MathFoundry

Grounded Generation of Secondary School Math Examination Practice Questions

Reina Peh Shu Ting (1005359)
Sean Chen Zhi En (1005122)
Tee Zhi Zhang (1005136)

Singapore University of Technology and Design
50.045 Information Retrieval
Prof. Soujanya Poria
December 2024

Abstract

MathFoundry is a Retrieval-Augmented Generation (RAG) system created for math question-and-answer generation. We introduce a novel database of questions from recent Singapore secondary school examination papers on Elementary Mathematics and Additional Mathematics subjects. Our RAG model consists of two main components: retrieval of existing questions from our database, and grounded generation of new questions based on, and cited with, existing questions. For the retrieval pipeline, we introduce a novel hybrid search approach to retrieve the most relevant documents. Our results show that this hybrid search approach significantly outperforms traditional vectorised search when evaluated against classical retrieval metrics like F1-score and Mean Average Precision. Also, we introduce a novel approach that uses a combination of Large Language Model (LLM) judges to verify whether a mathematical question is answerable. We further propose a new metric, MathTRUST-SCORE, to evaluate generated mathematical questions and answers. This metric aggregates answerability, trustworthiness, and the diversity of the generated question. Finally, we ran experiments to optimise our RAG pipeline for MathTRUST-SCORE.

Introduction

Background

The Singapore-Cambridge General Certificate of Education Ordinary Level (or Singapore-Cambridge GCE O-Level) is a national examination held annually by the Ministry of Education (MOE), Singapore Examinations and Assessment Board (SEAB) and the University of Cambridge for Secondary School students (typically aged 15 to 16). According to Black et al. (2023), there is a strong correlation between increased practice of questions and improved grades, with the average course grade increasing quadratically with the logarithm of the total number of questions completed (NQ, tot). In Singapore, teachers spend 46 hours per week working. This is significantly higher than the average of 39 hours per week across 48 global education systems (Wong, 2019). Teachers are found to spend a significant amount of time (7.5 hours per week) preparing materials for lessons (Wong, 2019), leading to increased workload stress.

Problem Statement

To improve students' grades, students often practice additional questions in topics that they are weaker in. In Singapore, teachers typically compile topic-specific questions for students to practice. However, it is difficult to compile topic-specific practice questions. Hence, our team aims to create a Retrieval-Augmented Generation (RAG) system called MathFoundry, which generates relevant topic-specific practice questions and answers.

Objectives

MathFoundry is designed to generate relevant and answerable mathematics questions with accurate answers. Currently, we are focusing on middle-school math-level subjects in Singapore: Elementary Mathematics and Additional Mathematics. We chose to focus on math because of the novelty of verifying the answerability of generated math questions and solutions. Referencing research done by Wei (2024), GPT-4o demonstrates strong performance in middle and high-school-level mathematics, particularly in algebra, but it is weaker in areas like geometry. The Elementary and Additional Mathematics subjects in Singapore overlap significantly with the middle and high school math syllabus in the United States. Hence, we decided to use GPT-4o for generation and focus on questions from the Elementary and Additional Mathematics subjects.

Dataset

Data Collection & Preprocessing

The source of our data is Holy Grail (n.d.), an online platform containing past year school exam papers aimed at students who are preparing for various national exams in Singapore. From the Holy Grail platform, we downloaded PDF files of 6 papers per target subject. The PDF for each paper was parsed using PyPDF Parser to process each file page by page and used GPT-4o to identify the text of individual questions, along with the topic and sub-topic. We parsed 331 questions across both subjects. The processed data was stored in MongoDB, a free-to-use NoSQL database that comes with vector search functionality through Vector Atlas Search. For each question parsed, we stored the text of the question in a field called `question_body`. Then, we passed the value stored in `question_body` through OpenAI's `text-embedding-3-small` model to create vector embeddings for each question. This setup enables us to perform vectorised searches easily. (see Appendix 1)

Database Schema

The database schema is designed to store metadata and content for mathematics questions used in the MathFoundry system. Each entry includes details such as the `question_body` (the text of the retrieved question) and its associated `topic` and `sub_topic`, enabling targeted retrieval based on user queries. To ensure transparency and traceability, the schema stores the web addresses for the question and answer papers' file paths under `question_paper_filepath` and `answer_paper_filepath`, along with the `page_start` and `page_end` fields to pinpoint the location of each question in the original documents. (see Appendix 2)

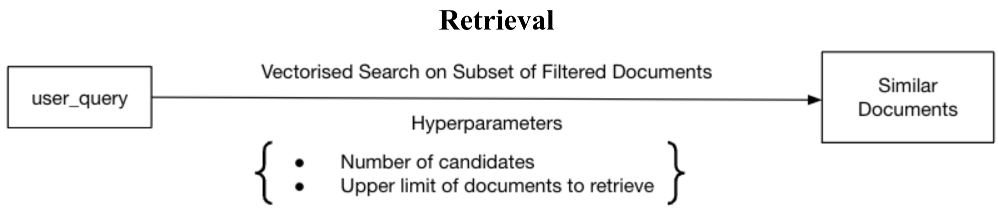


Figure 1: Diagram of Vector Search Process

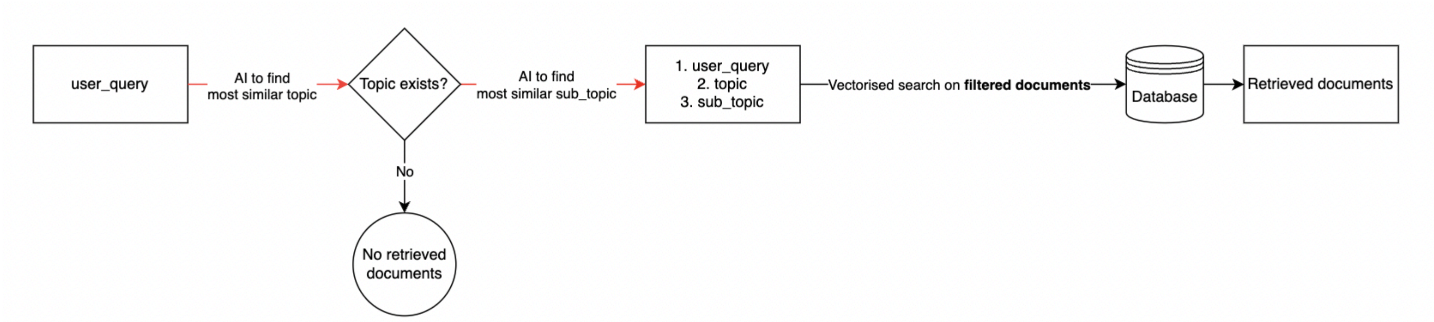


Figure 2: Diagram of Hybrid Search Process

Our retrieval process augments the traditional vectorised search by introducing a pre-filter based on the question's topic and sub-topic. Specifically, as shown in Figure 1, we first pass the user query into a LLM. Using OpenAI's structured outputs, this LLM compares the user query with a predefined list of topics to determine. If the user query is similar to any of them, the LLM extracts the most relevant topic. Then, we find the list of sub-topics nested under the extracted topic and pass the user query into another LLM to extract the most relevant sub-topic.

