



MathFoundry

Grounded Generation of Secondary School
Math Examination Practice Questions

Team **ZSR**:

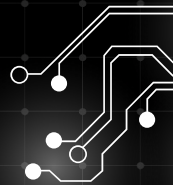
Tee **Z**hi Zhang (1005136), **S**ean Chen Zhi En (1005122), **R**eina Peh (1005359)

50.045 Information Retrieval

Content

1. Problem
2. Solution
3. Data collection
4. Ensure that retriever returns most relevant documents
5. Ensure that generated questions are answerable
6. Maximise number of answerable generated questions
7. [STRETCH] Generating questions of different difficulty
8. Ensure generated questions are citation grounded + LLM refuses to answer (TRUST)
9. Limitations
10. Future Work

.....► Experiments



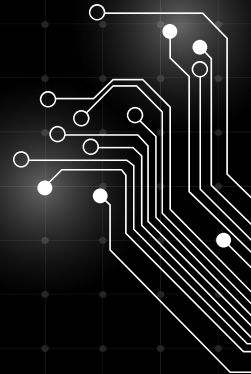
Problem

Time-consuming to create new topic-specific practices for students



Students often request for extra
practice questions for weaker topics

= more workload for educators



Problem

Topics & Question Types

Example: Mathematics

Topic 14:

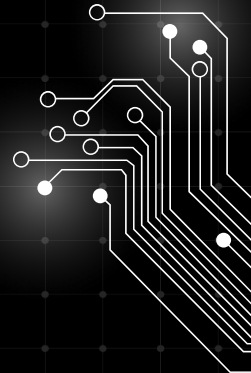
Applications of integration

Question Types:

- Definite integrals involving algebraic functions
- Definite integrals involving trigonometric functions
- Properties of definite integrals
- Integration as a reverse process of differentiation
- Integration as a reverse process of differentiation involving multiple functions
- Area under curve
- Area above and below x-axis
- Area bounded by curve, y-axis, and line
- Area bounded by curve and line
- Applications of tangent and normal to shaded area



How might we allow teachers to easily **create & compile topic-specific (and question type-specific) questions?**



Solution

Target User: Teachers

Inputs

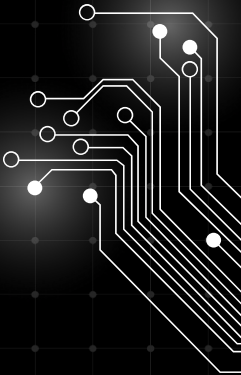
Subject & User Query

Optional:

- Question Topic
- Question Type
- School
- Year

Outputs

- Generated Question
- Generation Answer



Data Collection & Processing

Data Source

Find recent exam papers at :
<https://grail.moe/>

Download PDF files for parsing:

- 6 each for both Additional and Elementary Mathematics

Convert PDF to JSON for each paper:

- PyPDF Parser to process input page by page
- GPT-4o identifies individual questions along with the topic and sub-topic

Filter By:

Category

GCE '...'

Subject

Eleme...

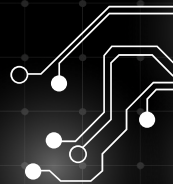
Type

Document N...

Year

Document Name	Category	Subject	Type	Uploaded By	Year	Uploaded On	Download
<div><div></div><div>2023 Springfield 3G3 Mathematics (4052) (P2-QP)</div></div>	GCE 'O' Levels	Elementary Mathematics	MYEs/CAs/Other Tests	rayden03	2023	3 December 2024	<div></div>
<div><div></div><div>2023 Springfield 3G3 Mathematics (4052) (P1-QP)</div></div>	GCE 'O' Levels	Elementary Mathematics	MYEs/CAs/Other Tests	rayden03	2023	3 December 2024	<div></div>

```
{
  "question_paper_filepath": "https://document.grail.moe/fad1557809cb4c8c8f9b7e450cef190c.pdf",
  "answer_paper_filepath": "https://document.grail.moe/9798972321764462873d7c189d16c14f.pdf",
  "meta_info": {
    "subject": "additional_mathematics",
    "school": "Cedar Girls\u20192019 Secondary School",
    "level": "o_level",
    "year": "2024",
    "exam_type": "preliminary_exam",
    "paper": "1"
  },
  "questions": [
    {
      "question": "Two cylinders are such that the ratio of their heights is 7 : 1. The height of 1",
      "question_number": 1,
      "question_part": "1",
      "page_start": 3,
      "page_end": 3,
      "category": "Surds",
      "marks": 4,
      "difficulty_level": "difficult",
      "question_type": "Word problems involving surds"
    },
  ]
}
```



