FINAL PROJECT 2024

DEEP LEARNING FOR DATA SCIENCE

Eedi

Mining Misconceptions in Mathematics

deepstudy

21120035
Nguyễn Hoài An

21120103
Phan Thảo Nguyên

21120179
Nguyễn Đặng
Đăng Khoa

21120275
Huỳnh Cao Khôi

21120308
Phạm Lê Tú Nhi

Content

1	Introduction	What is the competition description?
2	Data Discovery	What insights can be gained from the data?
3	Model Building	What model do we use to solve the problem?
4	Experimentation	How did the proposed solution perform?
5	Conclusion	Project summary

01
Introduction

Eedi: Mining Misconceptions in Mathematics

Predict affinity between misconceptions and incorrect answers (distractors) in multiple-choice questions

Purpose: To develop an NLP model capable of identifying misconceptions reflected in incorrect choices within mathematical multiple-choice questions. Developing such model would assits human labelers in accurately selecting suitable misconceptions from existing and newly identified options.

02
Data Discovery

Data generation process

How is the data collected?

The data is taken from Eedi - a learning platform where students answer Diagnostic Questions(DQ). DQ are multiple-choice questions featuring

- 1 correct
- 3 incorrect answers

Dataset overview

Misconception mapping overview

Number of rows: 2587

2 columns:

- MisconceptionId: Used for mapping for train dataset
- MisconceptionName: Corresponding meaning for each Misconception

Dataset overview

Train dataset overview

Number of rows: 1869

15 columns, in which:

- 7 Columns are ID columns for ConstructName, SubjectName, QuestionText, Misconception[A/B/C/D]
- 8 Columns are value columns for ConstructName, SubjectName, Question Text, Correct Answer, Answers[A/B/C/D]

Column	Unique values
Construct Name	757
Subject Name	163
Question Text	1857
Misconception [A/B/C/D]	1604

Semantic columns in the dataset

Dataset overview

Train dataset overview

Number of rows: 1869

15 columns, in which:

- 7 Columns are ID columns for ConstructName, SubjectName, QuestionText, Misconception[A/B/C/D]
- 8 Columns are value columns for ConstructName, SubjectName, Question Text, Correct Answer, Answers[A/B/C/D]

Column	Unique values
Construct Name	757
Subject Name	163
Question Text	1857 contains duplications
Misconception [A/B/C/D]	1604 62% misconceptions

Semantic columns in the dataset

Our approaches

We takes 2 approaches to doing **Data Discovery**

1. Standard statistical analysis for NLP
2. Further analysis with semantical clustering using LLM

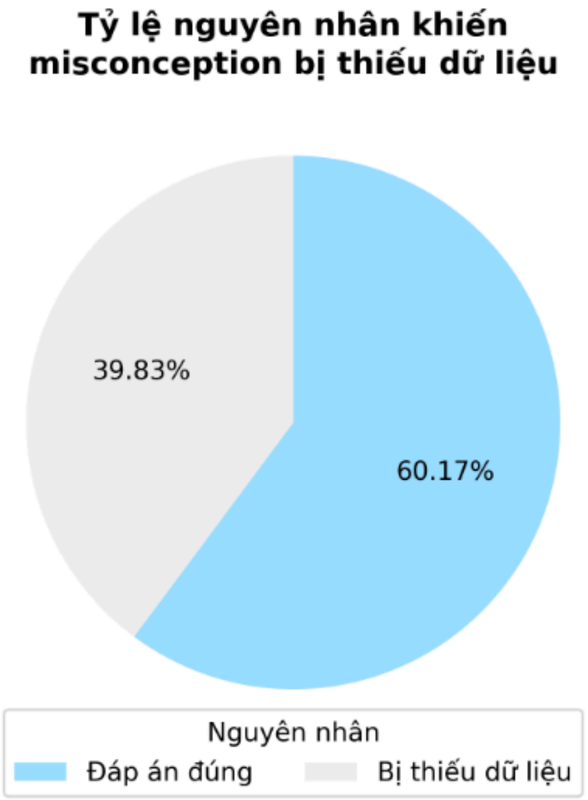
First approach: Statistical Analysis for NLP

Misconception in train dataset

Question 01: Why are there missing misconception in the data?

Misconception in **train dataset**

Missing Misconceptions are either because correct answer, or because they really are misisng.

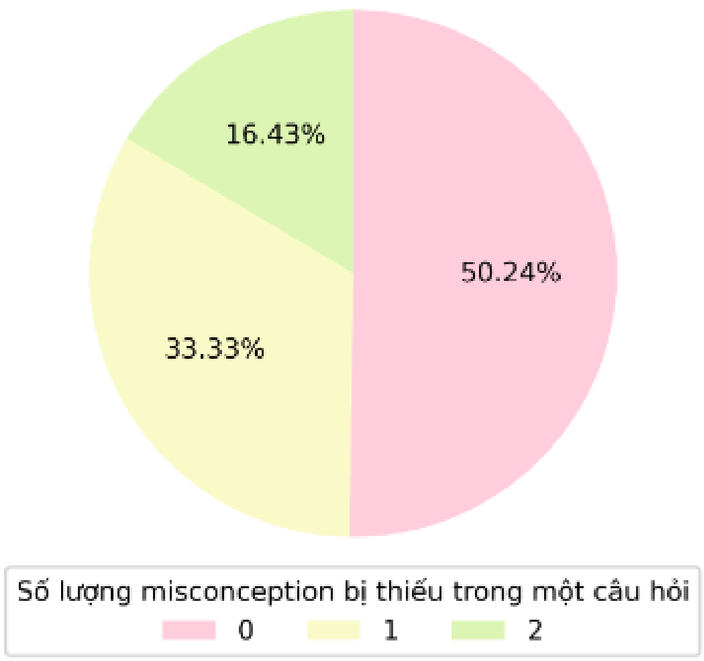


Question 02: What is the ratio of missing Misconception in a question?

Misconception in **train dataset**

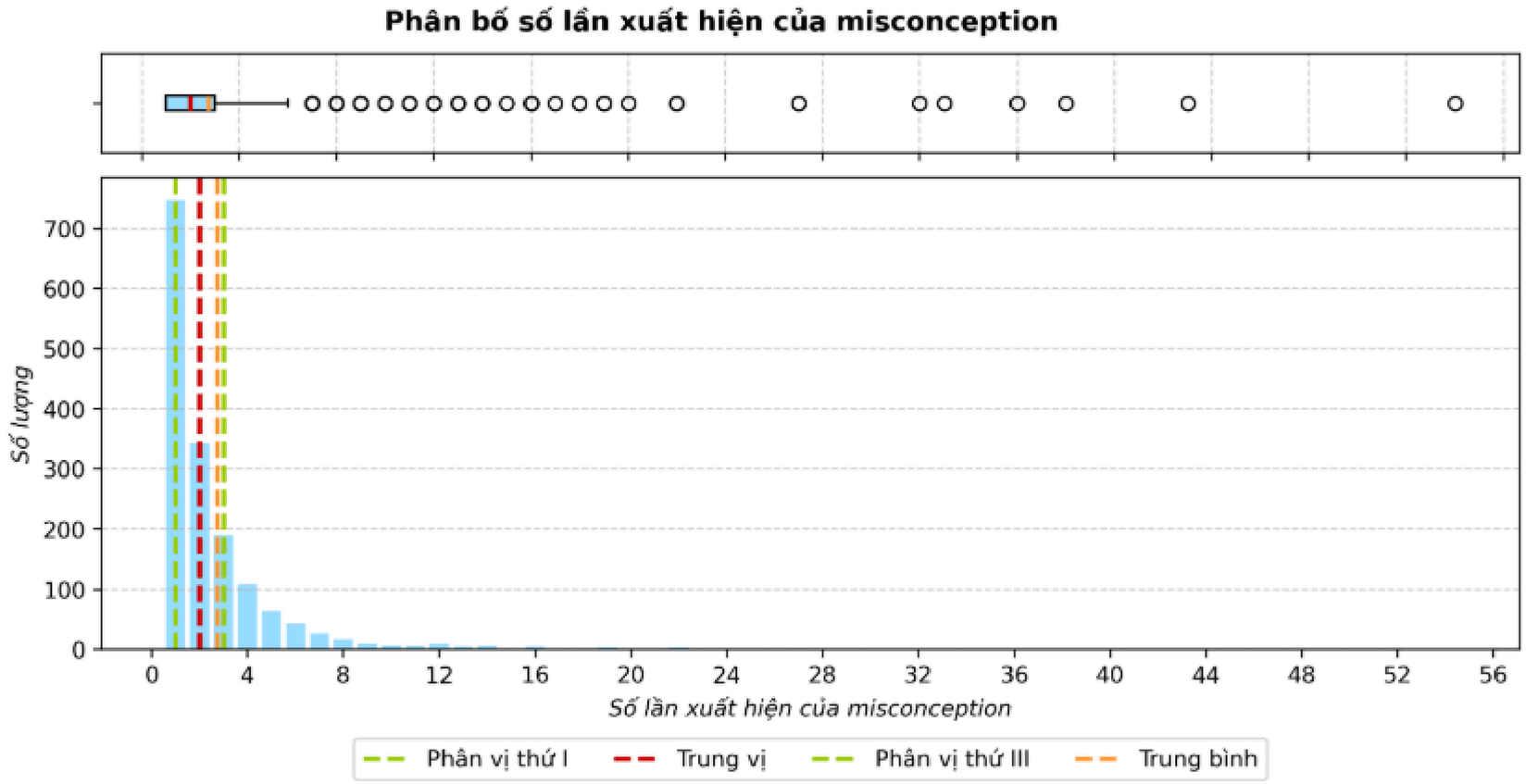
There are no Questions with 0 Misconceptions

Tỷ lệ số lượng misconception trong các câu hỏi



Question 03: What is the distribution of Misconception appearance?