Evaluation Metrics Used

To evaluate the performance of our retrieval, we first constructed a ground truth dataset consisting of 35 user queries. Each user query is paired with a set of documents ranked in order of relevance. This dataset serves as the benchmark for assessing retrieval quality. Then, we conducted a grid search across two hyperparameters: the number of candidates considered during vector search and the upper limit of documents retrieved for each query. Finally, we assessed the performance of our retrieval using the following evaluation metrics: F1 score, Jaccard Similarity, Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR).

Experiments

Vector Search Hyperparameters

We first conducted experiments on a scaled-down evaluation dataset of just 10 queries across a grid search over a range of 5 values for both the number of candidates (from 20 to 60) and return limit (from 1 to 15) hyperparameters for vector search. We observe that the number of candidates is a trivial hyperparameter for vector search (see Appendix 3a). This was corroborated by experiments conducted with the full evaluation dataset of 35 queries across the same range for the number of candidates considered (see Appendix 3b). Since the number of candidates did not result in any change in our retrieval metrics, we excluded it from our experiments.

Hybrid Search

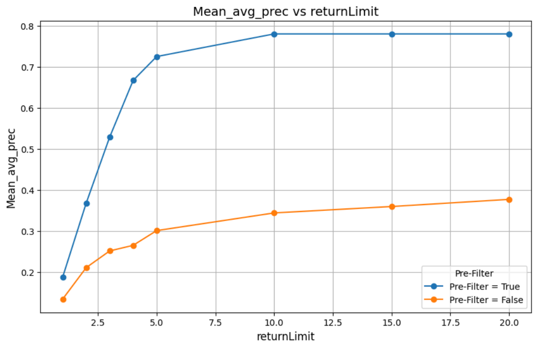


Figure 4a: Line Graph of Mean Avg. Precision (MAP) Against Return Limit

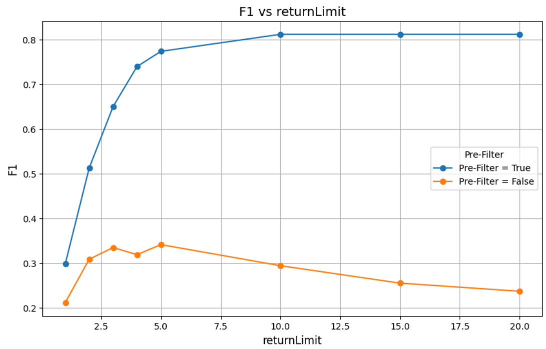


Figure 4b: Line Graph of Mean Reciprocal Rank (MRR) Against Return Limit

Next, we compared vector search against hybrid search while varying the upper limit of documents to retrieve. With reference to Figure 4a and Figure 4b, we can observe that retrieval metrics are significantly higher when the database of questions is first pre-filtered on the "topic" and "sub-topic" fields using the MongoDB Query Language. Also, the retrieval metrics are maximised when the upper limit of retrieved documents is set to 10.

Generation

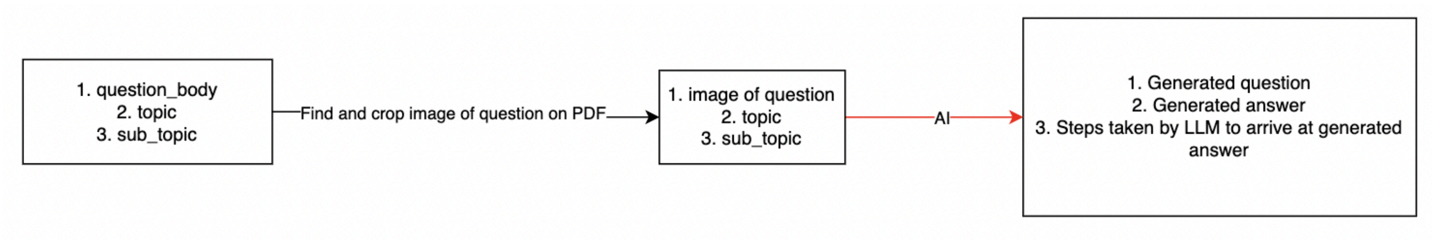


Figure 3: Diagram of Generation Process

In the generation pipeline, given a user query, we pass retrieved documents into GPT-4o to generate a contextually accurate and answerable question. Our model also outputs the steps taken to arrive at the answer. We define MathTRUST-SCORE, a metric which measures the quality of the model's output, and run a grid search to obtain the best configuration for generation. Each generated output is also cited with the base existing question(s) used as a reference.

Output Format with Citation

Example user query

"I need some exercises involving matrix multiplication; can you help?"

Example output from RAG pipeline

" Here is a question newly generated for your students to practice matrix multiplication:
`A science museum offers tickets at different prices: $45 for adults, $35 for students, and $25 for children. On a Saturday, the museum recorded 200 adults, 320 students, and 50 children visitors. On Sunday, there were 250 adults, 290 students, and 40 children visitors. Represent the number of visitors on Saturday and Sunday using the matrix \(V = \begin{pmatrix} 200 & 320 & 50 \\ 250 & 290 & 40 \end{pmatrix} \).`
`Determine the total revenue generated from ticket sales over the weekend.` **[1]**
Answer to this generated question: `43850`
[1] BUKIT PANJANG GOVERNMENT HIGH SCHOOL PRELIMINARY EXAMINATION 2023 MATHEMATICS PAPER 1 "

For each user query, our RAG pipeline generates a single generated question and answer. A citation is placed at the end of the generated question and at the end of the model's response, to allow the user to verify that the question is based on an actual practice question. The actual citation will be hyperlinked to the exact page in the PDF of the source exam paper, where the base question can be found.

