi_perimeter, side_1, side_2, side_3)` # Calculates the area of a triangle given three sides and the semi-perimeter, using Heron's formula.
28. `area_of_trapezoid(base_1, base_2, height)` # Calculates the area of a trapezoid given the two bases and the height.

29. `area_of_equilateral_triangle(side)` # Calculates the area of an equilateral triangle given the side.
30. `area_of_sector(angle, radius)` # Calculates the area of a sector given the angle and radius.
31. `side_of_square_given_perimeter(perimeter)` # Calculates the length of a side of a square given the perimeter.
32. `perimeter_of_shape(sides)` # Calculates the perimeter of a shape given a list of sides.
33. `perimeter_of_square(side)` # Calculates the perimeter of a square given a side.
34. `perimeter_of_rectangle_with_side_removed(side_1, side_2)` # Calculates perimeter of a rectangle with one side (`side_2`) removed given two sides.
35. `perimeter_of_square_given_diagonal(diagonal)` # Calculates the perimeter of a square given the diagonal.
36. `perimeter_of_trapezoid(base_1, base_2, side_1, side_2)` # Calculates the perimeter of a trapezoid given both the bases and the lateral sides.
37. `perimeter_of_rectangle(width, height)` # Calculates the perimeter of a rectangle given both the width and the height.
38. `perimeter_of_parallelogram(side_1, side_2)` # Calculates perimeter of a parallelogram given two sides.
39. `perimeter_of_triangle(side_1, side_2, side_3)` # Calculates perimeter of a triangle given all three sides.
40. `perimeter_of_semi_circle(diameter)` # Calculates the perimeter/circumference of a semi-circle given the diameter.
41. `third_angle_of_triangle(angle_1, angle_2)` # Calculates the third of a triangle given the other two angles.
42. `tangent_of_angle(angle)` # Calculates the tangent of an angle. The tangent of an angle in a right triangle gives the ratio between the opposite and the adjacent sides.
43. `angle_of_right_triangle(opposite_side, hypotenuse)` # Calculates the angle (in degrees) of a right triangle given the opposite side and hypotenuse.
44. `complementary_angle(angle)` # Calculates the complementary angle for a given angle.
45. `angle_of_sector(arc_length, radius)` # Calculates the angle of a sector given the radius and the arc length.
46. `ngle_of_parallelogram_given_area(side_1, side_2, area)` # Calculates angle between two sides of a parallelogram given two sides and the area.
47. `solve_for_x(eq_1, eq_2)` # Solves a system of equations to get value of x.

A.2 Hyper Parameters Used

In this work, we used 4-bit QLoRA (Dettmers et al., 2024) to finetune all our models. The LoRA Rank and LoRA Alpha (Hu et al., 2021) values were set to 128 and 256, respectively, to control the rank of the low-rank adaptation layers, allowing efficient fine-tuning. A Projection Learning Rate of $2e-5$ was chosen to ensure stable gradient updates during training. The model was trained for 10 epochs, with an initial learning rate of $2e-4$, following a cosine decay schedule that progressively reduced the learning rate for smoother convergence. A learning rate warmup ratio of 0.03 was employed to avoid sharp transitions at the start of training, preventing instability. Additionally, no weight decay was applied, as regularization through this method was unnecessary for the dataset and the size of the model used.