Data Collection & Processing

Database

No-SQL database used: MongoDB

- Free-to-use
- Comes with vector database functionality: Vector Atlas Search

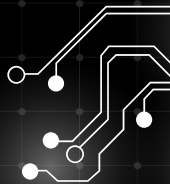
Vector embeddings created for each question using the question text

- OpenAI embeddings

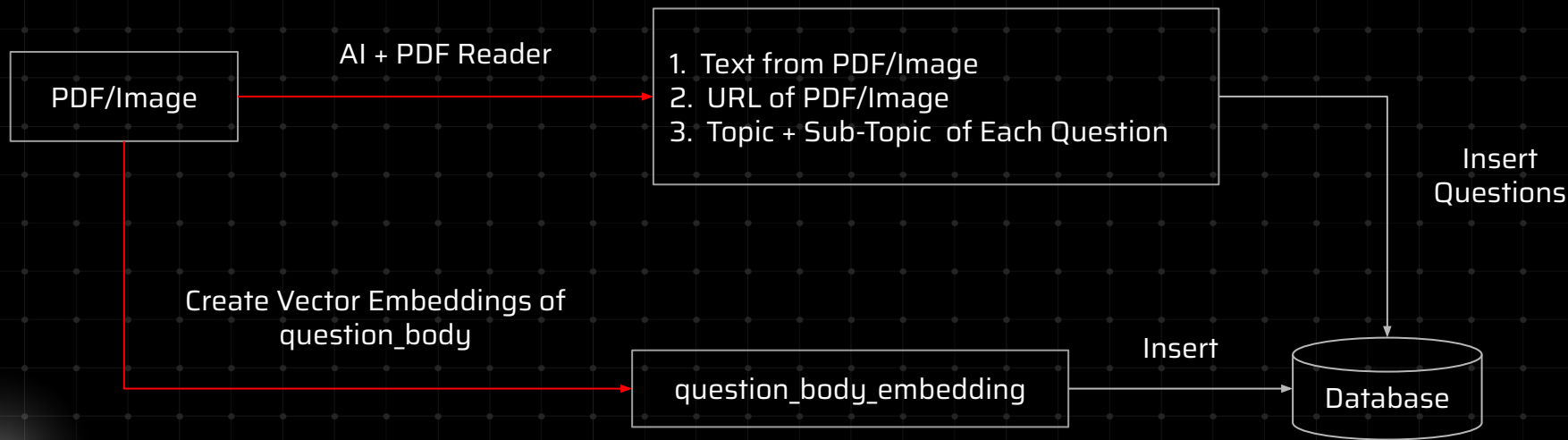
Meta-information of the question

331 questions in database across both subjects - our **citations** for generated questions

```
_id: ObjectId('674dcd0eb00b977d048c92aa')
page_start : 3
page_end : 3
question_number : 1
question_part : "1"
question_body : "It is given that  $a$  and  $b$  are the roots of the quadratic equation  $x^2 - 5x + 6 = 0$ "
question_body_embedding : Array (1536)
topic : "Quadratic Functions, Equations and Inequalities"
sub_topic : "Unknown"
answer_body : ""
question_paper_filepath : "https://document.grail.moe/ee5096dfab614b04b525f87a5..."
answer_paper_filepath : "https://document.grail.moe/06020d9323f04ea0bf13be6caab..."
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 : 2024-12-02T15:06:54.035+00:00
updated_utc : 2024-12-02T15:06:54.035+00:00
```