Evaluation Metrics Used

Answerability of Generated Questions and Solutions

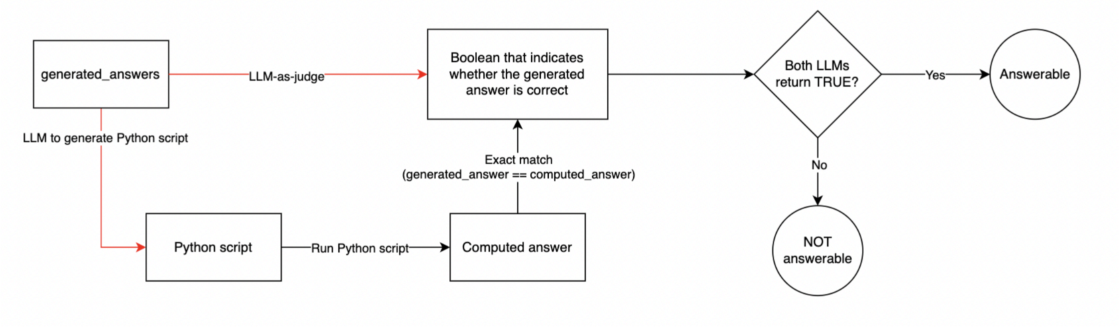


Figure 5: Diagram of Answerability Verification Process

Currently, there is no robust approach to verifying if a mathematical question is answerable. An answerable mathematical question is defined as a question that has a valid mathematical solution. We thus introduced a novel approach to classifying if a mathematical question is answerable by using a combination of LLM judges. Additionally, each of our judges made use of the GPT-4o model because they had to minimally be as good as the LLM that generated the question.

In choosing the LLM judges to use, we experimented with three different judges: a "Python script" judge, a "Step-by-step" judge, and a "black box" judge. When evaluating our LLM judges, we emphasised the reduction of false positive rates. Even a single false positive would be devastating because it could mean that a student would waste unnecessary time trying to solve a question that cannot be solved. Meanwhile, a false negative is less significant because it simply means that the end user might miss out on some solvable questions. At worst, this user would spend more computing resources generating the same number of solvable questions.

"Python Script" Judge is Not Sufficient

With this in mind, we first investigated if the "Python script" judge alone was sufficient. The "Python script" judge generates Python code to solve each question and compares the computed answer with the generated answer. Then, it classifies the question as solvable if there is an exact match between the computed answer and the generated answer. Intuitively, this judge would result in the least number of questions erroneously marked as answerable (false positives) as it is the most stringent. However, we still had to investigate if this judge was capable of producing even a single false positive. As such, we manually inspected all cases whereby the "Python script" judge was the only judge that identified a generated question as answerable. We found at least one instance where the "Python script" judge introduced a false positive (see Appendix 4). Hence, we needed to add at least one other LLM judge to reduce false positive rates.

"Step-By-Step" Judge Helps to Reduce False Positive Rates

To do this, we manually inspected all cases whereby the "step-by-step" judge was the only judge to classify a question as incorrect. We found at least one instance where the "step-by-step" judge correctly classified the question as incorrect (see Appendix 5). This shows that false positive rates will decrease if we define a question to be answerable only when both the "step-by-step" and "Python script" judges classify a question to be answerable.

"Black Box" Judge Does Not Reduce False Positive Rates

Finally, we manually inspected all cases whereby the "black box" judge was the only judge to classify a question as incorrect. In all cases, we found that the "black box" judge was hallucinating. Specifically, in all cases, we found that the generated question was indeed answerable. However, the "black box" judge consistently erroneously classified the questions as unanswerable. This means that adding this judge would only increase false negative rates without decreasing false positive rates. Thus, this judge was not used.

Based on these findings, **we define a question to be answerable if, and only if, both the "Python script" and "step-by-step" judges classify a question to be answerable**. We take the percentage of generated questions and solutions that pass both judges as our answerability metric.

Refusal Ability

For our model to produce questions and answers that are grounded by real-world questions in exam papers, it should only answer user queries that can be related to existing questions in our database. Hence, we need to measure our model's ability to refuse to answer queries that are irrelevant and not supported by our database's questions.

To do so, we relied on the Grounded Refusal metric component of the TRUST-SCORE, which takes the F1 of the quality of answering and refusal (Song et al., 2024). To measure the quality of refusal, we added 25 random and invalid user queries unrelated to math question and answer generation, to our evaluation dataset. Each of these queries has randomly generated retrieved documents.

Citation Groundedness

In instances where our model proceeds with generation, the new questions generated need to be based on real-world questions found in exam papers, which are retrieved as documents and used as context. To measure the level of groundedness of our model's output to its retrieved documents, we relied on the Citation Groundedness metric component of the TRUST-SCORE (Song et al., 2024). In our use case, the output is the generated math question, and the citations in the output are the retrieved documents that were passed into the generation prompt. To save cost and complexity, we limit the output to just one generated question per user query, and the generated question is treated as one entire statement.

Due to the lack of GPU access, we did not use natural-language inference (NLI) models for entailment classification but instead used GPT-4o as our entailment judge (see Appendix 6). As a result of these differences, we followed the definitions of Citation Recall and Precision in the Grounded Citations metric (see Appendix 7) but created new functions to calculate them.

Due to the novelty of using an LLM such as GPT-4o for entailment classification, we measured the accuracy of this method. On a manually created dataset of 40 groups of generated questions and their citations or retrieved documents, we labelled each citation as 1 (entailment) or 0 (neutral). **Comparing GPT-4o's labels against our ground-truth labels, the accuracy of GPT-4o's labels was 61.90%**. This is decent for a general-purpose model but leaves room for improvement with further fine-tuning.

Creativity

While generated questions should be grounded in real-world questions, they should not be mere regurgitations of existing questions in our database with slight changes in the values or names used. In other words, we also need a measure for the "creativity" of the model and its ability to generate output that differs significantly from the examples it was provided. To do so, we researched String Metric algorithms, which provide general and robust ways to measure the similarity of two strings. We first calculated the Damerau–Levenshtein distance (Damerau, 1964), which quantifies the minimal number of character-level changes needed to transform one string into another, between the generated question and each of the retrieved questions. Since math problems often differ by specific symbols, numbers, or terms, a higher distance indicates greater variation in structure, operations, or components, capturing their distinctiveness effectively. We then normalise this number by dividing the length of the longer string, to get a diversity metric ranging from 0 (exact match) to 100 (completely different).

MathTRUST-SCORE

To encapsulate the above metrics and build off the TRUST-SCORE (Song et al., 2024), we define MathTRUST-SCORE - an aggregation of a model's refusal ability, output groundedness in citation, answerability and creativity while generating math questions and solutions. The formula for MathTRUST-SCORE is defined as such, where each component is scaled from 0 to 100:

$$MathTrust = 0.5 \times Answerability + 0.2 \times CitationGroundednessF1 + 0.2 \times RefusalAbilityF1 + 0.1 \times Creativity$$

We allocated the weights for each component according to their importance in our use case.

Experiments

Plain text vs LaTeX

Since the "Python script" LLM judge required an exact match between the generated and computed answers, we hypothesised that the format of the generated answer would significantly impact the performance of the "Python script" LLM judge. We thus experimented with two different system prompts that instruct the LLM to generate a question and answer in plain text and LaTeX format respectively. Appendix 8 shows the system prompts we experimented with. At the same time, for each system prompt, we experimented with converting the generated answer from LaTeX format into plain text format using the Python latex2text library before passing the generated answer to our LLM judges.

This experiment was not conducted with the 25 out-of-topic questions. This is because the experiment was not used to test the refusal accuracy of our LLM. This experiment was repeated 3 times. In other words, each system prompt was evaluated against 105 questions.