A.3 Additional Figures

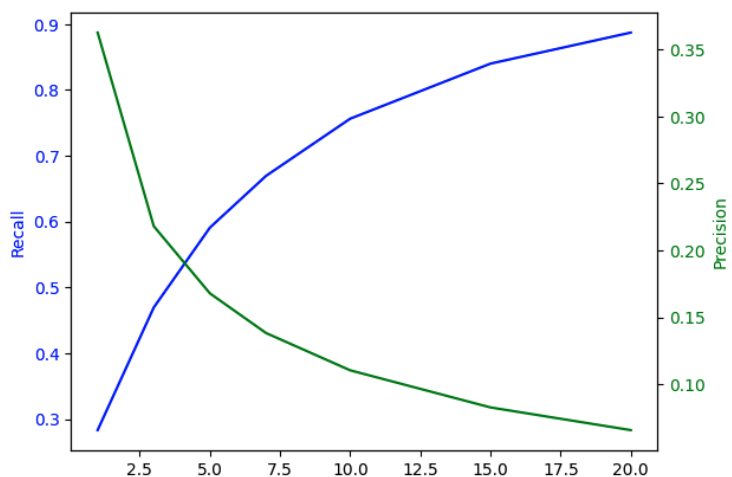


Figure 6: Precision-Recall for multimodal retrieval in RAG-GeoCoder on GeomVerse dataset.

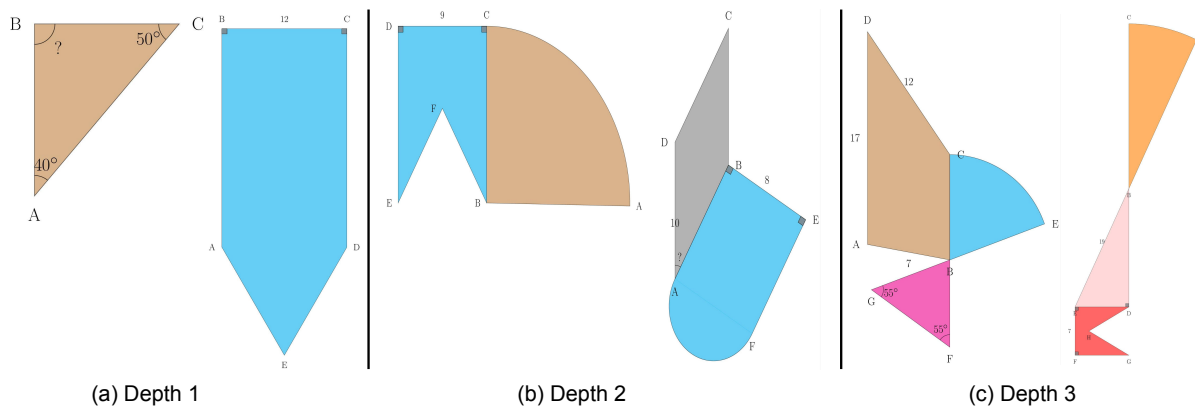


Figure 7: Complexity levels at different depths in the GeomVerse dataset.

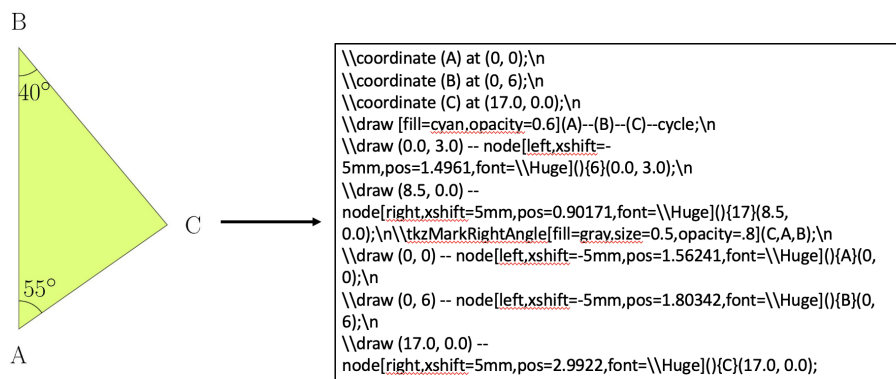


Figure 8: TikZ representation of an image from the GeomVerse dataset.