Misconception in **train dataset**

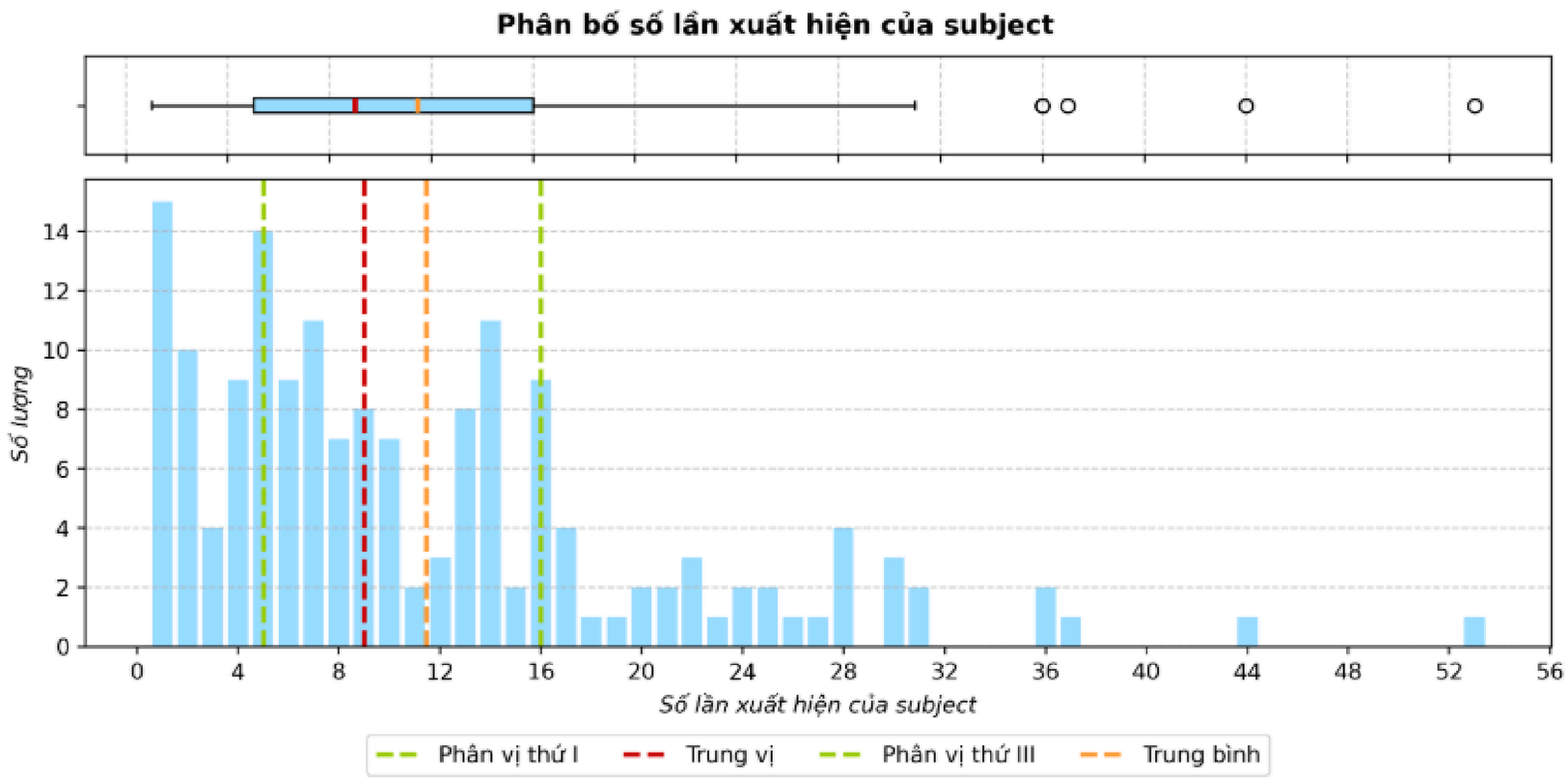


Most misconceptions only appear **once**

Subject in train dataset

Question 04: What is the distribution of Subject appearance?

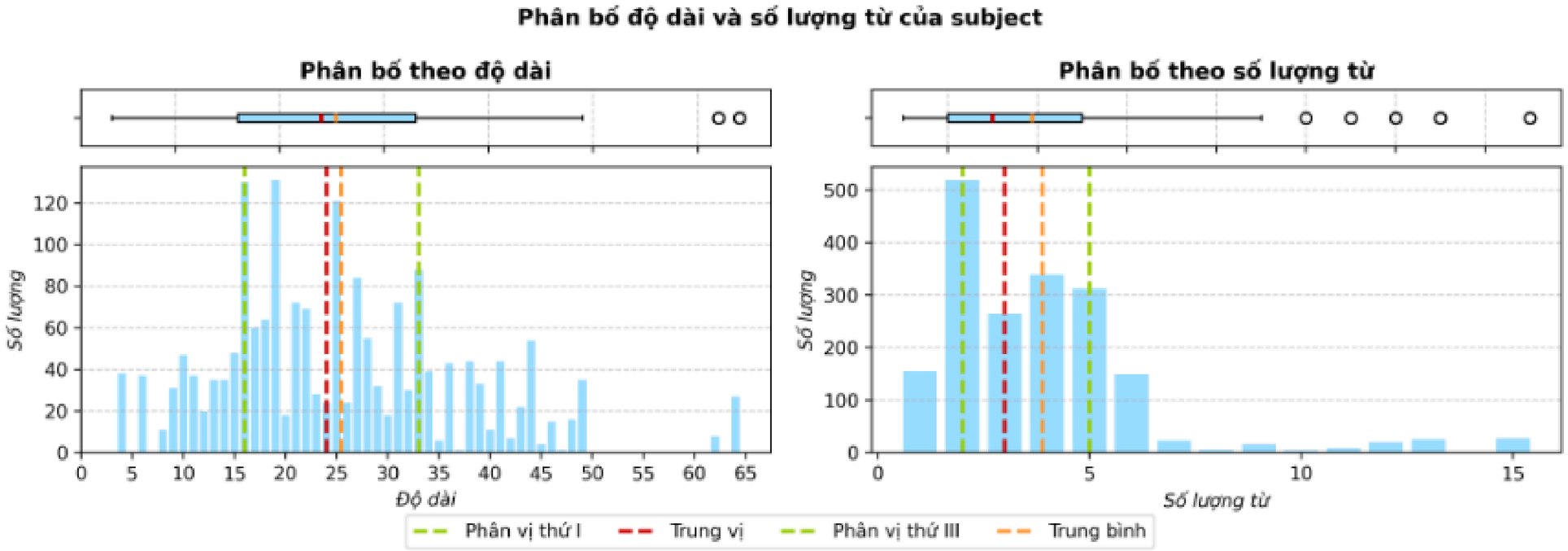
Subject in **train dataset**



Mean and medium both doesn't represent distribution centrality

Question 05: What is the distribution of length and word count of Subject?

Subject in **train dataset**



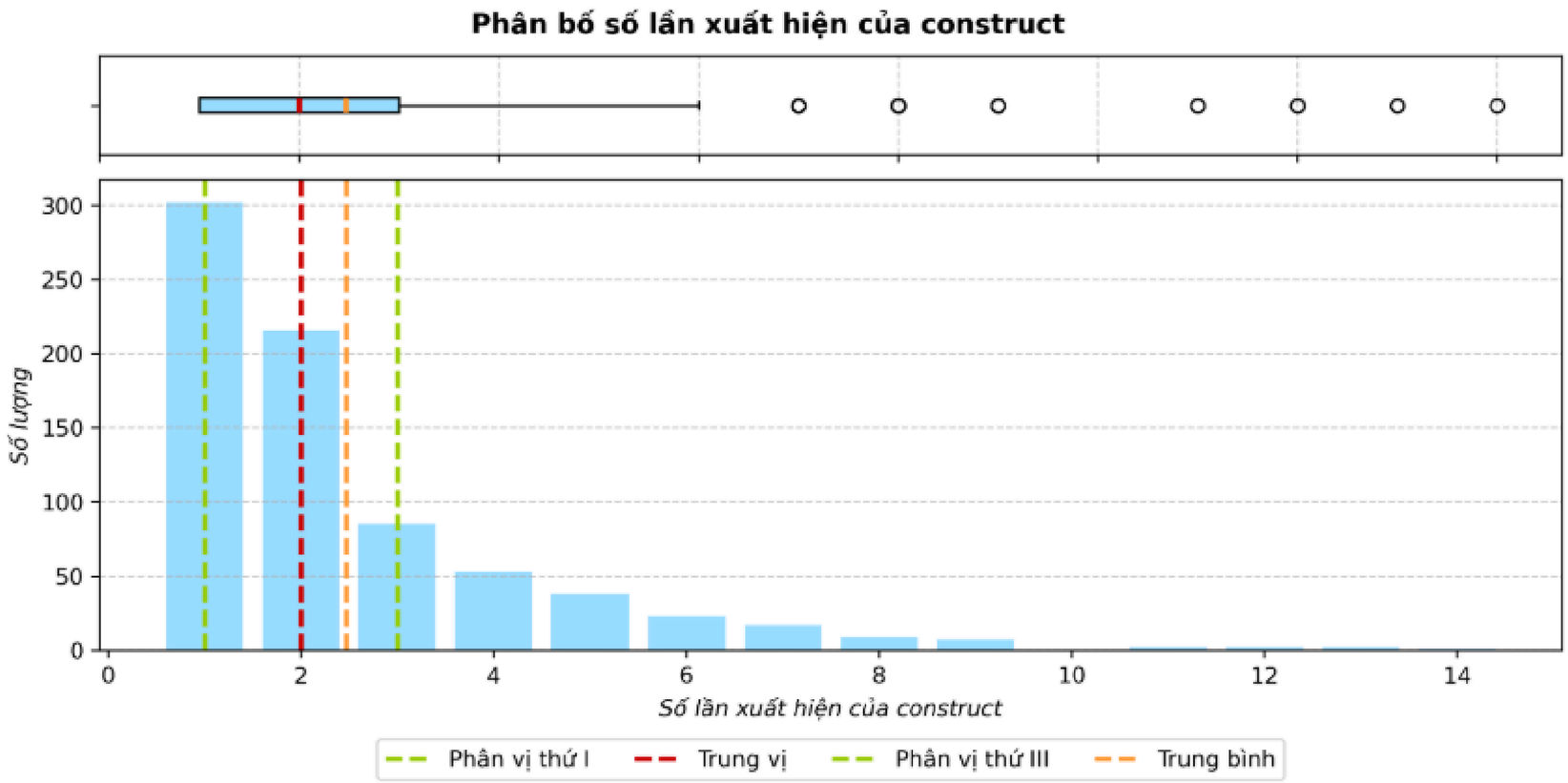
Both have long concentrated range and many outliners

First approach: Statistical Analysis for NLP

Construct in train dataset

Question 06: What is the distribution of Construct appearance?

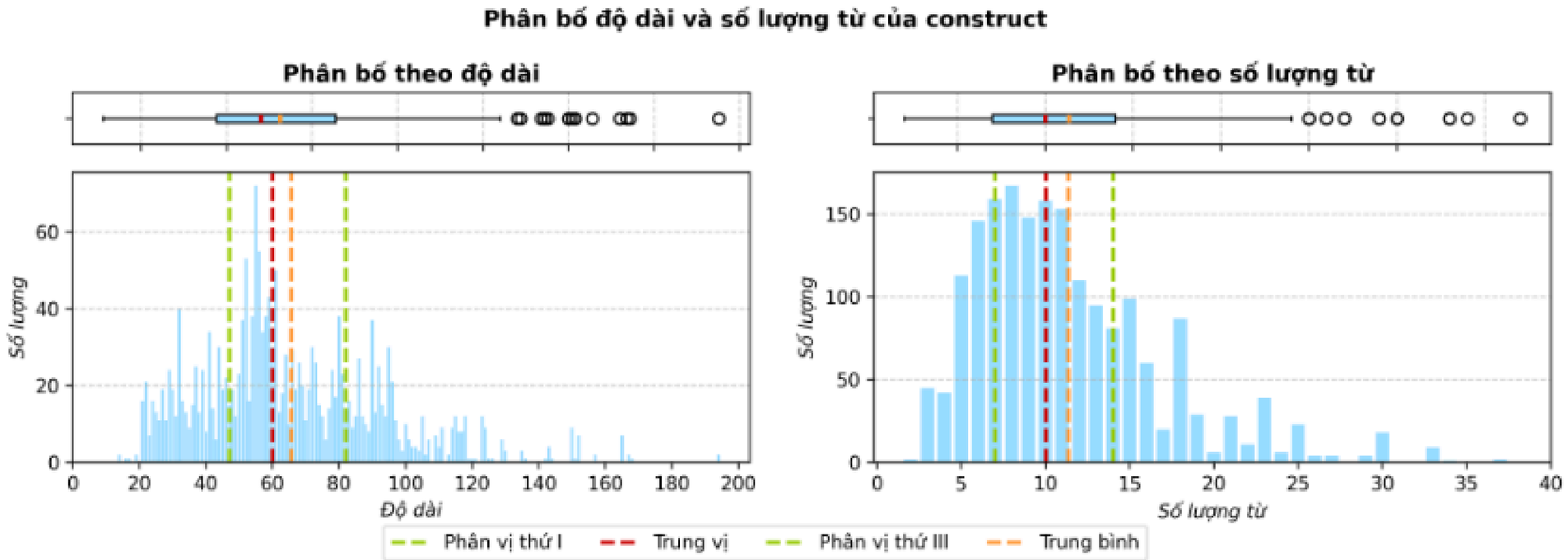
Construct in **train dataset**



Most Construct appears **1-2 times**

Question 07: What is the distribution of length and word count of Construct?