Hyperparameters		Results		
System Prompt	Convert to Plain Text?	Number of Test Cases Passed	Total Number of Test Cases	Percentage of Test Cases Passed
Plain text	F	49	105	47.00%
Plain text	T	50		48.00%
LaTeX	F	49		47.00%
LaTeX	T	46		44.00%

Figure 6: Percentage of answerable questions for different output formats

As shown in Figure 6, the **format of the generated questions and answers does not matter**. Specifically, there is no significant difference in the percentage of test cases passed between all hyperparameter configurations.

Experimental Setup

Evaluation datasets

We created two ground truth datasets. The Elementary Mathematics dataset consists of 35 Elementary Mathematics questions and 25 out-of-topic questions. The Additional Mathematics dataset consists of 25 Additional Mathematics questions and 25 out-of-topic questions. There is a lower number of Additional Mathematics questions due to a limited number of Additional Mathematics sub-topics with more than 3 questions in our database.

Experiments

Grid Search

To find the optimal hyperparameter configuration for generating new questions, we performed a grid search over the following variables:

1. Number of retrieved documents by retriever
2. Single-shot vs Zero-shot system prompt used in generation
3. Text vs Multimodal input during generation

Due to cost limitations, the grid search experiment was only conducted once. While we had empirically found the optimal number of retrieved documents by the retriever to be 10, we only varied the number of retrieved documents from 1 to 4 for our experiments. This was attributed to observations that user queries in the ground truth datasets only retrieved at most 4 relevant documents. Hence, we limited the range from just 1 to 4 instead.

The single-shot prompt was implemented by adding a gold standard example question and answer at the end of our system prompt based on the sub-topic identified for a given user query (see Appendix 9). We experimented with generating a new question using a text and multimodal approach (see Appendix 10). The text pipeline simply passes the retrieved question's text, topic, and sub-topic directly into our LLM to generate a new question. Meanwhile, the multimodal pipeline first retrieves the image of the question using the retrieved question's text and the PDF which the question was sourced from. Then, the image of the question, the question's topic and sub-topic is passed into our LLM to generate a new question.

Question difficulty

After finding the optimal hyperparameter configuration, we investigated the impact of question difficulty on our RAG pipeline. We used Elementary Mathematics as a proxy for easy questions and Additional Mathematics for difficult questions. A proxy was used as it was challenging to manually evaluate the difficulty of a question since there is no objective measure for question difficulty. For example, a student could label a statistics question as easy while another student could label the same question as hard. Meanwhile, according to the official MOE syllabus (MOE Singapore, n.d.), Additional Mathematics builds on top of the content taught in Elementary Mathematics. Hence, it is fair to assume that Additional Mathematics is generally more difficult than Elementary Mathematics.

Results and analysis

Grid Search (Answerability)

We observe **no significant difference between the multimodal and text generation pipelines** (see Appendix 10). This means the text generation pipeline should be used for subsequent experiments as it is less computationally expensive. Also, we can observe that **a single retrieved document consistently performed better**. This might suggest that the LLM tends to hallucinate more if provided with more than one retrieved document. Finally, while there is no distinct separation in results between a single-shot prompt and a zero-shot prompt, the **experiments with a single-shot prompt generally performed better** than those with a zero-shot prompt. Specifically, the bottom three results were experiments which used a zero-shot prompt.

Grid Search (MathTRUST-SCORE)

Hyperparameters			**Results**								
Use Image?	No. of Retrieved Docs	Use Single-Shot Prompt?	Quality of Answering	Quality of Refusal	Grounded Refusal F1	Citation Precision	Citation Recall	Citation Grounded-ness F1	Answerability of Question	Creativity	Math TRUST-SCORE
F	1	F	98.6	98	98.3	<u>94.4</u>	<u>86.1</u>	<u>90.1</u>	<u>62.9</u>	53	<u>74.43</u>
F	1	T	98.6	98	98.3	66.7	61.1	63.8	60	60	68.42
T	3	T	98.6	98	98.3	38	58.3	53.9	60	<u>69.4</u>	67.38

Figure 7: Top 3 MathTRUST-SCORE Results for E-Math Dataset (35 Valid and 25 Invalid Queries)

Referring to Figure 7 and Appendix 11 for the full experiment results, grounded refusal metrics were consistently high across all hyperparameter configurations. This was likely due to our hybrid search retrieval refusing to generate an output if no relevant documents were retrieved. On the other hand, citation groundedness metrics varied significantly across experiments. Generally, adding more retrieved documents or employing single-shot prompts reduced citation groundedness. This might suggest that the LLM was unable to integrate the information present in multiple retrieved documents.

Also, when images were utilised or multiple documents were retrieved, there was often a noticeable drop in citation-related scores, diversity, and the MathTrust Score. In terms of citations and fidelity to sources, this suggests that simpler input conditions with no images and minimal retrieval may lead to more grounded and trustworthy responses.

Another interesting pattern observed is the trade-off between answerability and creativity metrics. Configurations with a higher answerability score tend to yield a lower creativity score. This is sensible as when the model deviates from the base, answerable questions for more diverse generations, its output is less likely to be answerable.

Overall, the best hyperparameter configuration is no image, no single-shot prompt and only 1 retrieved document. This suggests that **a minimalist approach yields better results**. Adding more context through images, multiple retrieved documents and single-shot prompts does not uniformly improve performance. Instead, it can cloud the citation groundedness and trustworthiness metrics. Future tuning efforts will need to balance the richness of the retrieval process with the clarity and trustworthiness of the final outputs.

Elementary and Additional Mathematics

Hyperparameters				**Results**		
Subject	Use Image?	Number of Retrieved Docs	Use Single-Shot Prompt?	Number of Test Cases Passed	Total Number of Test Cases	% of Test Cases Passed
E-MATH	F	1	F	22	36	62.86%
A-MATH	F	1	F	1	25	4.00%

Figure 8: Answerability for E-Math and A-Math Datasets

Using A-Math questions as a proxy for difficult questions compared to questions from E-Math, our findings suggest that it is significantly harder to generate difficult questions that were answerable and valid as compared to easier questions. Out of the 25 Additional Mathematics questions in our experiment, only 1 answerable question was generated.

Limitations

High Costs

The lack of GPU access restricted our ability to run Natural Language Inference (NLI) locally, which made us reliant on OpenAI (GPT-4o) API calls which are costly (50 questions x 4 citations per question = 200 API calls per experiment). Also, for developer convenience, the OpenAI API was used throughout all experiments. Furthermore, we used the latest model, GPT-4o, to improve the quality of LLM responses. All of these factors significantly increased our costs, limiting the total number of experiments we could conduct.

Difficulty in Creating Our Dataset

Our experiments required a new dataset to be created. This required us to invest a significant amount of time and effort to create our database of questions and evaluation datasets.