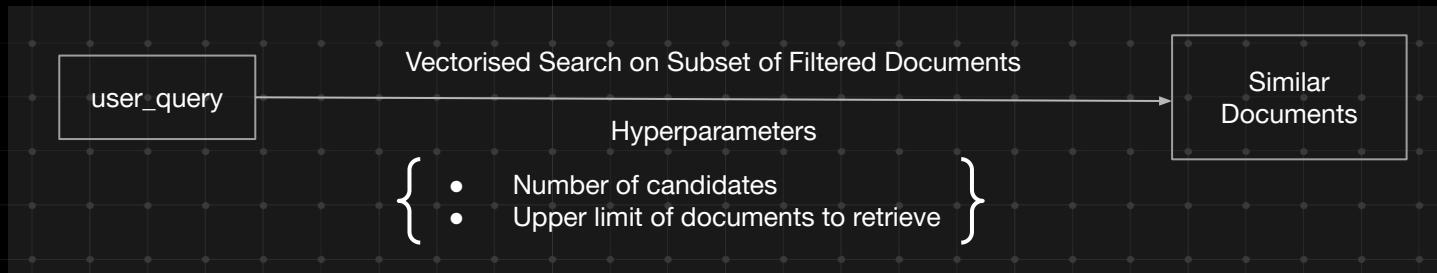


Data Collection & Preprocessing

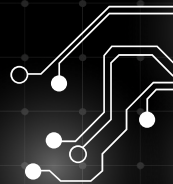
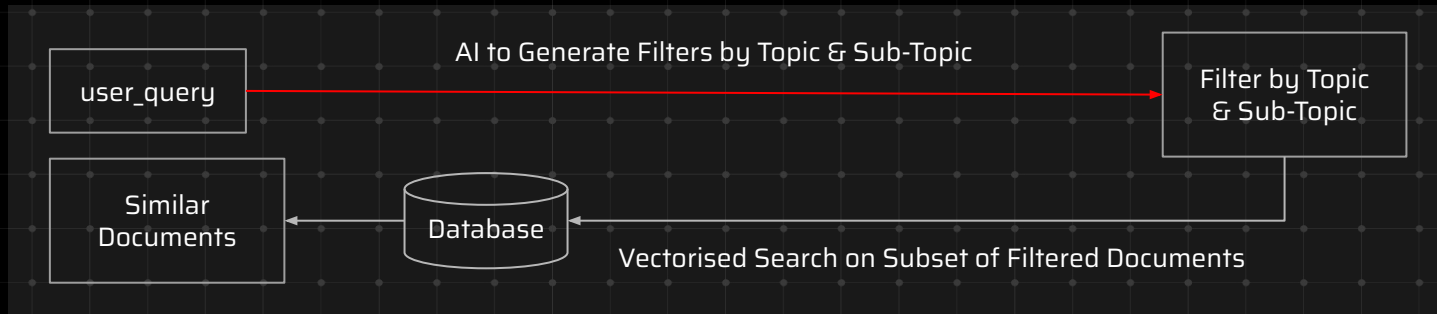


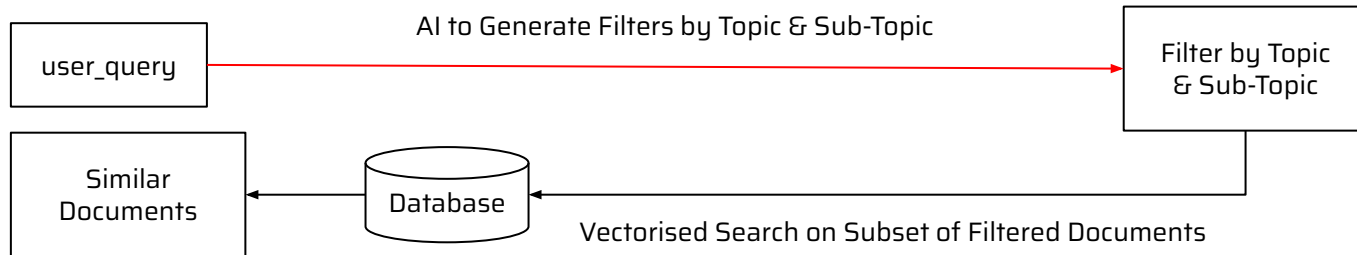
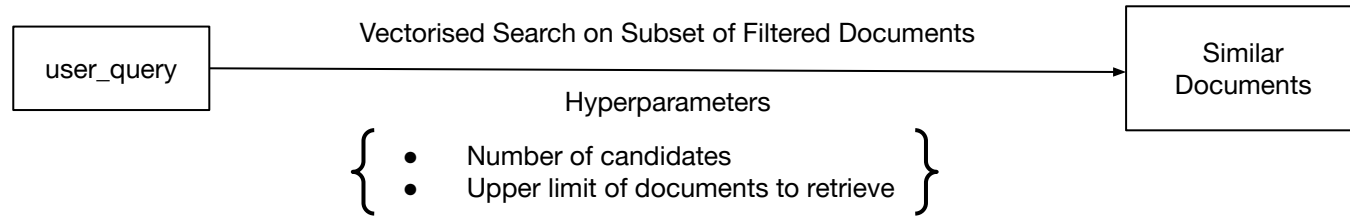
Retrieval

Vector Search



Hybrid Search





Evaluation of vector search

Ground truth dataset of 35 user queries with documents relevant to each, ranked in order of relevance