Construct in **train dataset**

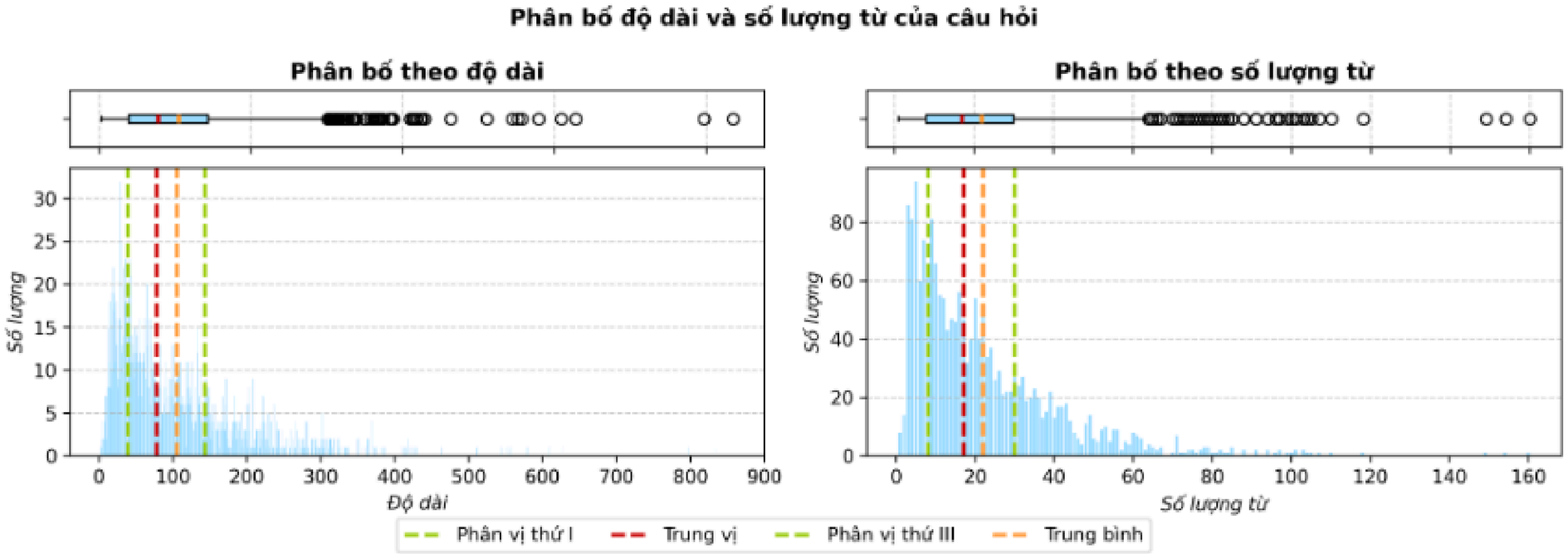


Both have long concentrated range and many outliners

Questions in train dataset

Question 08: What is the distribution of Question's length and word counts?

Question in **train dataset**



Both are right-skewed

Question 09: How complicated is the language used in Question?

Misconception in **train dataset**

We use **The Flesch Reading Ease** to calculate the readability of Questions.

0-30 College Graduate → 90 - 100 Grade 5

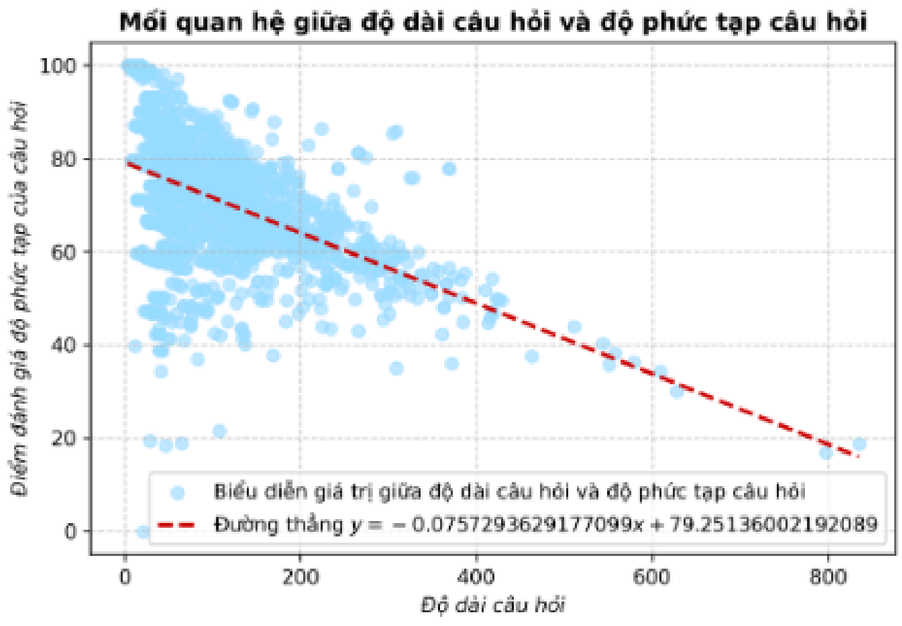
+

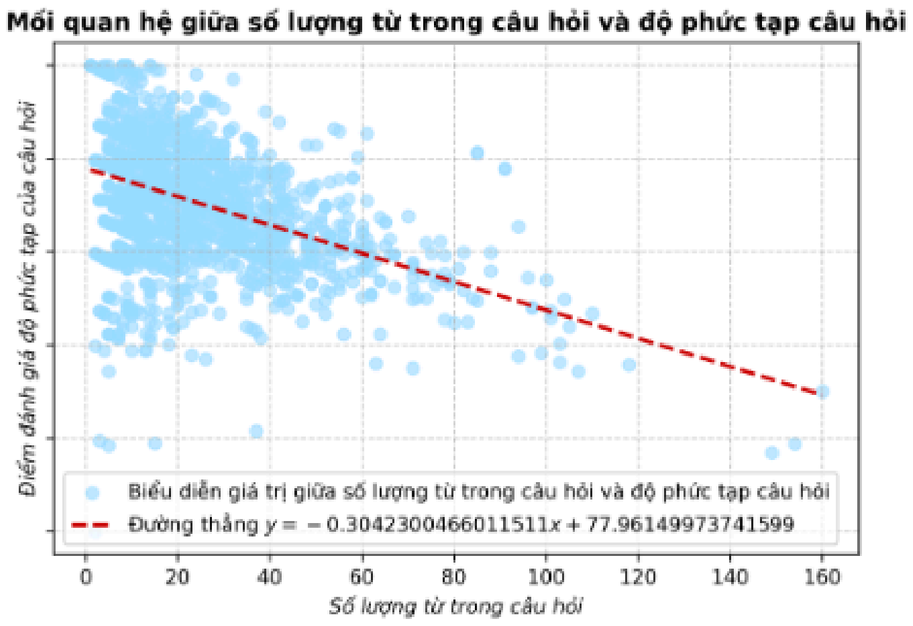
Process the Question text:

- Transform/Remove Latex notations
- Remove white space, stop words
- Lemmatize

Question 09: How complicated is the language used in Question?

Question in **train dataset**





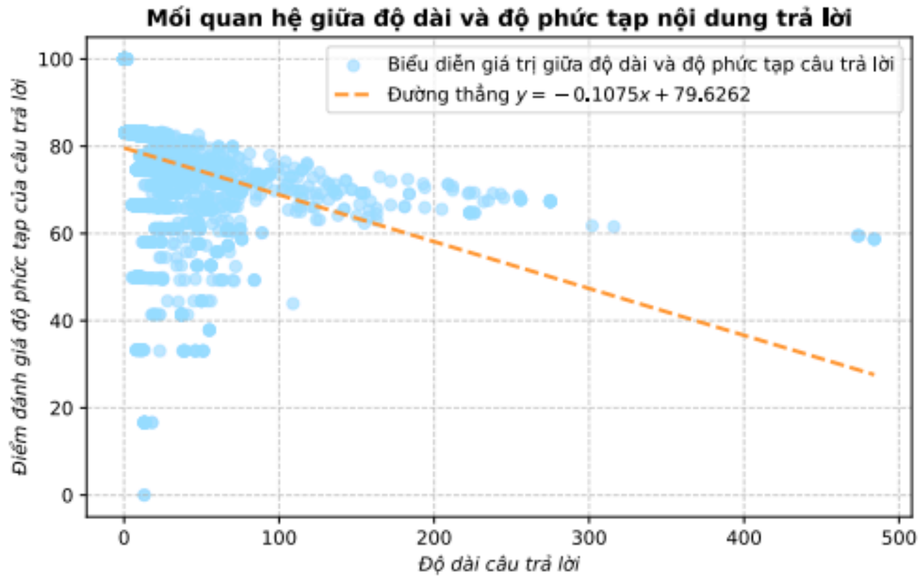
The longer the Question, the more complex it seems to be

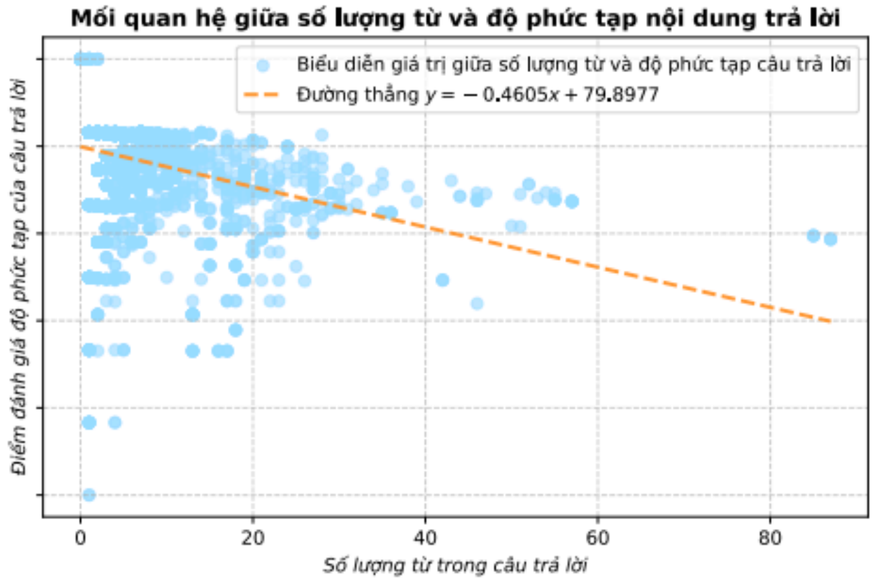
First approach: Statistical Analysis for NLP

Answer in train dataset

Question 10: How complicated is the language used in Answers?