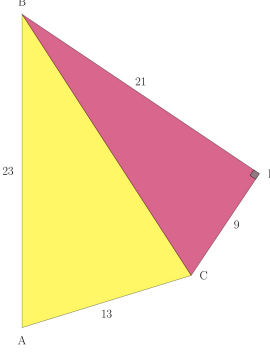
Image	Chain-of-Thought
	<p>Question: Compute the area of the ABC triangle. Round computations to 2 decimal places.</p> <p>Chain-of-Thought: The lengths of the BD and CD sides of the BCD triangle are 21 and 9, so the length of the hypotenuse (the BC side) is $\sqrt{21^2 + 9^2} = \sqrt{441 + 81} = \sqrt{522} = 22.85$. We know the lengths of the AB, AC and BC sides of the ABC triangle are 23 and 13 and 22.85, so the semi-perimeter equals $(23 + 13 + 22.85) / 2 = 29.43$. So the area is $\sqrt{29.43 * (29.43-23) * (29.43-13) * (29.43-22.85)} = \sqrt{29.43 * 6.43 * 16.43 * 6.58} = \sqrt{20458.07} = 143.03$. Therefore the final answer is 143.03.</p>
	<p>Question: If the ABCDE shape is a rectangle where an equilateral triangle has been removed from one side of it, the BAFG shape is a rectangle where a semi-circle has been removed from one side of it, the perimeter of the BAFG shape is 72, the AFIJ shape is a combination of a rectangle and a semi-circle and the perimeter of the AFIJ shape is 50, compute the perimeter of the ABCDE shape. Assume $\pi=3.14$. Round computations to 2 decimal places.</p> <p>Chain-of-Thought: The perimeter of the AFIJ shape is 50 and the length of the FI side is 8, so $2 * OtherSide + 8 + \frac{8}{2} * 3.142 = 50$. So $2 * OtherSide = 50 - 8 - \frac{8}{2} * 3.142 = 50 - 8 - 12.56 = 29.44$. Therefore, the length of the AF side is $\frac{29.44}{2} = 14.72$. The diameter of the semi-circle in the BAFG shape is equal to the side of the rectangle with length 14.72 so the shape has two sides with equal but unknown lengths, one side with length 14.72, and one semi-circle arc with diameter 14.72. So the perimeter is $2 * UnknownSide + 14.72 + \frac{14.72}{2} * \pi$. So $2 * UnknownSide + 14.72 + \frac{14.72}{2} * 3.142 = 72$. So $2 * UnknownSide = 72 - 14.72 - \frac{14.72}{2} * 3.142 = 72 - 14.72 - 23.11 = 34.17$. Therefore, the length of the AB side is $\frac{34.17}{2} = 17.09$. The side of the equilateral triangle in the ABCDE shape is equal to the side of the rectangle with width 14 so the shape has two rectangle sides with length 17.09, one rectangle side with length 14, and two triangle sides with lengths 14 so its perimeter becomes $2 * 17.09 + 3 * 14 = 34.18 + 42 = 76.18$. Therefore the final answer is 76.18</p>

Table 7: Sample Chain-of-Thought from the GeomVerse dataset

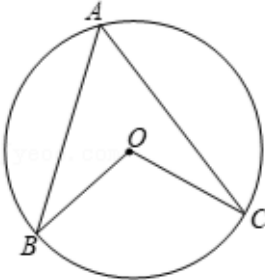
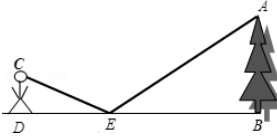
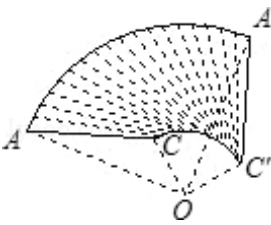
Image	Chain-of-Thought
	<p>Question: In the provided diagram, if points A, B, and C all lie on the circle O and angle BAC measures 54°, what is the measure of angle BOC?</p> <p>Chain-of-Thought: since angle A = 54°, therefore, angle BOC = 2 angle A = 108°. Therefore, the answer is 108°</p>
	<p>Question: What method did the mathematics interest group use to measure the height of a tree on the horizontal ground of the campus as depicted in the figure?</p> <p>Chain-of-Thought: Explanation: Given that angle AEB=angle CED, we can conclude that Right Triangle ABE is similar to Right Triangle CDE (using the property of angles). Therefore, the ratio of their corresponding sides must be equal: $AB/CD = BE/DE$. Substituting the given values, we have $AB/1.6 = 7.8/3.2$. Solving for AB, we get $AB = 3.9$ (meters).</p>
	<p>Question: Given the configuration, let the length of the windshield wiper AC be denoted as 'L'. If $AO = 65.0$, $CO = 15.0$, and the wiper AC rotates 90.0 around the point O, determine the expression for the area swept by the wiper AC in terms of 'L' and π.</p> <p>Chain-of-Thought: The area swept by the wiper AC can be calculated using the concept of congruent triangles and sectors. Substituting the given values, the area can be represented as $((65^2 - 15^2)/4) \pi$. By substituting 'L' for 65, the area becomes $((L^2 - 15^2)/4) \pi$, expressed in terms of 'L'. Thus, the area swept by the wiper AC is $1000 \pi \text{ cm}^2$.</p>