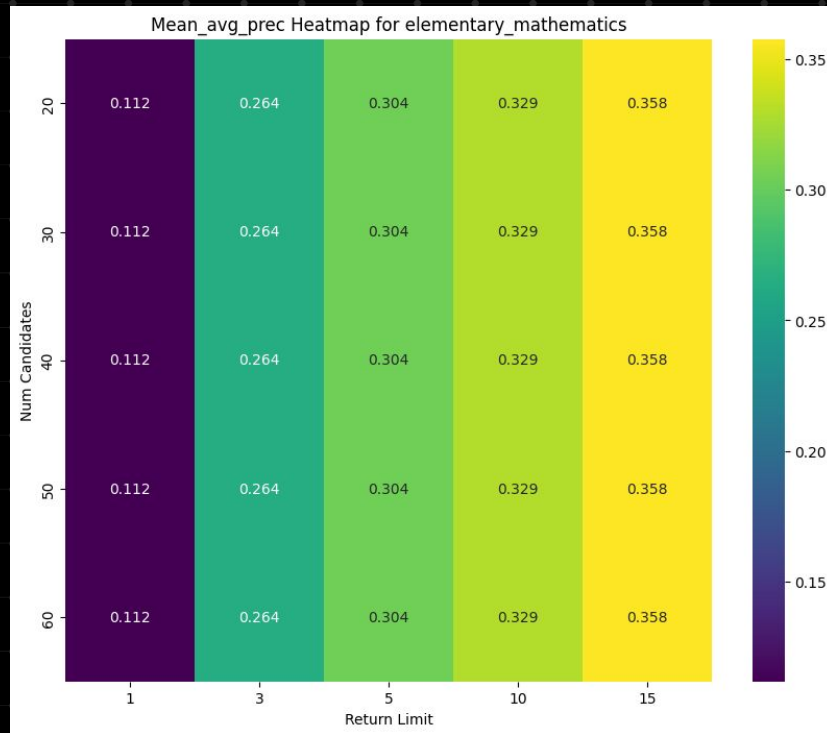
Grid Search across 2 hyperparameters:

- Number of candidates
- Upper limit of documents to retrieve

Metrics used:

- F1
- Jaccard Similarity
- Mean Average Precision
- Mean Reciprocal Rank

Number of candidates is a trivial hyperparameter for vector search



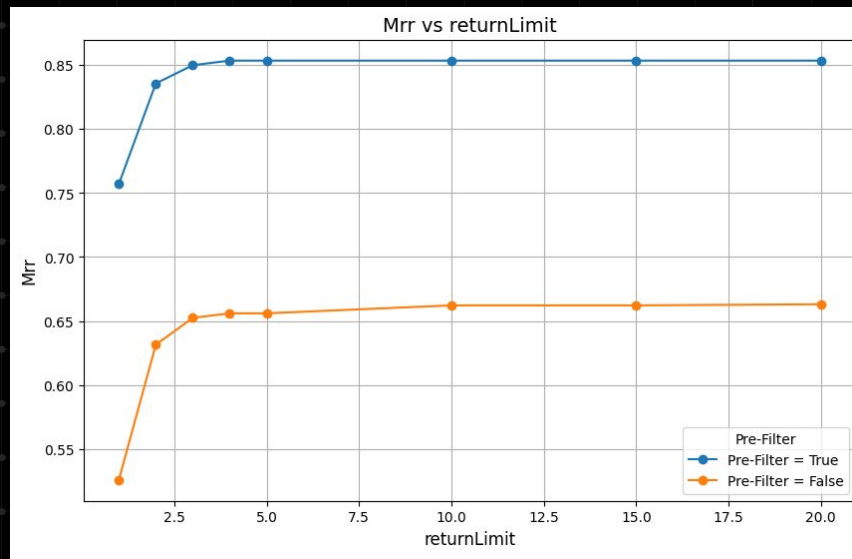
Evaluation of hybrid search

Compared hybrid search with vector search to evaluate the effect of pre-filtering of questions

Varied the upper limit of retrieved documents to identify optimal number

Optimal configuration for retriever:

- Hybrid search with pre-filtering
- Upper limit of documents to return = 10



HMW evaluate whether a question is answerable?

Generate Python script

Inputs:

1. Generated question
2. Generated answer

Ask LLM to generate Python script. Mark question as answerable iff. answer from Python script == generated answer

Step-by-step

Inputs:

1. Generated question
2. Generated steps
3. Generated answer

Ask LLM to evaluate if each step is correct before evaluating if the generated answer is correct

Black box

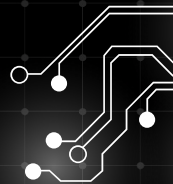
Inputs:

1. Generated question
2. Generated answer

Ask LLM to evaluate if generated answer is correct based on generated question

LLM Judges - Is “Python script” sufficient?

	Python script	Step-by-step	Black box	Remarks
Is “Python script” passing correctly?	PASS	FAIL	FAIL	No - need another LLM



LLM Judges - Is “Python script” sufficient?

Example **unanswerable** generated question:

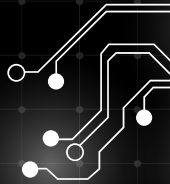
Matrix M represents the original prices of three items: toothpaste (\$5), shampoo (\$10), and soap (\$3). Write this as a **1x3 matrix M** . Matrix D represents the discount percentages applied at two different stores, where the first row shows a 15% discount and the second row shows a 5% discount. Write this as a **2x3 matrix D** . **Calculate the matrix product MD** to determine the discounted price for each item at both stores.