Answer in **train dataset**



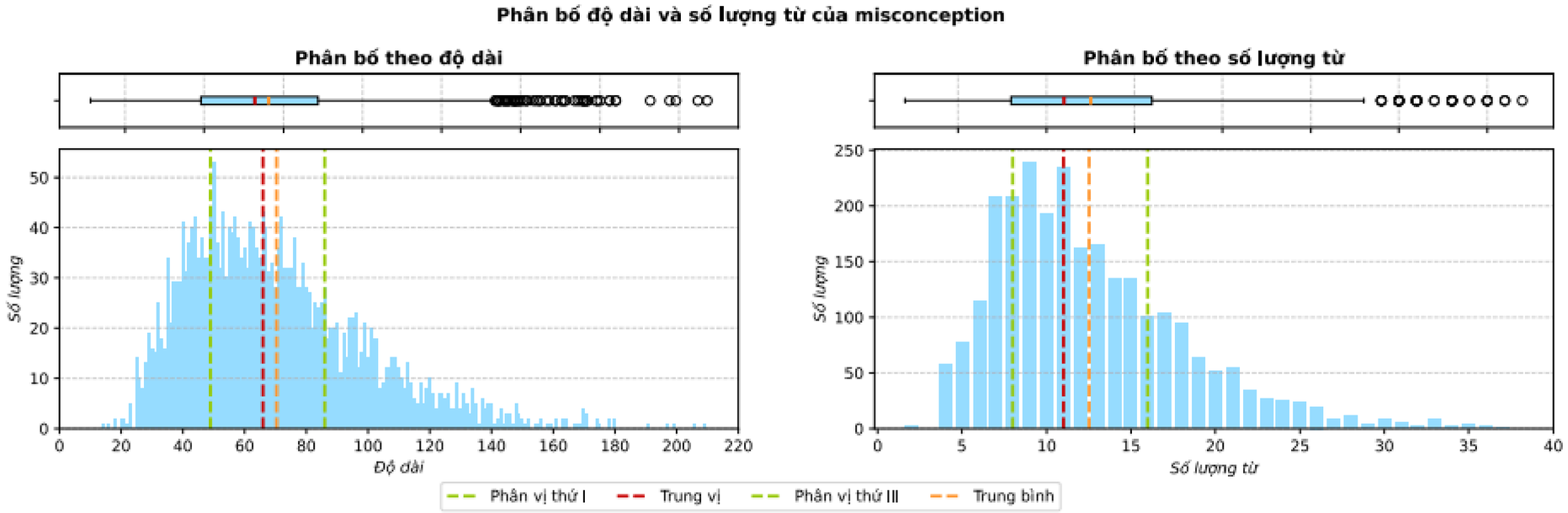


The longer the Answer, the more complex it seems to be

Misconception in Misconception mapping

Question 11: What is the distribution of Misconception's length and word counts?

Misconception in **Misconception mapping**



Both have long concentrated range and many outliners

Second approach: Semantic Analysis with LLM clustering

LLM Clustering Process

How do we use LLM for semantic clustering?

Data Preperation

Prepare data that fit into LLM context window

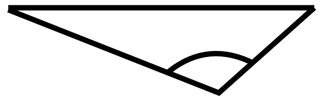
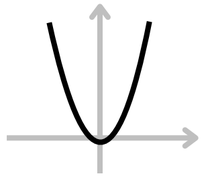
↓

Input data → Claude 3.5 Sonnet

Iterate

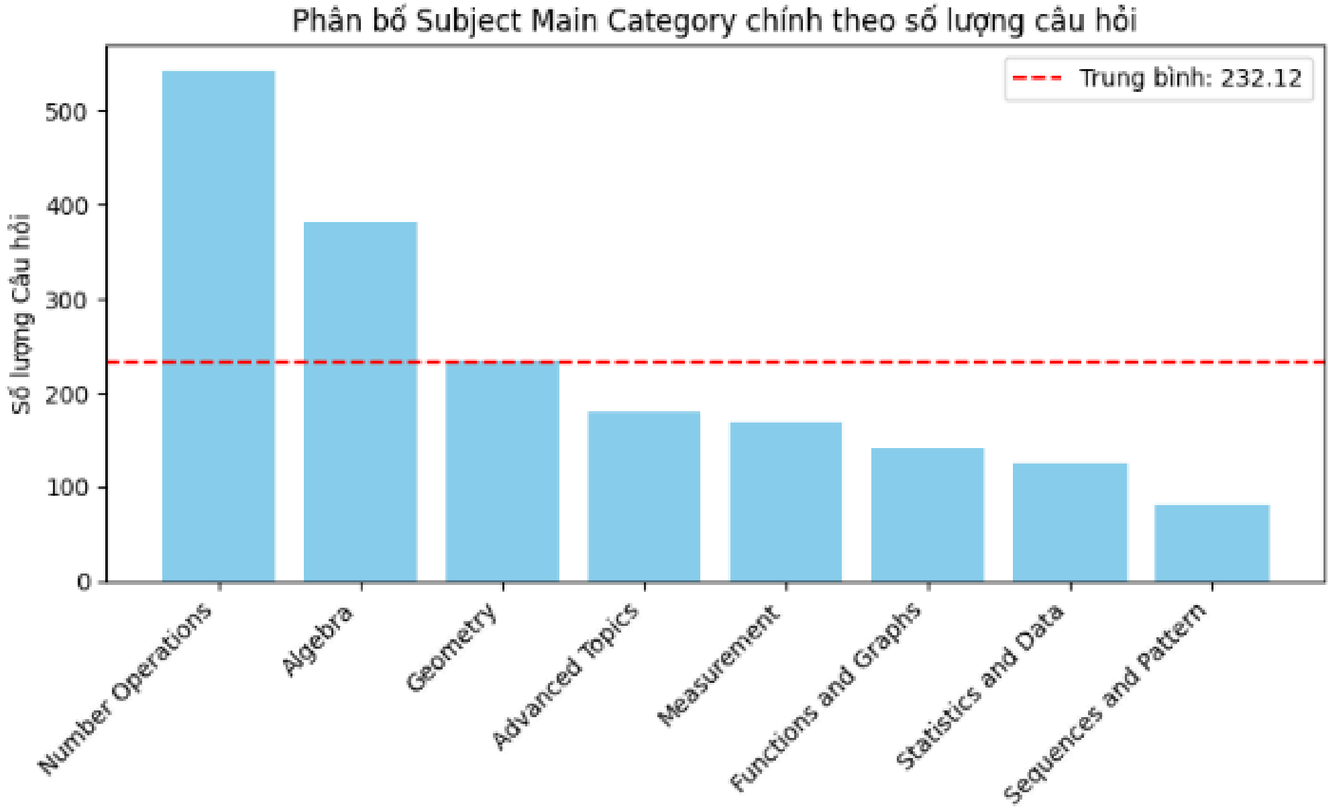
Subject Name Analysis

What are the Mathematic topics that exist in the dataset?

2 + 3 = 5 Number Operations	2**x** + 5 = 0 Algebra	 Gepmetry	 Functions and Graphs
 Measurement	 Statistics and Data	2,4,8,16,32... Sequence and Pattern	sin(x), cos(x)... Advanced Topics

Question 12: How is Subject Main Category distributed?

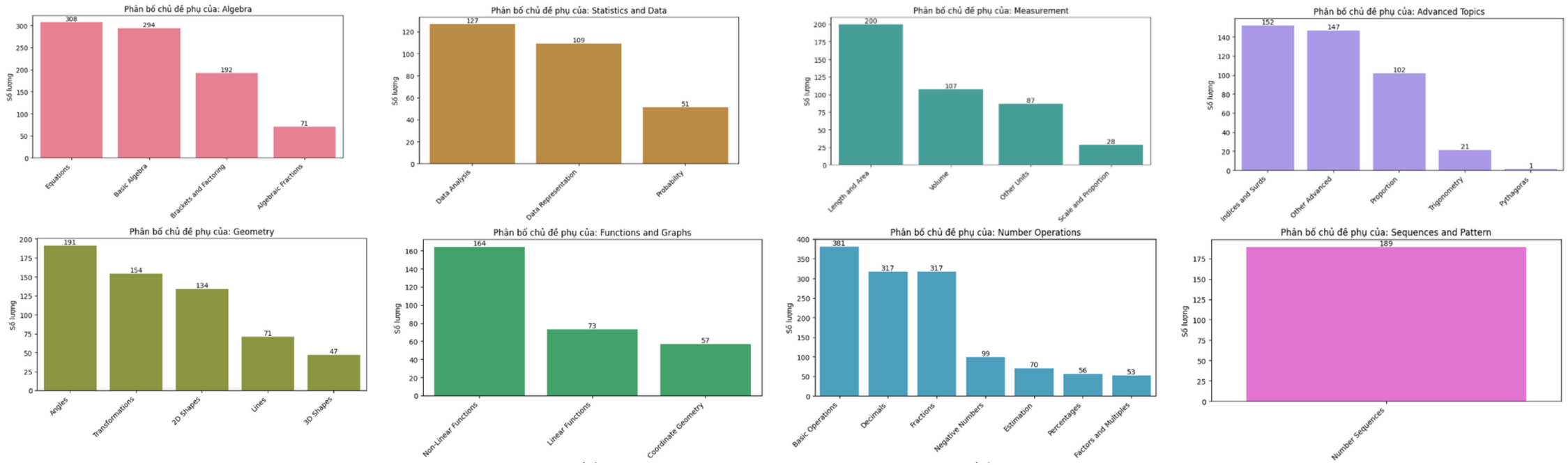
Subject Category Analysis



50% of questions are either in Number Operations or Algebra

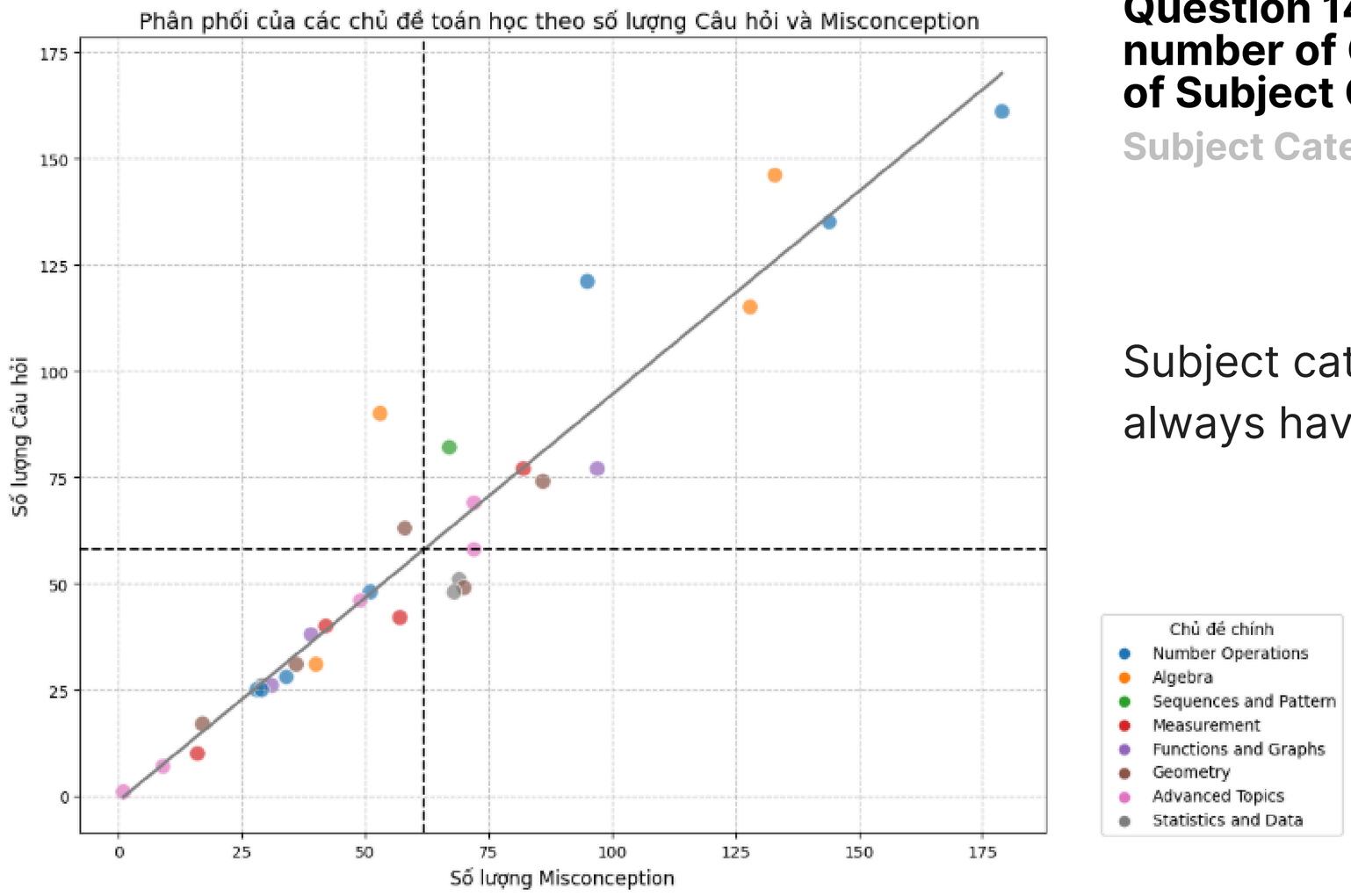
Question 13: How is Subject Sub Category distributed?