Table 8: Sample Chain-of-Thought from the Geo170k-QA-NO and GeoQA-NO datasets.


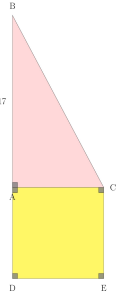
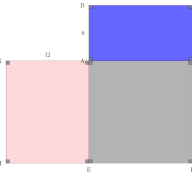
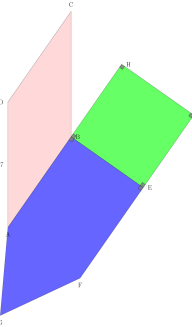
Image	RAG-GeoCoder Output
	<p>Question: If the area of the red semi-circle is 56.52, compute the diagonal of the ABCD rectangle.</p> <p>Ground Truth: 25.94</p> <p>RAG-GeoCoder:</p> <pre>def solution(): area_red_semicircle = 56.52 diameter = diameter_of_semi_circle_given_area(area_red_semicircle) side_AD = 23 diagonal = diagonal_of_rectangle(side_AD, diameter) return diagonal</pre> <p>Code Output: 25.94</p>
	<p>Question: If the area of the ADEC square is 81, compute the area of the ABC right triangle.</p> <p>Ground Truth: 76.5</p> <p>RAG-GeoCoder:</p> <pre>def solution(): area_ADEC = 81 side_ADEC = side_of_square_given_area(area_ADEC) side_AB = 17 area_ABC = area_of_triangle_given_base_height(side_ADEC, side_AB) output = area_ABC return output</pre> <p>Code Output: 76.5</p>
	<p>Question: If the diagonal of the AEFB rectangle is 21 and the diagonal of the AGHE rectangle is 19, compute the diagonal of the ABCD rectangle.</p> <p>Ground Truth: 16.97</p> <p>RAG-GeoCoder:</p> <pre>def solution(): diagonal_AEFB = 21 diagonal_AGHE = 19 side_AG = 12 side_AE = side_of_rectangle_given_diagonal(diagonal_AGHE, side_AG) side_AB = side_of_rectangle_given_diagonal(diagonal_AEFB, side_AE) side_AD = 8 output = diagonal_of_rectangle(side_AB, side_AD) return output</pre> <p>Code Output: 16.97</p>
	<p>Question: If the ABFG shape is a combination of a rectangle and an equilateral triangle, the perimeter of the ABFG shape is 66 and the perimeter of the BHIE square is 48, compute the perimeter of the ABCD parallelogram.</p> <p>Ground Truth: 64.0</p> <p>RAG-GeoCoder:</p> <pre>def solution(): perimeter_ABEFG = 66 perimeter_BHIE = 48 side_BE = side_of_square_given_perimeter(perimeter_BHIE) side_AB = (perimeter_ABEFG - 3 * side_BE) / 2 side_AD = 17 output = perimeter_of_parallelogram(side_AB, side_AD) return output</pre> <p>Code Output: 64.0</p>

Table 9: Sample code generations by the RAG-GeoCoder model on the GeomVerse test set.