“Python script” erroneously marked question as answerable



LLM Judges - Is “step-by-step” useful?

	Python script	Step-by-step	Black box	Remarks
Is “step-by-step” failing correctly?	PASS	FAIL	PASS	Sometimes - keep the LLM



LLM Judges - Is “step-by-step” useful?

Example **unanswerable** generated question:

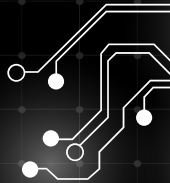
Charlotte has written five positive integers. The median of these numbers is 9, the mode is 10 and the mean is 12. The range of these numbers is 18. Find the five numbers.

ONLY “step-by-step” correctly marked question as unanswerable



LLM Judges - Is “black box” useful?

	Python script	Step-by-step	Black box	Remarks
Is “black box” failing correctly?	PASS	PASS	FAIL	No - discard it



HMW evaluate whether a question is answerable?

Generate Python script

Inputs:

1. Generated question
2. Generated answer

Ask LLM to generate Python script. Mark question as answerable iff. answer from Python script == generated answer

Step-by-step

Inputs:

1. Generated question
2. Generated steps
3. Generated answer

Ask LLM to evaluate if each step is correct before evaluating if the generated answer is correct

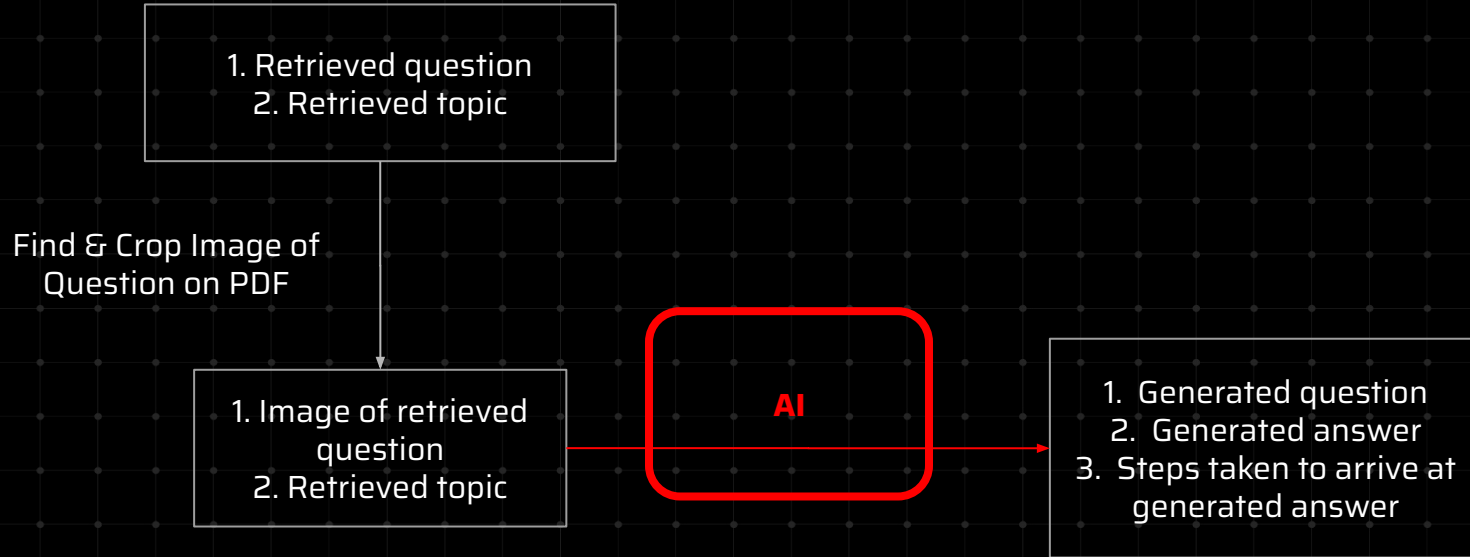
HMW maximise # of answerable questions?