Subject Category Analysis



Almost no Subject Sub Categories are distributed evenly

Second approach: Semantic Analysis with LLM clustering



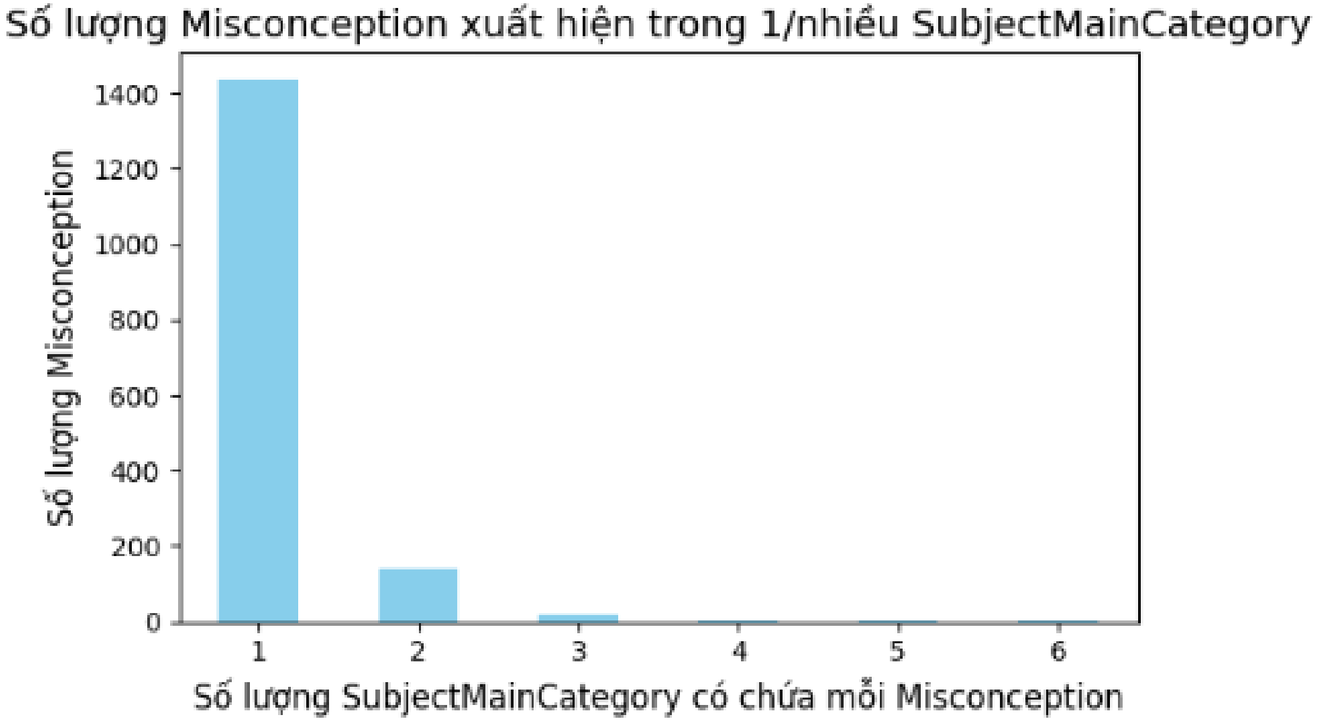
Question 14: What is the correlation of number of Questions and Misconceptions of Subject Categories?

Subject Category Analysis

Subject categories with more Quesitions always have more Misconceptions

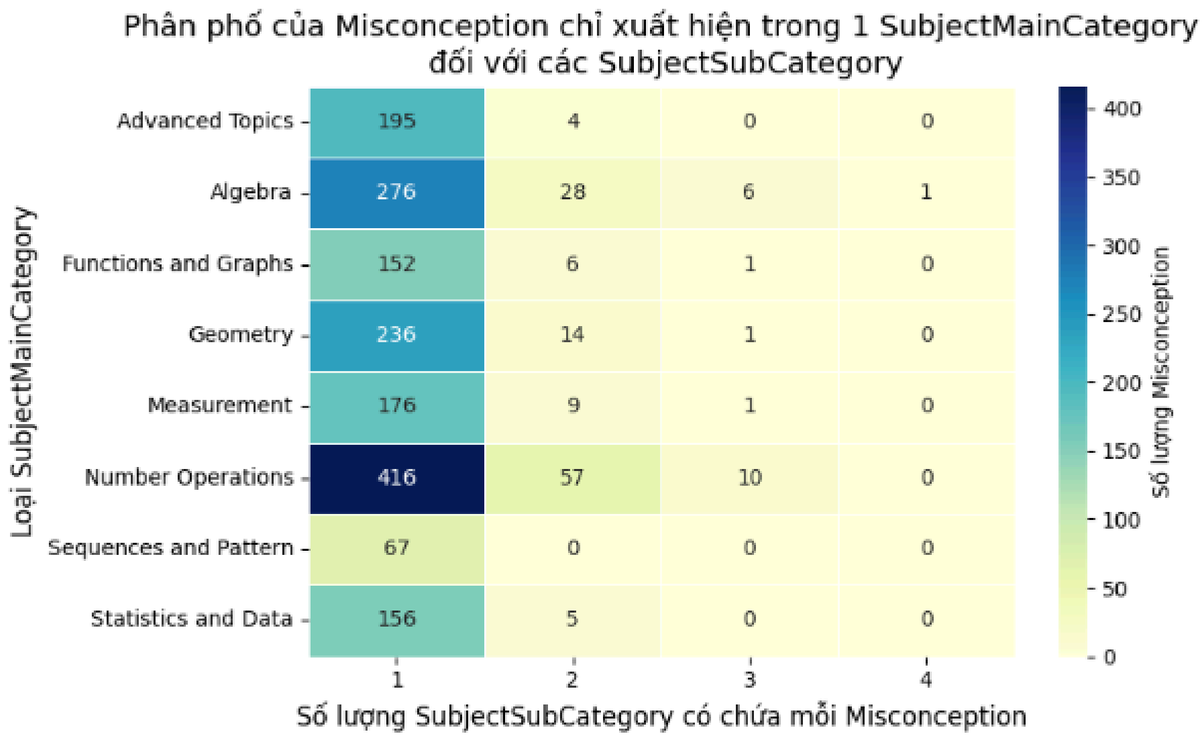
Question 15: How many misconceptions are there across any number of subject categories?

Subject Category Analysis



Most Misconceptions only appears in 1 Subject Category

Second approach: Semantic Analysis with LLM clustering

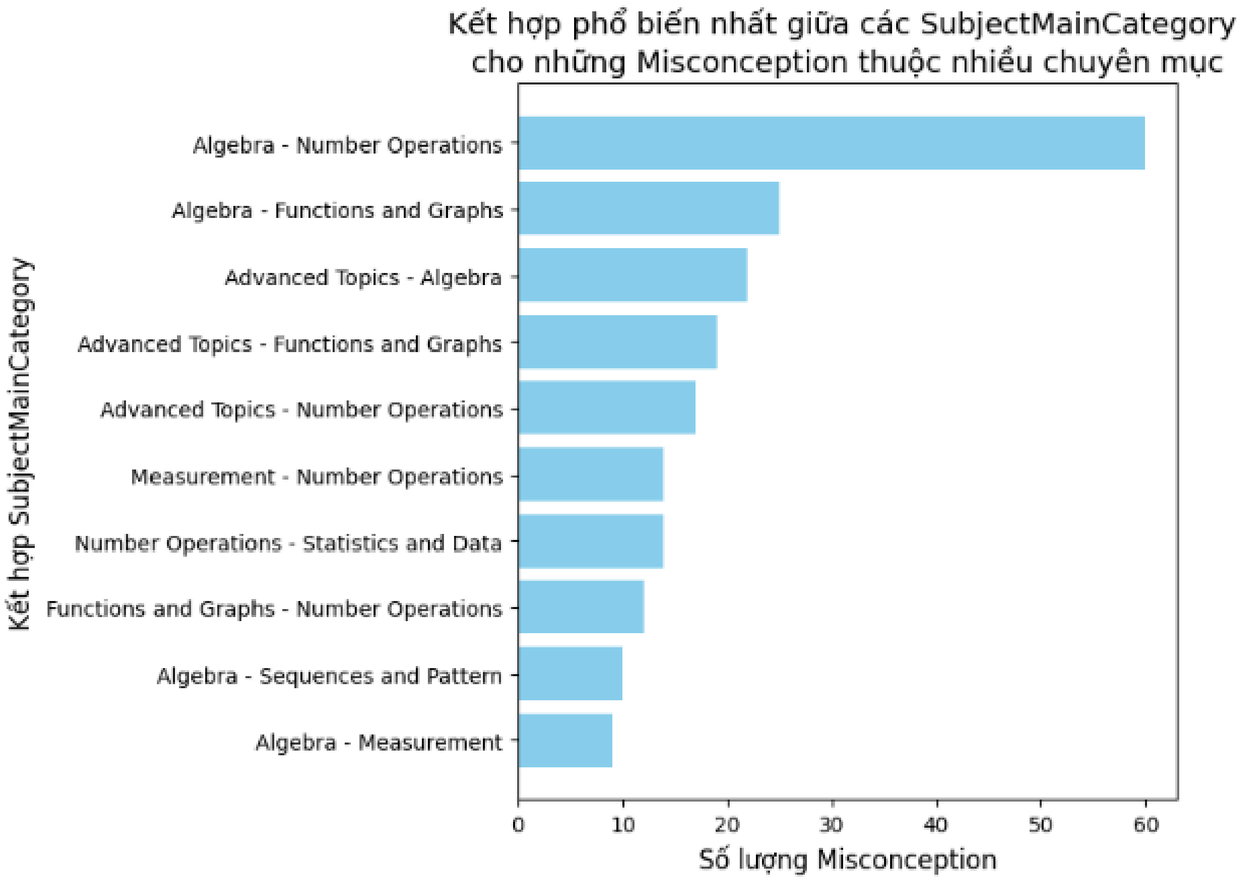


Question 16: For Misconceptions in 1 Subject Category, how many Misconceptions are there across Subject Sub Categories?

Subject Category Analysis

Most Misconceptions only appears in 1 Subject Sub Category

Second approach: Semantic Analysis with LLM clustering



Question 17: For Misconceptions in 2 Subject Categories, what are the most common combination of categories?

Subject Category Analysis

The most common combination is Algebra - Number Operation.

Which are also the most common Subject Types.

Construct Name Analysis

What is the structure of a Construct?

Calculate the **range** from a **list of data**

Action

Subject

Context

Construct Name Analysis

What is the structure of a Construct?