Results Are Only Valid For The Current MOE Syllabus

We will need to keep track of MOE syllabus updates to ensure questions in our database are up to date. Then, we will need to run our experiments against this new set of questions to ensure that our results remain relevant and that generated questions remain useful for our users.

Stochasticity of LLMs

Despite our efforts to ensure that generated questions are answerable, LLMs currently do not provide a foolproof, guaranteed verification of mathematical problems and solutions. There is still a risk that our LLM judges incorrectly classify a question as answerable. Hence, human intervention is required to confirm if a question is indeed answerable.

Future Work

Verification of Our Experiment Results

Due to the high costs of relying on OpenAI APIs, with dedicated GPU access, we can look towards re-running our experiments to ensure robustness in our results, using open-sourced and smaller models, as well as incorporating local NLI models to evaluate citation groundedness.

Increase The Diversity of Generated Questions

Also, more experiments should be conducted to increase the diversity of generated questions. This is to encourage our model to produce fresh and diverse questions, rather than simply swapping values within questions.

Generate More Questions Per User Query

Finally, our pipeline should be updated to generate more questions per user query. It was limited to a single generated question for ease of testing. However, generating more questions per user query will likely improve the user experience as it reduces the number of questions each user needs to ask.

Automate Data Collection Process

Moving forward, an automated script can be used to import significantly more questions from the Holy Grail website.

Conclusion

To conclude, the development of MathFoundry marks a meaningful stride in leveraging Retrieval-Augmented Generation (RAG) systems to address educational challenges. By automating the generation of topic-specific mathematics questions, MathFoundry provides a scalable solution to alleviate the workload of educators while enhancing students' learning experiences. The proposed hybrid retrieval pipeline and novel evaluation metric, MathTrust-Score, offer robust mechanisms to ensure answerability, trustworthiness, and creativity in generated outputs. Despite limitations such as high computational costs and reliance on current MOE syllabi, this work lays a solid foundation for future advancements. Moving forward, refining the verification processes, enhancing question diversity, and expanding automated data collection will further improve the system's utility and reliability. MathFoundry demonstrates the transformative potential of integrating AI into education, setting the stage for continued innovation in this domain.

The code and results from our experiments and projects are stored in this GitHub repository: https://github.com/ZhiZhangT/eduRAG

References

Black, W. K., Matz, R. L., Mills, M., & Evrard, A. E. (2023). Practice Makes Better: Quantifying Grade Benefits of Study. *ArXiv:2301.02927 [Physics]*. https://arxiv.org/abs/2301.02927

Damerau, F. J. (1964). A technique for computer detection and correction of spelling errors. *Communications of the ACM*, *7*(3), 171–176. https://doi.org/10.1145/363958.363994

Holy Grail. (n.d.). Grail.moe. https://grail.moe/

MOE Singapore. (n.d.). *ADDITIONAL MATHEMATICS SYLLABUSES Secondary Three to Four Express Course Normal (Academic) Course Implementation starting with 2020 Secondary Three Cohort*. Retrieved December 18, 2024, from https://www.moe.gov.sg/-/media/files/secondary/syllabuses/maths/2020-express_na-add-maths_syllabuses.pdf

Song, M., Sim, S. H., Bhardwaj, R., Chieu, H. L., Majumder, N., & Poria, S. (2024). *Measuring and Enhancing Trustworthiness of LLMs in RAG through Grounded Attributions and Learning to Refuse*. ArXiv.org. https://arxiv.org/abs/2409.11242

Wei, X. (2024). Evaluating chatGPT-4 and chatGPT-4o: performance insights from NAEP mathematics problem solving. *Frontiers in Education*, *9*. https://doi.org/10.3389/feduc.2024.1452570

Wong, C. (2019, June 19). *Teachers in Singapore work 46 hours per week, more than global average: survey*. Yahoo News. https://sg.news.yahoo.com/teachers-in-singapore-younger-and-better-trained-than-global-peers-survey-091105351.html?guccounter=1

Appendix

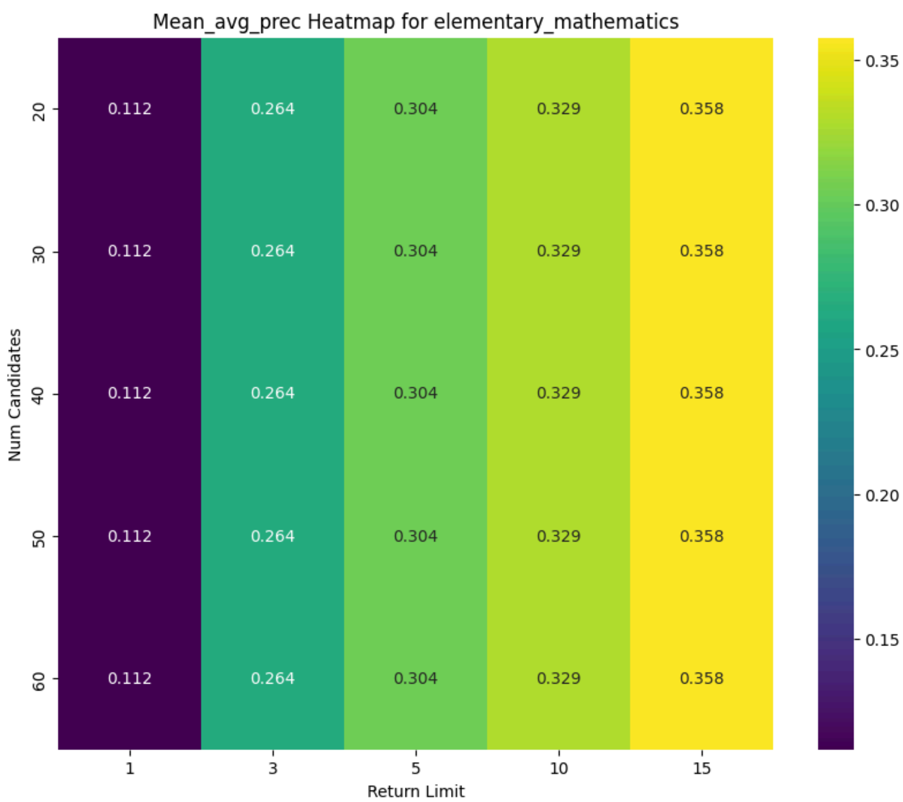
Appendix 1: Vector embeddings for semantic question retrieval


```
question_body : "A circle with centre $O$ passes through the points $P(-1, 7)$ and $Q(0…"
▾ question_body_embedding : Array (1536)
  0: 0.00812410656362772
  1: 0.02627989836037159
  2: 0.06066831201314926
  3: -0.02702837623655796
  4: -0.012744919396936893
  5: 0.04569874703884125
  6: -0.025302719324827194
  7: 0.0013371249660849571
  8: -0.024387912824749947
  9: 0.024304747581481934
  10: 0.04723728820681572
  11: 0.01905500516295433
  12: -0.017641214653849602
  13: 0.01260977704077959
  14: 0.005122397094964981
  15: -0.02567695826292038
  16: 0.022994911298155785
  17: 0.03887927904725075
  18: 0.00978219322860241
  19: 0.0811682939529419
  20: 0.03316174075007439
  21: 0.02632148005068302
  22: 0.02919064648449421
  23: 0.013077576644718647
  24: 0.017828334122896194
  25: 0.014553741551935673
  26: 0.004550643265247345
  27: 0.005137990694493055
  28: 0.013441420160233974
  29: 0.01735013909637928
```