Experiments

1. Plain text vs LaTeX vs LaTeX (format output)
2. # of retrieved documents
3. CoT (few-shot) vs CoT (zero-shot)
4. Text vs Multimodal generation



Plain text vs LaTeX



Plain text vs LaTeX

	LaTeX	Plain text
Format output	<p>Prompt: Generate question and answer in “LaTeX formatting”</p> <p>Output: Convert LaTeX to plain text</p>	<p>Prompt: Generate question and answer in “plain text”</p> <p>Output: Convert LaTeX to plain text</p>
Do not format output	<p>Prompt: Generate question and answer in “LaTeX formatting”</p>	<p>Prompt: Generate question and answer in “plain text”</p>



[Results] Plain text vs LaTeX

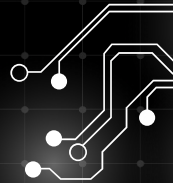
	LaTeX	Plain text
Format output	60.0% of test cases pass	54.3% of test cases pass
Do not format output	57.1% of test cases pass	57.1% of test cases pass



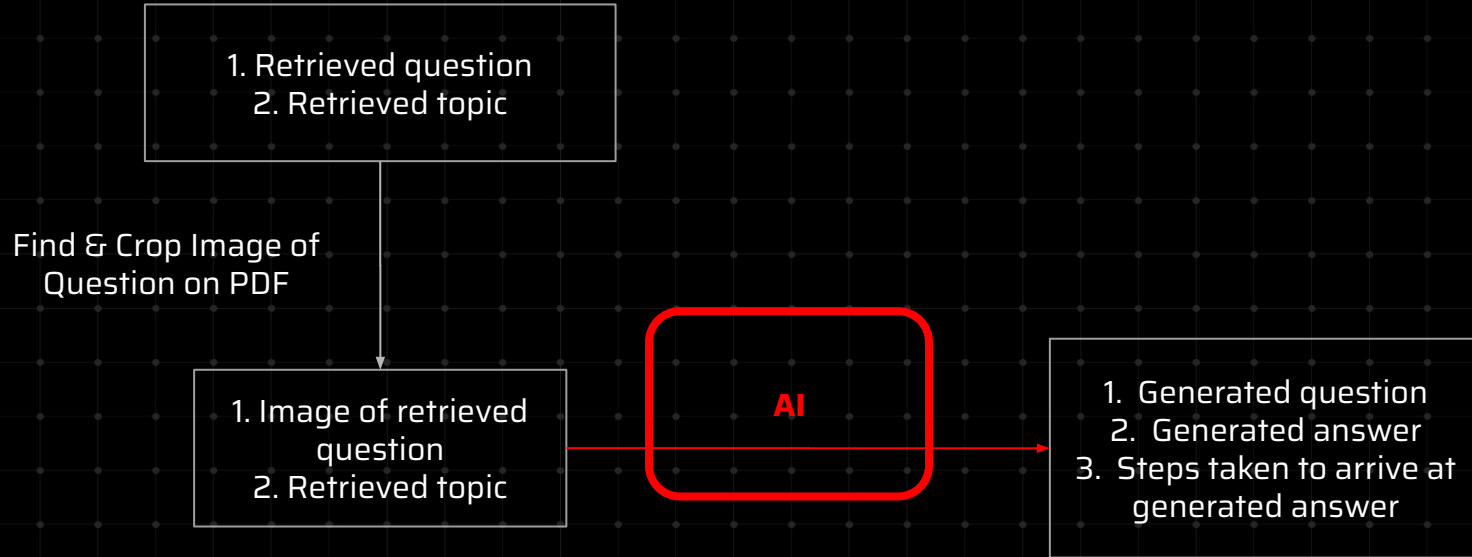
Grid search

Conducted **16** different experiments with 35 user queries each:

1. # of retrieved documents
 - a. From 1 to 4
2. Few-shot vs zero-shot prompts
3. Text vs Multimodal generation

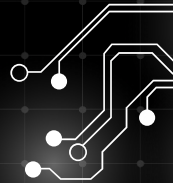


Few-shot vs zero-shot



Few-shot vs zero-shot

Few-shot	Zero-shot
<p data-bbox="218 388 917 472">Append 1 example question + answer to system prompt</p> <p data-bbox="218 525 942 656">Example question: Solve the simultaneous equations: $\begin{cases} 2x + 3y = 15 \\ -3y + 4x = 3 \end{cases}$</p> <p data-bbox="218 707 513 791">Example answer: $x = 3, y = 3$</p>	<p data-bbox="981 388 1043 423">NIL</p>