Calculate the range from a list of data

First 3-grams

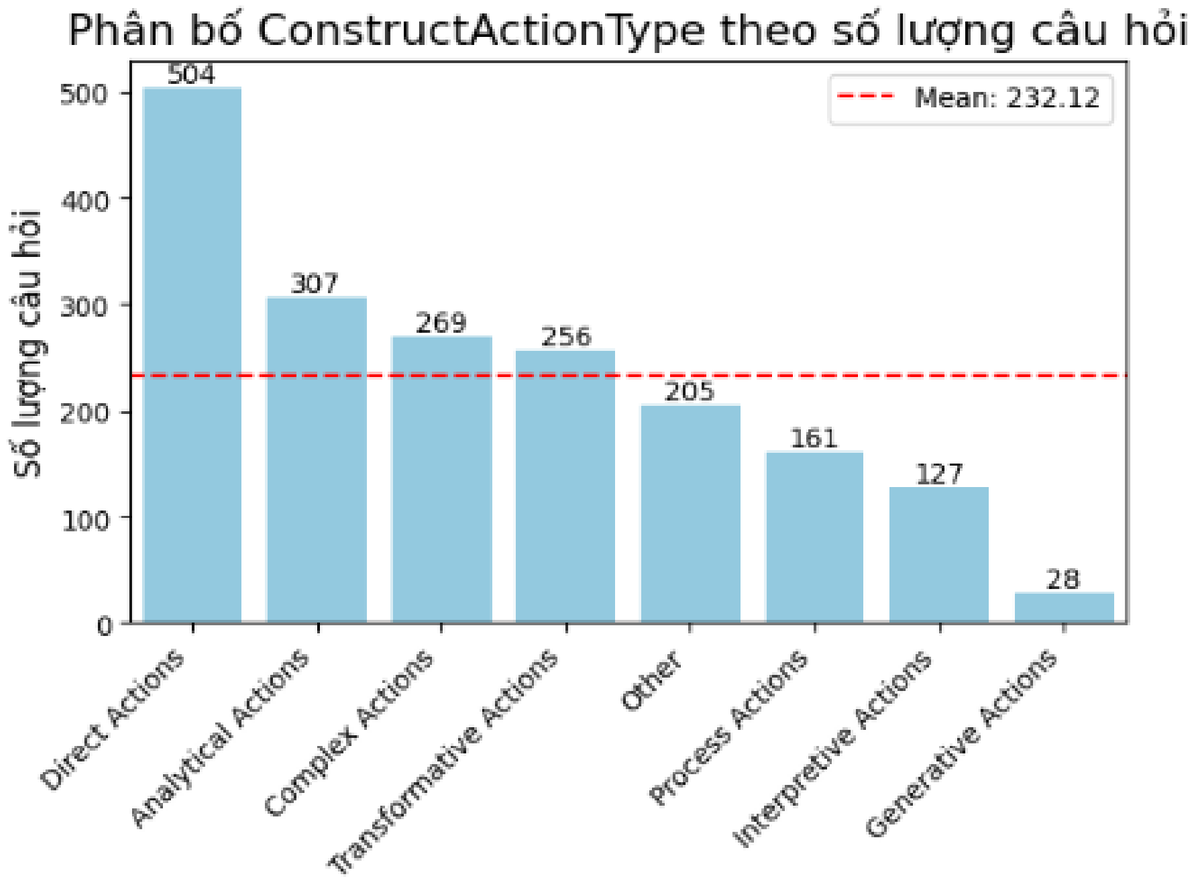
Construct Action Analysis

What are the Action Types in Constructs?

Do A + B Direct Actions	Prove A = B Complex Actions	Identify A Analytical Actions	Step a-> b-> c Process Actions
Make A -> B Transformative Actions	Describe A Interpretive Actions	Plot A Generative Actions	... Other

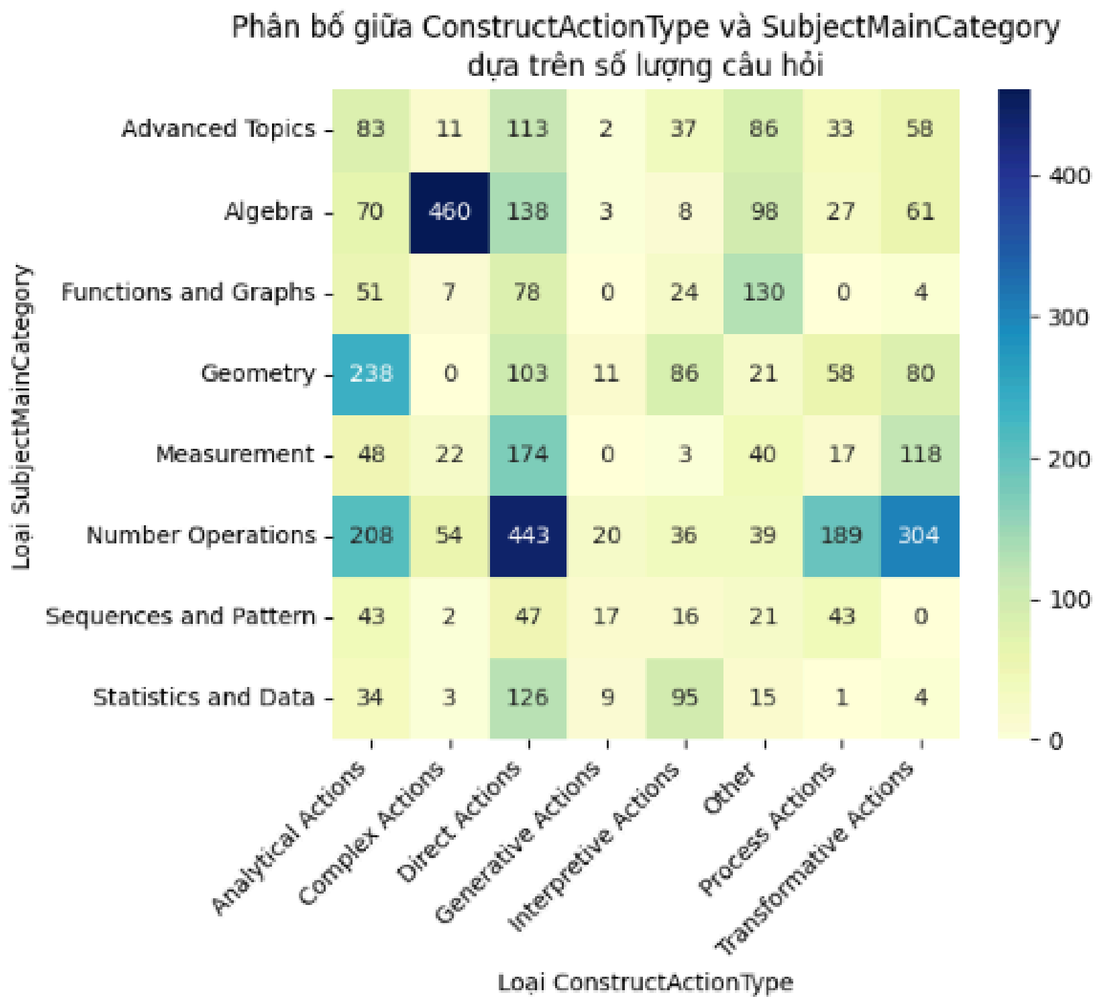
Question 18: How are Construct Action Type distributed?

Construct Action Analysis



The most common Constuct Action is Direct Action

Second approach: Semantic Analysis with LLM clustering



Question 19: What is the correlation between Construct Action Type and Subject Main Category based on number of Questions?

Construct Action Analysis

There are 3 types of Actions:
Popular, Focus and Unpopular

Misconception Analysis

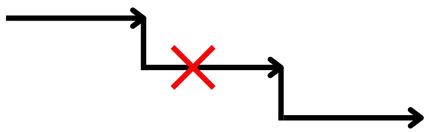
What is the structure of a Construct?

Believes if you changed all values by the same proportion the range would not change

First 3-grams

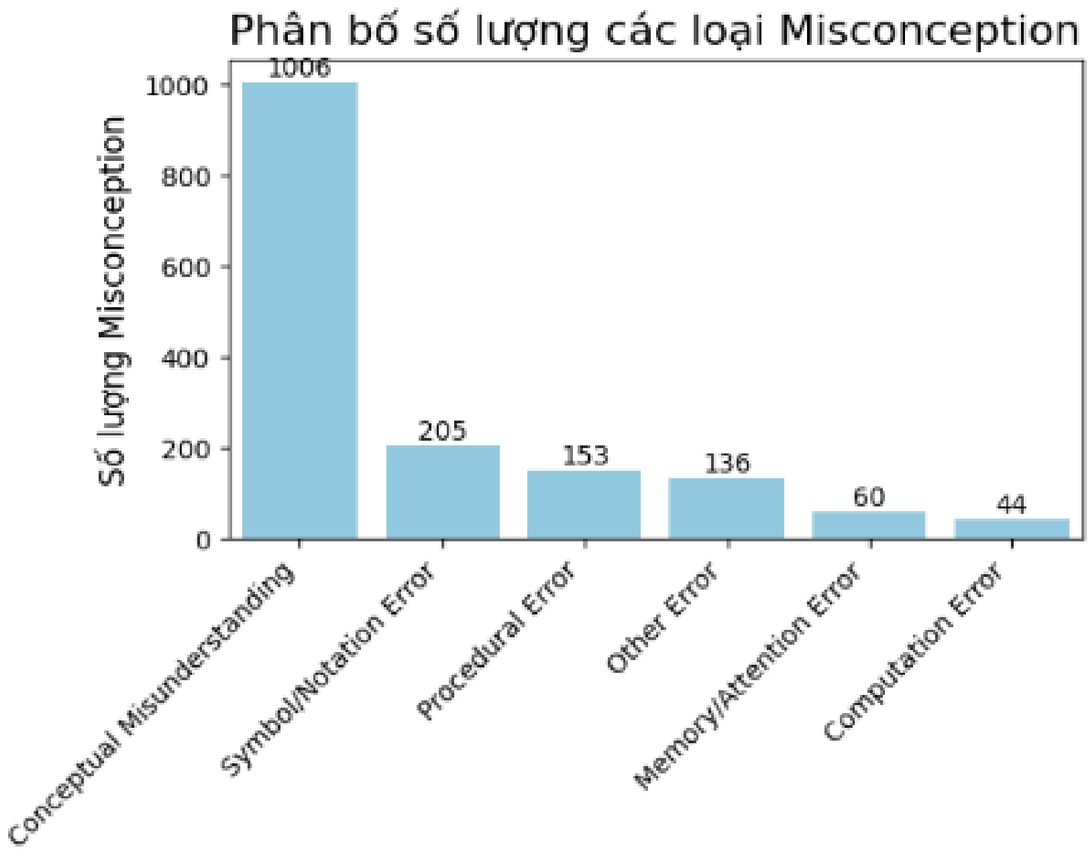
Misconception Analysis

What are the Misconception Types in Misconception mapping?