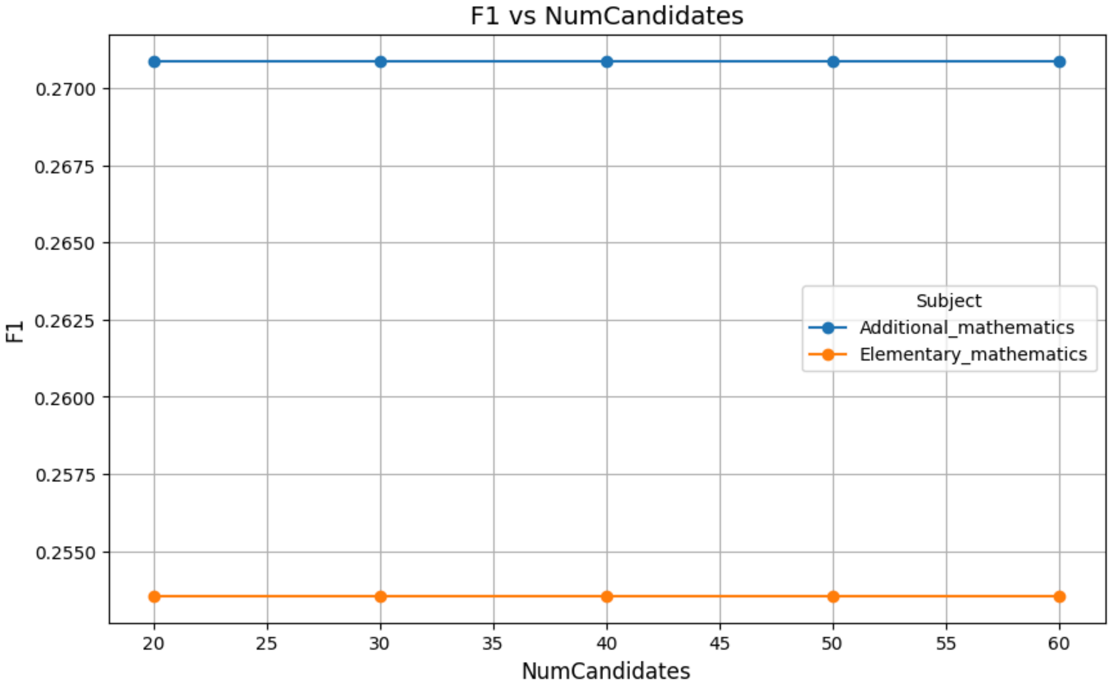
Appendix 2: Database schema storing PDF file paths, question metadata and details


```
_id: ObjectId('674dcd0fb00b977d048c92ac')
page_start : 4
page_end : 4
question_number : 2
question_part : "2b"
question_body : "A circle with centre $O$ passes through the points $P(-1, 7)$ and $Q(0…"
▸ question_body_embedding : Array (1536)
topic : "Coordinate geometry"
sub_topic : "Finding coordinates of unknown collinear point"
answer_body : ""
question_paper_filepath : "https://document.grail.moe/ee5096dfab614b04b525f87a54128941.pdf"
answer_paper_filepath : "https://document.grail.moe/06020d9323f04ea0bf13be6caab5e092.pdf"
subject : "additional_mathematics"
paper_number : "2"
level : "o_level"
exam_type : "final_exam"
year : 2023
school : "anglo__chinese_school_independent"
marks : 5
difficulty : "difficult"
created_utc : 2024-12-02T15:06:55.800+00:00
updated_utc : 2024-12-02T15:06:55.800+00:00
```


Appendix 3a:
Heatmap of Mean Average Precision Scores



Appendix 3b:
Line Graph of F1 against NumCandidates



Note for Appendix 3: The results on the triviality of NumCandidates as a parameter were the same across all evaluation metrics used: F1, Jaccard Similarity, Mean Avg. Precision and Mean Reciprocal Rank. For full results on this, refer to the ***retrieval_gridsearch.ipynb*** *notebook in our repository.*

Appendix 4: "Python script" judge incorrectly classified a question as answerable when all other judges correctly classified it to be answerable

Example unanswerable generated question
Matrix M represents the original prices of three items: toothpaste ($5), shampoo ($10), and soap ($3). Write this as a 1x3 matrix M. Matrix D represents the discount percentages applied at two different stores, where the first row shows a 15% discount and the second row shows a 5% discount. Write this as a 2x3 matrix D. Calculate the matrix product MD to determine the discounted price for each item at both stores.
Remarks
Only the "Python script" judge hallucinated by erroneously classifying the question as answerable.
However, this question is not answerable because it is not mathematically possible to do a matrix product between a 1-by-3 and a 2-by-3 matrix.

Appendix 5: "Step-by-step" judge correctly classified a question as unanswerable when all other judges incorrectly classified it to be answerable

Example unanswerable generated question
Charlotte has written five positive integers. The median of these numbers is 9, the mode is 10 and the mean is 12. The range of these numbers is 18. Find the five numbers.
Remarks
Only the "step-by-step" judge correctly classified the question as unanswerable. This question is unanswerable as there is no mathematical solution.