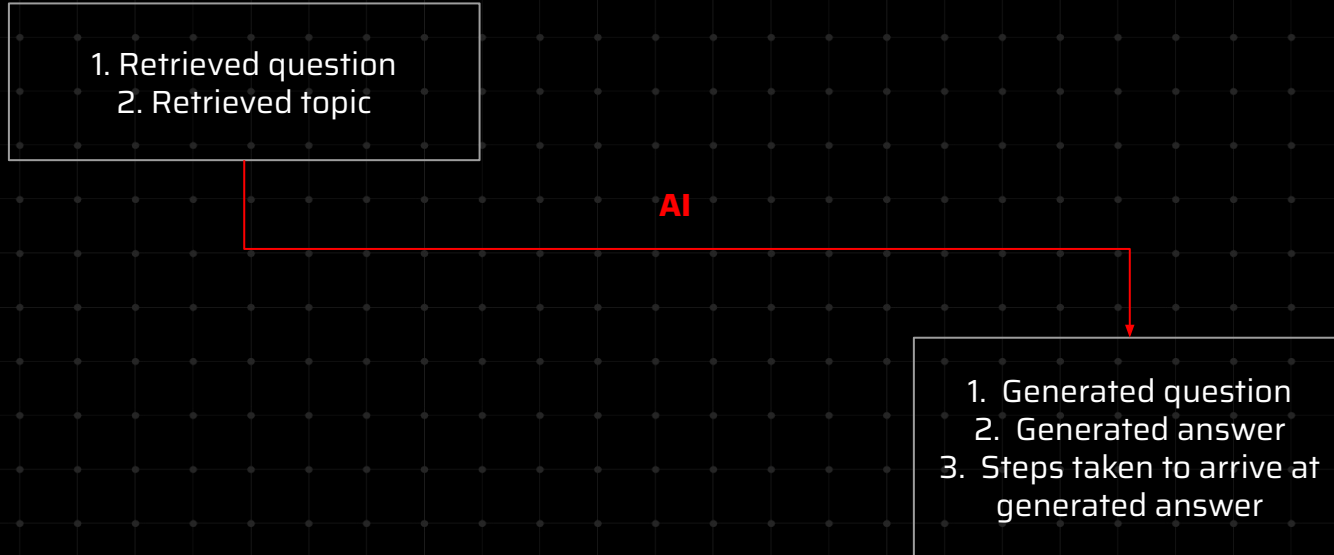


Text vs Multimodal

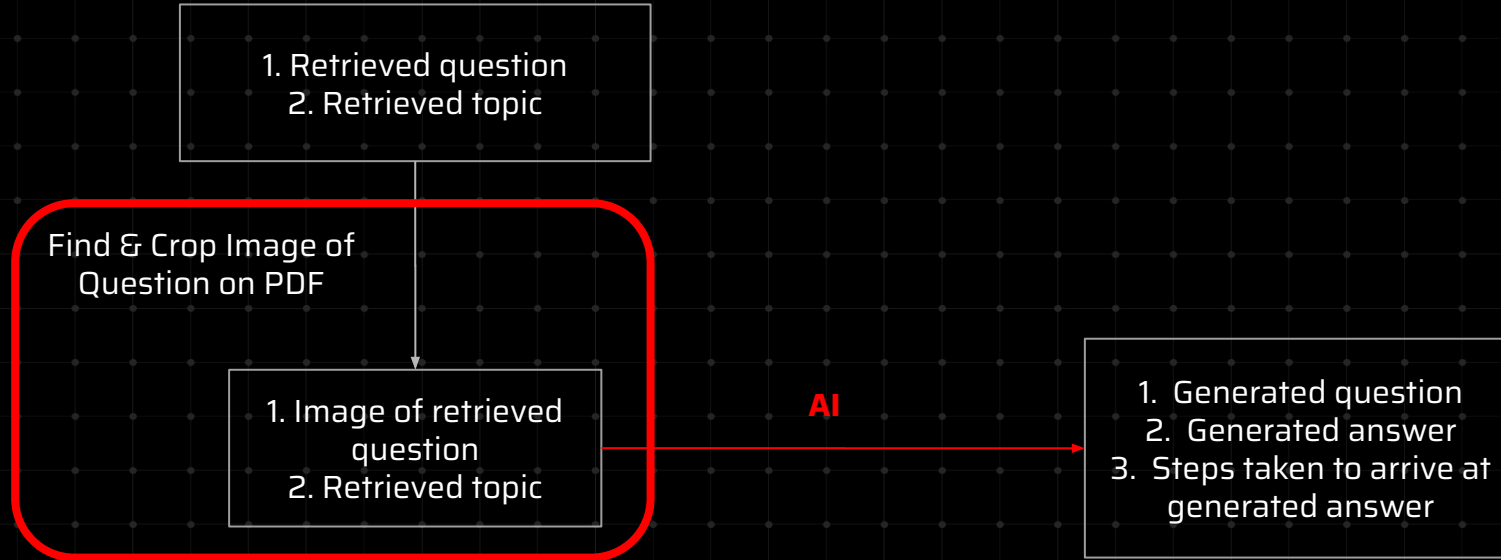
Text	Multimodal
<p>Text:</p> <ol style="list-style-type: none">1. Question2. Topic3. Sub-topic	<p>Image:</p> <ol style="list-style-type: none">1. Question <p>Text:</p> <ol style="list-style-type: none">1. Topic2. Sub-topic



Text pipeline

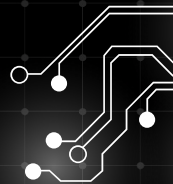


Multimodal pipeline



[Results] Grid search

Multimodal?	# of retrieved docs	Use few-shot?	% of tests pass
NO	1	YES	62.86%
NO	1	NO	60.00%
YES	1	NO	60.00%
YES	3	YES	60.00%



How does % of answerable questions differ based on question difficulty?