Conceptual Misunderstanding	 Procedual Error	12 / 3 = 5 Computation Error
a = 5 b = 2 -> a = 5 b = 7 Memory/Attention Error	octagon !=  Symbol/Notation Error	... Other Error

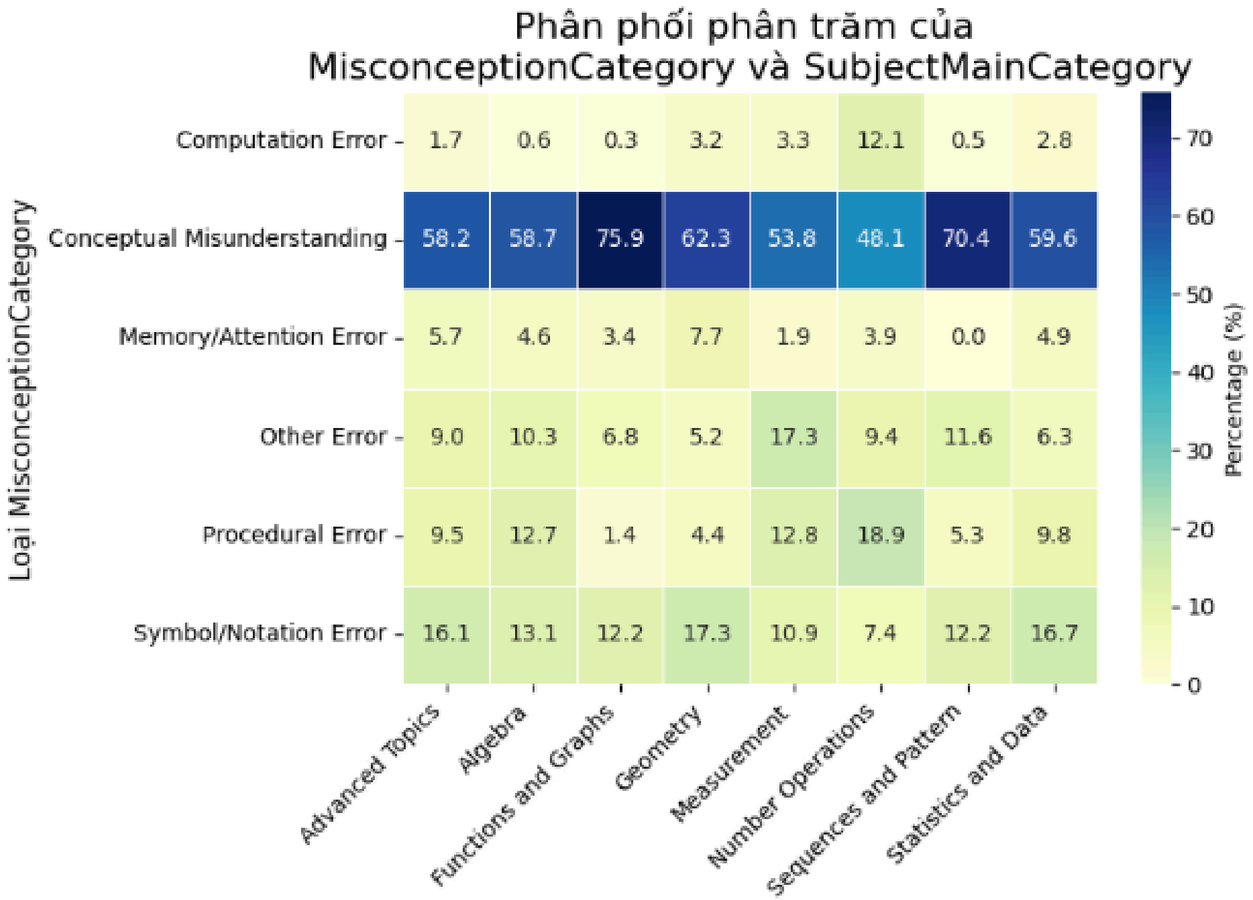
Question 20: How are Misconception Category distributed?

Misconception Analysis



The most common mistake is Conceptual Misunderstanding

Second approach: Semantic Analysis with LLM clustering

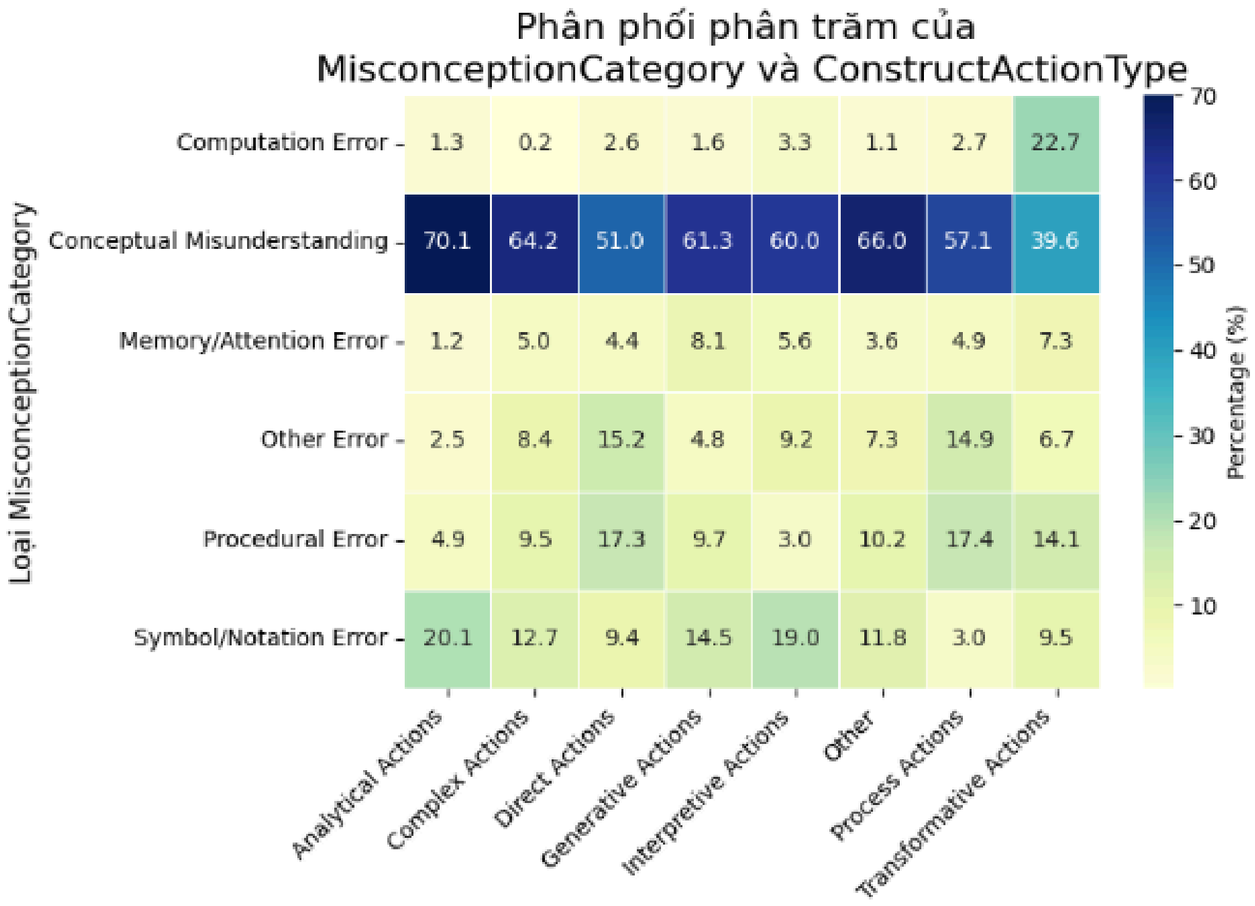


Question 19: What is the correlation between Misconception Category and Subject Main Category, expressed as a percentage, grouped by Subject Category?

Construct Action Analysis

Conceptual Misunderstanding is the most popular mistake for all Subject Category

Second approach: Semantic Analysis with LLM clustering



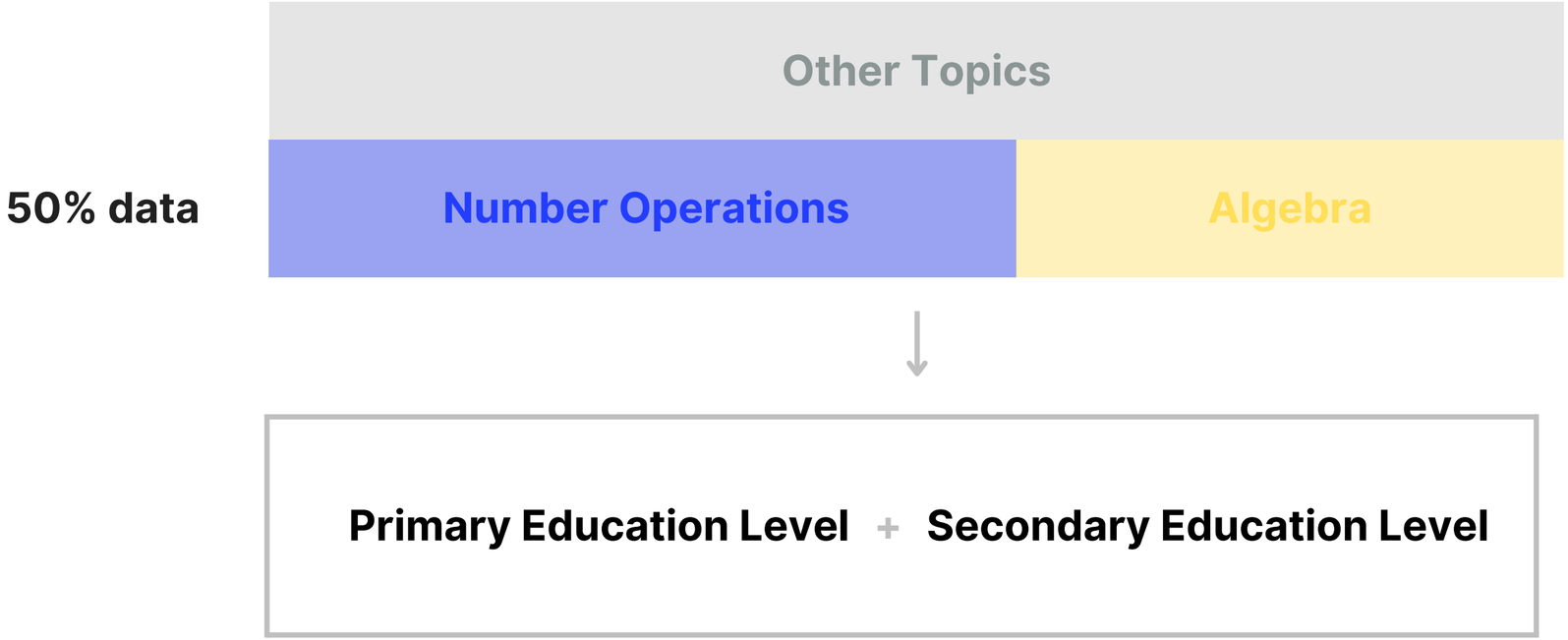
Question 20: What is the correlation between Misconception Category and Construct Action Type, expressed as a percentage, grouped by Construct Action?

Construct Action Analysis

Conceptual Misunderstanding is the most popular mistake for all Construct Action

Insights Summaries

What are the biggest insights we gained from our analysis?



Insights Summaries

What are the biggest insights we gained from our analysis?

Direct Actions →	Popular Construct Action	popular in most subject
Complex Action →	Focused Construct Action	focus primarily in Algebra
Generative Actions →	Unpopular Construct Action	rarely used in a subject

Insights Summaries

What are the biggest insights we gained from our analysis?

62.71% Mistakes comes from Conceptual Misunderstanding

↓

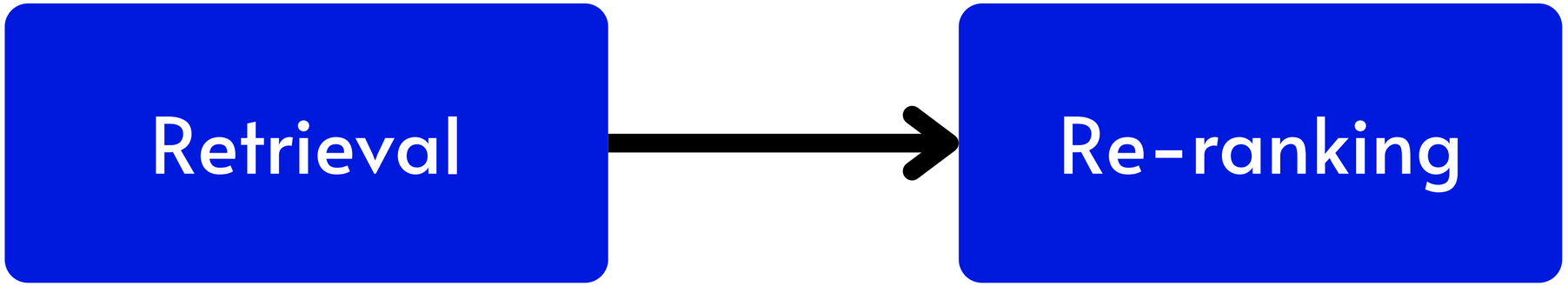
Problems with **learning process**.

03
Model Building

Information Retrieval

- Searching for and retrieving information (**documents**) relevant to a specific **query** from the user.
- In detailed:

Two-stages process for Information Retrieval:



Preprocessing

Test set

Question	Correct Answer	Correct Ans text	Answer_A	Answer_B	Answer_C	Answer_D
1	A		...	...	...	...
2	B		....	....	....	.....