Appendix 6: GPT-4o Entailment Classification Prompt

```
prompt = f"""
          You are a helpful assistant that determines if a newly generated math question (the
hypothesis) is grounded in the provided set of existing math questions (the premise).

       Instructions:
```

```
      - Return '1' if the new generated question (hypothesis) is related and grounded to the
retrieved existing questions (premise).
      - Return '0' if the generated question is completely different from or unrelated to the
retrieved existing questions.

      Premise:
      {passage}

      Hypothesis:
      {claim}
""".strip()
```

Structured Output defined for response:

```
class EntailEnum(str, Enum):
   ENTAILS = "1"
   NOT_ENTAIL = "0"

class Entails(pydantic.BaseModel):
   entails: EntailEnum

entail_completion = client.beta.chat.completions.parse(
           model="gpt-4o",
           messages=[
           {"role": "system", "content": "You are a helpful assistant that only returns '1' or
'0'."},
           {"role": "user", "content": prompt}
           ],
           response_format=Entails,
       )
```

Appendix 7: Formulae for Citation Precision (CP), Citation Recall (CR), and F1 score (F1CG)

$$\mathrm{CR} = \frac{1}{|A_r|} \sum_{S \in A_r^s} \frac{1}{|S|} \sum_{s_i \in S} \mathrm{CR}^{s_i} \qquad \mathrm{CP} = \frac{1}{|A_r|} \sum_{C \in A_r^c} \frac{1}{|C|} \sum_{c_j \in C} \mathrm{CP}^{c_j}$$

$$\mathrm{F1}_{\mathrm{CG}} = \frac{2 \cdot \mathrm{CP} \cdot \mathrm{CR}}{\mathrm{CP} + \mathrm{CR}}$$

A_r	Set of questions where model provided an answer
S	Set of statements in a generated response
C	Set of citations in a generated response
A_r^s	Set of responses (only statements, no citations) in the dataset
A_r^c	Set of responses (only citations, no statements) in the dataset

Appendix 8: Plain Text and LaTeX System Prompts

Plain Text System Prompt

```
Given input containing:
- Multiple images, where each image displays a mathematical problem
- ID of each image in <image_id> tags
- Topic in <topic> tags
- Sub-topic in <sub_topic> tags

Instructions:
1. Generate one similar but distinct question that:
- Retains the same topic and sub-topic focus
- Is similar in difficulty and complexity level
- Uses plain text, not LaTeX
2. Introduce diversity by:
- Using varied contexts or scenarios while keeping the mathematical principles intact
- Exploring slightly different representations or formats for the same type of problem (e.g.,
equations vs. word problems)

Return a JSON response with:
{
"question_text": "Full question text",
"topic": "Mathematical topic",
"sub_topic": "Specific sub-topic",
"steps": [
"Step 1: Description and calculation",
"Step 2: Description and calculation",
],
"citations": "ID of images used to generate the new question",
"answer": "Final numerical or algebraic answer in plain text. The format should be 'Answer:
<answer>'"
```

```
}
```

LaTeX System Prompt

```
Given input containing:
- Multiple images, where each image displays a mathematical problem
- ID of each image in <image_id> tags
- Topic in <topic> tags
- Sub-topic in <sub_topic> tags

Instructions:
1. Generate one similar but distinct question that:
- Retains the same topic and sub-topic focus
- Is similar in difficulty and complexity level
- Uses LaTeX formatting for mathematical expressions
2. Introduce diversity by:
- Using varied contexts or scenarios while keeping the mathematical principles intact
- Exploring slightly different representations or formats for the same type of problem (e.g.,
equations vs. word problems)

Return a JSON response with:
{
"question_text": "Full question text",
"topic": "Mathematical topic",
"sub_topic": "Specific sub-topic",
"steps": [
"Step 1: Description and calculation",
"Step 2: Description and calculation",
],
"citations": "ID of images used to generate the new question",
"answer": "Final numerical or algebraic answer. The format should be 'Answer: <answer>'"
}
```

Appendix 9: Single-Shot and Zero-Shot System Prompts

Zero-Shot System Prompt

```
Given input containing:
- ID of a mathematical question in <question_id> tags
- Question in <question> tags
- Topic in <topic> tags
- Sub-topic in <sub_topic> tags

Instructions:
1. Generate one similar but distinct question that:
- Retains the same topic and sub-topic focus
- Is similar in difficulty and complexity level
- Uses LaTeX formatting for mathematical expressions
2. Introduce diversity by:
- Using varied contexts or scenarios while keeping the mathematical principles intact
```

```
- Exploring slightly different representations or formats for the same type of problem (e.g.,
equations vs. word problems)

Return a JSON response with:
{
"question_text": "Full question text",
"topic": "Mathematical topic",
"sub_topic": "Specific sub-topic",
"steps": [
"Step 1: Description and calculation",
"Step 2: Description and calculation",
],
"citations": "ID of the questions used to generate the new question",
"answer": "Final numerical or algebraic answer. The format should be 'Answer: <answer>'"
}
```

Single-Shot System Prompt

```
Given input containing:
- ID of a mathematical question in <question_id> tags
- Question in <question> tags
- Topic in <topic> tags
- Sub-topic in <sub_topic> tags

Instructions:
1. Generate one similar but distinct question that:
- Retains the same topic and sub-topic focus
- Is similar in difficulty and complexity level
- Uses LaTeX formatting for mathematical expressions
2. Introduce diversity by:
- Using varied contexts or scenarios while keeping the mathematical principles intact
- Exploring slightly different representations or formats for the same type of problem (e.g.,
equations vs. word problems)

Return a JSON response with:
{
"question_text": "Full question text",
"topic": "Mathematical topic",
"sub_topic": "Specific sub-topic",
"steps": [
"Step 1: Description and calculation",
"Step 2: Description and calculation",
],
"citations": "ID of the questions used to generate the new question",
"answer": "Final numerical or algebraic answer. The format should be 'Answer: <answer>'"
}

Example question:
```

```
Evaluate the matrix \\( T = AB \\), where \\( A \\) and \\( B \\) are given matrices:\n\\[ A = \\begin{bmatrix} 3 & 2 \\\\ 7 & 8 \\end{bmatrix}, \\quad B = \\begin{bmatrix} 3 & 2 \\\\ 7 & 8 \\end{bmatrix}. \\]

Example answer:
\\[ \\begin{bmatrix} 23 & 22 \\\\ 77 & 78 \\end{bmatrix} \\]
```