Conducted **2** different experiments with 35 user queries each:

1. E-Math vs A-Math
 - a. Subject used as proxy for question difficulty



TRUST-Score



Grounded Refusal

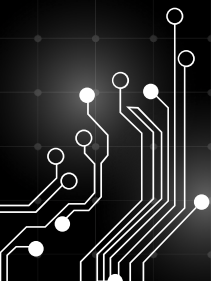
How capable is the model of refusing to answer user queries which are not supported by our scope?

Math Validity

How capable is the model of generating valid math questions that are actually answerable and solvable?

Citations Groundedness

How capable is the model of generating new math questions that are grounded by existing, real-world questions?



Grounded Refusal

- Added random questions unrelated to math questions into our evaluation dataset
 - No relevant documents
 - Randomly generate some irrelevant documents
- Utilised TRUST-Align function to calculate
 - Quality of Answering
 - Quality of Refusal
 - F1 of both
- Preliminary Results:
 - A-Math: 100%
 - E-Math: 93.2%

```
"question": "I need some exercises involving tangents and circles; can you help?",
"answers": [],
"docs": [
  {
    "title": "674dcd43b00b977d048c92f0",
    "text": "The equation of a circle is  $x^2 + y^2 - 4x - 12y + 36 = 0$ . The equation of",
    "answers_found": [
      1
    ]
  },
  {
    "title": "674dcd2cb00b977d048c92d9",
    "text": "The diagram shows a circle, centre O. Points A, B, D and E lie on the circle",
    "answers_found": [
      1
    ]
  },
  {
    "title": "674dcd10b00b977d048c92ad",
    "text": "A circle with centre O passes through the points P(-1, 7) and Q(0, 8).",
    "answers_found": [
      0
    ]
  },
  {
    "title": "674dce8db00b977d048c93c6",
    "text": "Given that  $\begin{pmatrix} 1 & 2 & 0 \\ 3 & 0 & 3 \\ 2 & 2 & 1 \end{pmatrix}$ ",
    "answers_found": [
      0
    ]
  }
]
```

Grounded Citations

- Each user query only generates one statement - the generated math question
 - unlike typical RAG evaluation cases
 - Hence length of S is only 1
- Lack of access to GPU -> unable to run NLI models from HuggingFace
 - Used GPT-4o for entailment classification

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

$$F1_{CG} = \frac{2 \cdot CP \cdot CR}{CP + CR}$$

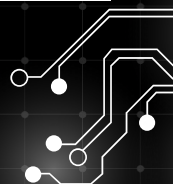
A_r Set of questions where model provided an answer

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



Limitations

Lack of GPU access

- Relied on OpenAI APIs for our project
- Unable to run NLI models locally, and reliance on GPT-4o is extremely costly (50 questions * 10 documents = **500** API calls / experiment)

Limited Relevant Data

- Questions are relevant up until the last syllabus change
- Requires us to keep track of MOE syllabus updates for each subject

Lack of Existing Datasets

- Goal was to tackle a niche, local use case and create something useful
- A lot of time dedicated to creating our database and evaluation datasets

Stochasticity of LLMs

- The current state of LLMs do not provide a foolproof, guaranteed verification of math problems and solutions
- Human intervention will always be required to confirm if a question is answerable

Future Work

Expand Pool of Existing Questions

- Ingest more questions from Holy Grail website for both subjects
- Increase number of subjects
- Develop script that automates this process

More Generated Questions

- Currently limited to 1 generated question per user query to save costs and complexity
- More generated questions can be useful for practice

Diversity in generated questions

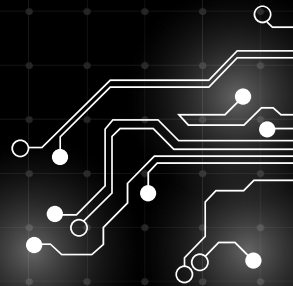
- Ensuring that new questions differ from existing questions to a certain extent
 - instead of just swapping values in a question

Optimising for TRUST

- Due to costly nature of evaluating TRUST-Score with GPT-4o, we did not conduct it for all experiments; only for the 4 best
- With GPU access, we can optimise our RAG hyperparameters for TRUST



Demo Time!





Thank You!

Team **ZSR**:

Tee **Zhi** Zhang (1005136), **Sean** Chen Zhi En (1005122), **Reina** Peh (1005359)