Gold Standard Examples For Each Sub-Topic

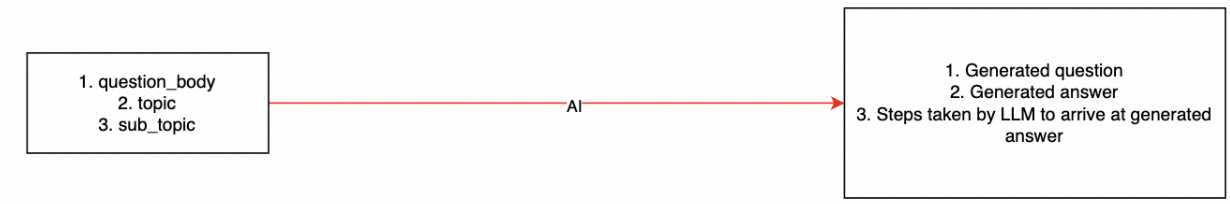
The example question and answer inserted depends on the sub-topic of the question. Using the question's sub-topic, we will find the relevant example question and answer from the following JSON:

```
{
 "Multiplication of 2 matrices": {
   "question": "Evaluate the matrix \\( T = AB \\), where \\( A \\) and \\( B \\) are given matrices:\n\\[ A = \\begin{bmatrix} 3 & 2 \\\\ 7 & 8 \\end{bmatrix}, \\quad B = \\begin{bmatrix} 3 & 2 \\\\ 7 & 8 \\end{bmatrix}. \\]",
   "answer": "\\[ \\begin{bmatrix} 23 & 22 \\\\ 77 & 78 \\end{bmatrix} \\]"
 },
 "Finding the mean, median and mode": {
   "question": "The table shows the shoe sizes of 60 people:\n\\[\n\\begin{array}{|c|c|c|c|c|c|}\n\\hline\n\\text{Shoe size} & 5 & 6 & 7 & 8 & 9 \\\\\n\\hline\n\\text{Frequency} & 24 & x & 13 & y & 5 \\\\\n\\hline\n\\end{array}\n\\]\nGiven that the mean shoe size is 6.3, determine the median and mode shoe sizes.",
   "answer": "The median shoe size is 6, and the mode shoe size is 5."
 },
 "Applying the angles properties of circles": {
   "question": "In the diagram, \\( A \\), \\( B \\), and \\( C \\) are three points on a circle with center \\( O \\). Given that \\( AOP \\) is a diameter of the circle and angle \\( \\angle COP = 68^\\circ \\), find \\( \\angle ABC \\).",
   "answer": "124 degrees"
 },
 "Applying rates in real-world contexts": {
   "question": "The rates of renting a car by two car rental companies are as shown:\n\\[\n\\begin{array}{|c|c|}\n\\hline\n\\text{Company A} & \\text{Company B} \\\\\n\\hline\n\\$11 \\text{ per hour for the first 4 hours} & \\$30 \\text{ for the first 2 hours} \\\\\n\\$8 \\text{ per half-hour thereafter} & \\$12 \\text{ per hour thereafter} \\\\\n\\hline\n\\end{array}\n\\]\nKenneth plans to rent a car for 6 hours. Which company should he rent the car from so that he pays less?",
   "answer": "Company A"
 },
 "Finding the probability of single events": {
   "question": "A bag contains three types of coffee sachets: 9 are caramel flavoured, 11 are hazelnut flavoured, and 12 are vanilla flavoured.\n(a) A sachet is chosen at random.\nWhat is the probability that it is not vanilla flavoured?",
   "answer": "5/8"
 },
 "Solving simultaneous equations using the method of elimination": {
   "question": "Solve the simultaneous equations:\n\\[ 2x + 3y = 15 \\]\n\\[ -3y + 4x = 3 \\]",
   "answer": "x = 3, y = 3"
```

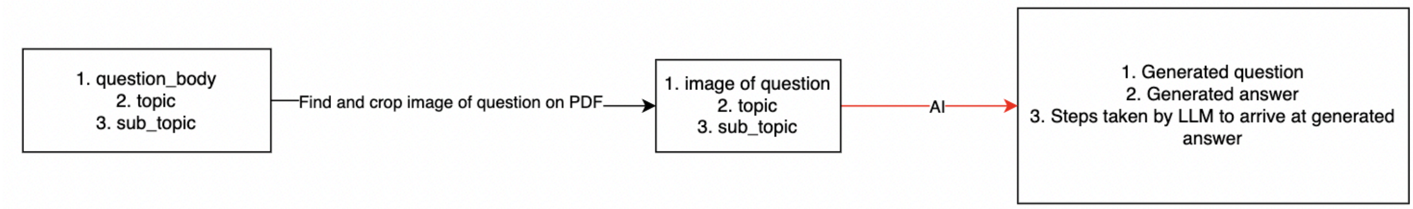
```
 },
 "Evaluating an algebraic formula": {
   "question": "The length, breadth, and height of a cuboid are \\( l \\) cm, \\( b \\) cm, and \\(
h \\) cm respectively. Find the volume \\( V \\) when \\( l = 8 \\), \\( b = 3 \\), and \\( h = 5
\\).",
   "answer": "120"
 }
}
```

Appendix 10: Text and Multimodal Generation Pipeline

Text Generation Pipeline



Multimodal Generation Pipeline



Appendix 11: Full MathTRUST-SCORE Results For E-Math Dataset (35 Valid and 25 Invalid Queries)

Hyperparameters			Results								
Use Image?	No. of Retrieved Docs	Use Single-Shot Prompt?	Quality of Answering	Quality of Refusal	Grounded Refusal F1	Citation Precision	Citation Recall	Citation Grounded-ness F1	Answerability of Question	Creativity	Math TRUST-SCORE
F	1	F	98.6	98.0	98.3	94.4	86.1	90.1	62.9	53.0	74.4
F	1	T	98.6	98.0	98.3	66.7	61.1	63.8	60.0	60.0	68.4
T	3	T	98.6	98.0	98.3	38.0	58.3	53.9	60.0	69.4	67.4
T	4	T	98.6	98.0	98.3	45.8	75.0	56.9	57.1	70.0	66.6
F	2	T	98.6	98.0	98.3	59.7	69.4	64.2	51.4	68.6	65.1
T	1	F	98.6	98.0	98.3	44.4	41.7	43.0	60.0	60.0	64.3
F	3	T	98.6	98.0	98.3	48.1	61.1	53.9	54.3	66.4	64.2
T	4	F	98.6	98.0	98.3	40.7	68.6	51.1	54.3	68.9	63.9
F	2	F	98.6	98.0	98.3	65.7	71.4	68.5	48.6	61.4	63.8
F	4	T	98.6	98.0	98.3	64.1	69.4	66.7	48.6	61.3	63.4
T	2	T	98.6	98.0	98.3	36.1	41.7	38.7	57.1	72.2	63.2
F	3	F	98.6	98.0	98.3	59.5	65.7	62.5	48.6	61.8	62.6
F	4	F	98.6	98.0	98.3	53.2	69.4	60.3	42.9	62.2	59.4
T	3	F	98.6	98.0	98.3	34.3	63.9	44.6	42.9	71.4	57.2
T	1	T	98.6	98.0	98.3	38.9	41.7	40.2	45.7	61.7	56.7
T	2	F	98.6	98.0	98.3	38.9	55.6	45.8	40.0	72.8	56.1

Appendix 12: Invalid query in evaluation dataset that elicited output from our model

```
        USER_QUERY: What is the Fibonacci sequence used for in real life?,

        GENERATED_QUESTION: "Emma has a total of 240 mangoes, 320 bananas, and 400 cherries. She
wants to arrange them into identical fruit trays. What is the maximum number of fruit trays that
Emma can prepare?",

        GENERATED_ANSWER: 80,

        CITATION_TEXT: "Mary needs to pack 210 oranges, 252 apples and 294 pears into identical
fruit baskets. (i) What is the largest possible number of fruit baskets that can be packed?",
```

This is an example of an invalid, random user query which incorrectly caused our model to generate an output instead of refusing to answer. Our model incorrectly identified the topic and sub-topic for this query as “Number Theory” and “Greatest Common Divisor (GCD)” respectively, which led to the generation of output.