Preprocessing

Test set

Question	Correct Answer	Correct Answer text	Answer	Answer_text
1	A		A	.....
1	A		B	....
1	A		C	....
1	A		D	....
2	B		....	

Preprocessing

Test set

Question	Correct Answer	Correct Answer text	Answer	Answer_text
1	A		B	....
1	A		C	....
1	A		D	....
2	B		....	

Query creation for a sample

Construct Name

Subject Name

Question Text

Correct Answer Text

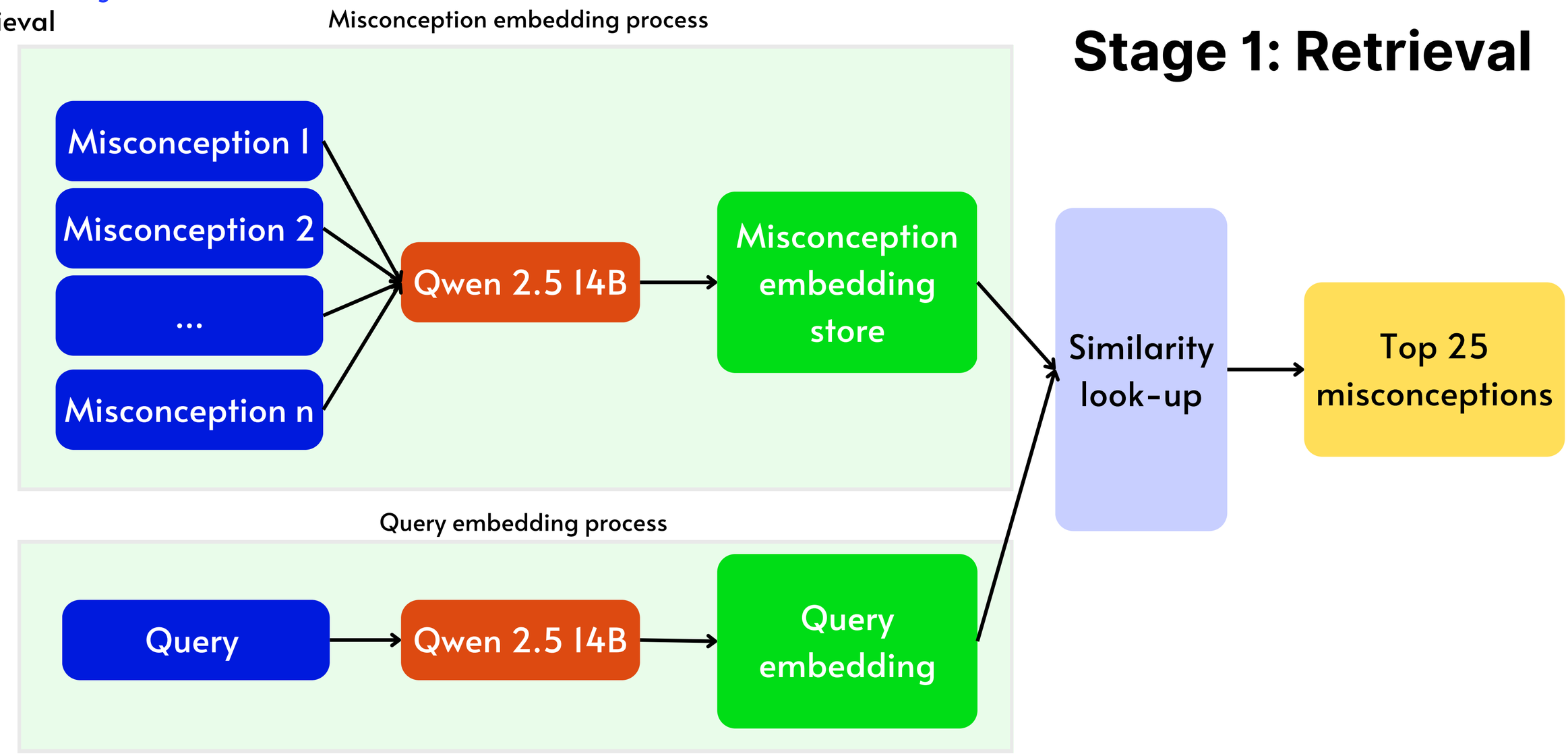
Wrong Answer Text

Template → Query

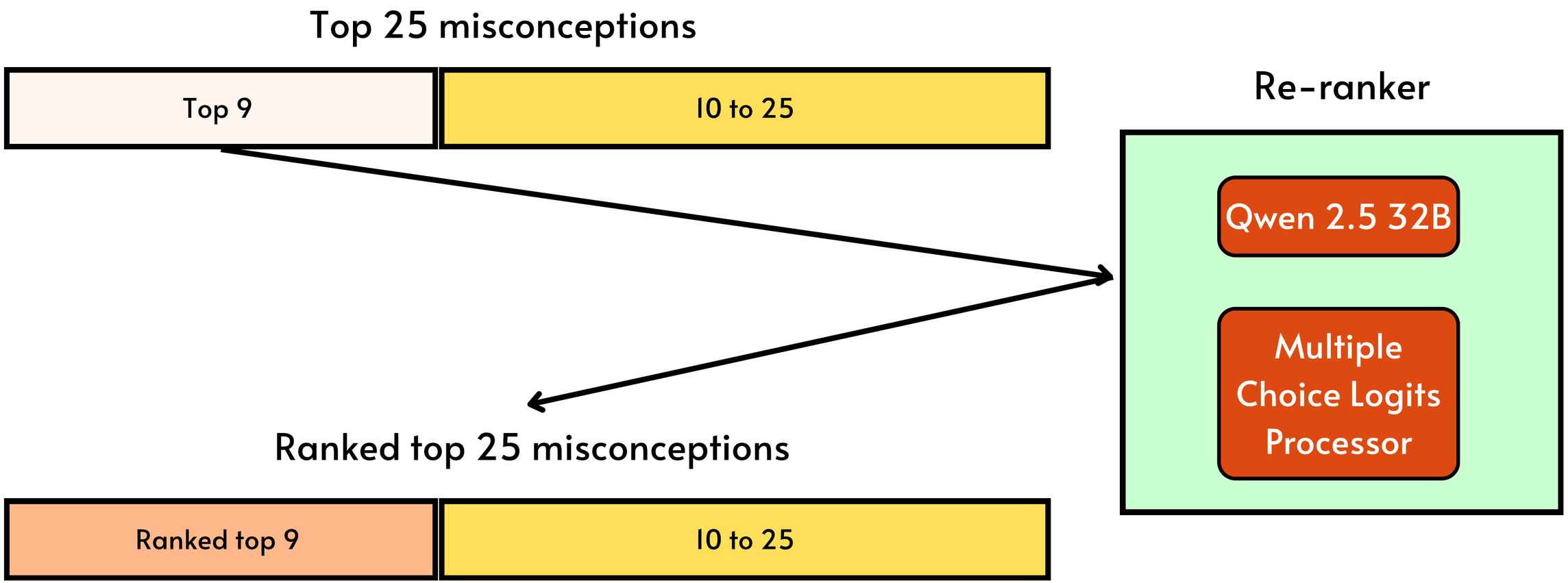
<instruct>Given a math multiple-choice problem with a student's wrong answer, retrieve the math misconceptions
<query>Question: {question}

SubjectName: {subject}
ConstructName: {construct}
Correct answer: {correct}
Student wrong answer: {wrong}
<response>

Stage 1: Retrieval



Stage 2: Re-ranking

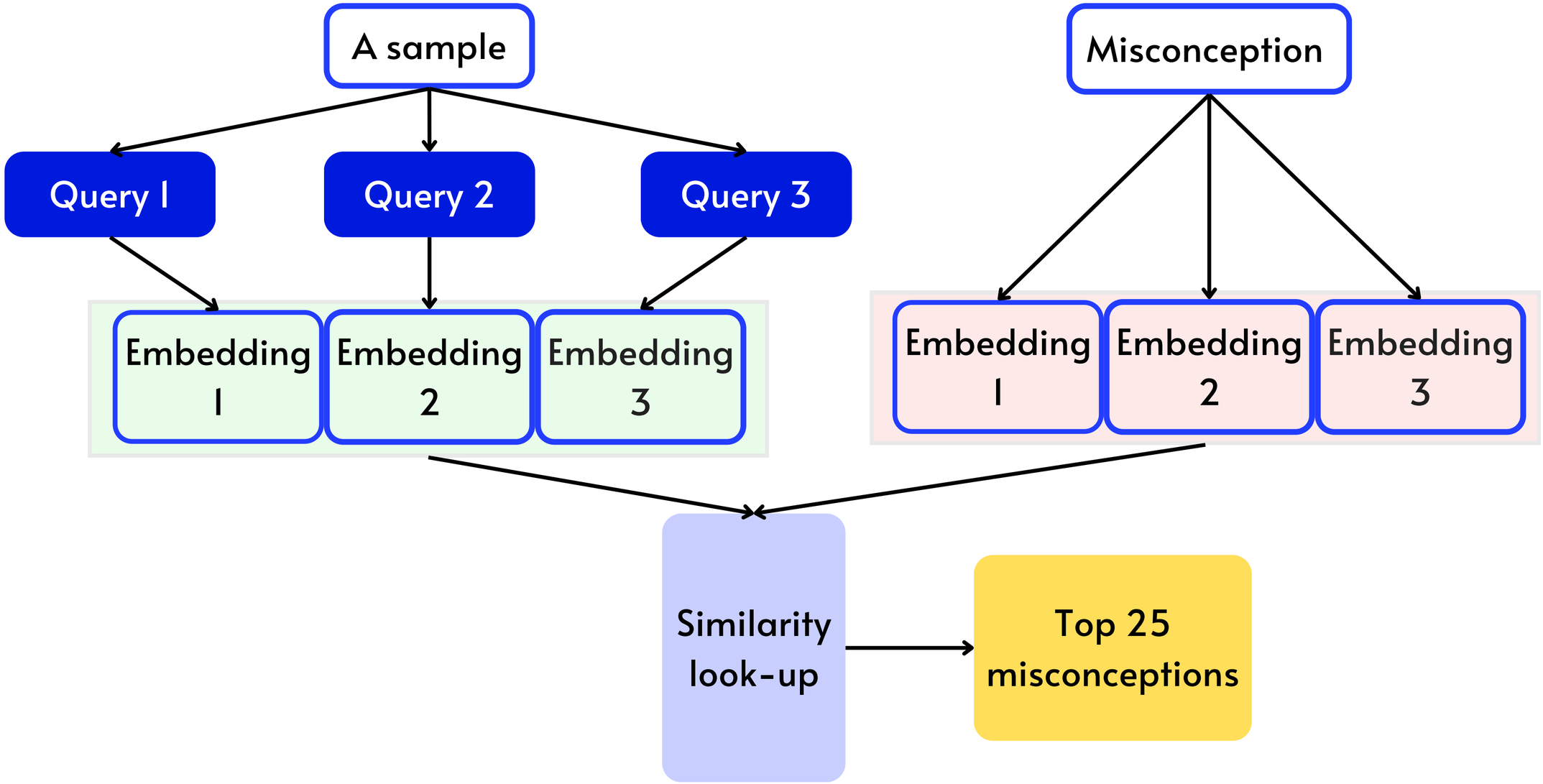


04
Experimentation

1.Approaches

Set up	Public score	Private Score
TF-IDF vectorizor + LGBM classifier	0.00713	0.00518
TF-IDF + Language model embeddings + Cosine SImilarrity	0.18383	0.17065
Qwen 2.5 14B retreival	0.45712	0.43415
Qwen 2.5 14B retreival + Qwen 2.5 32B re-rank	0.48270	0.44615

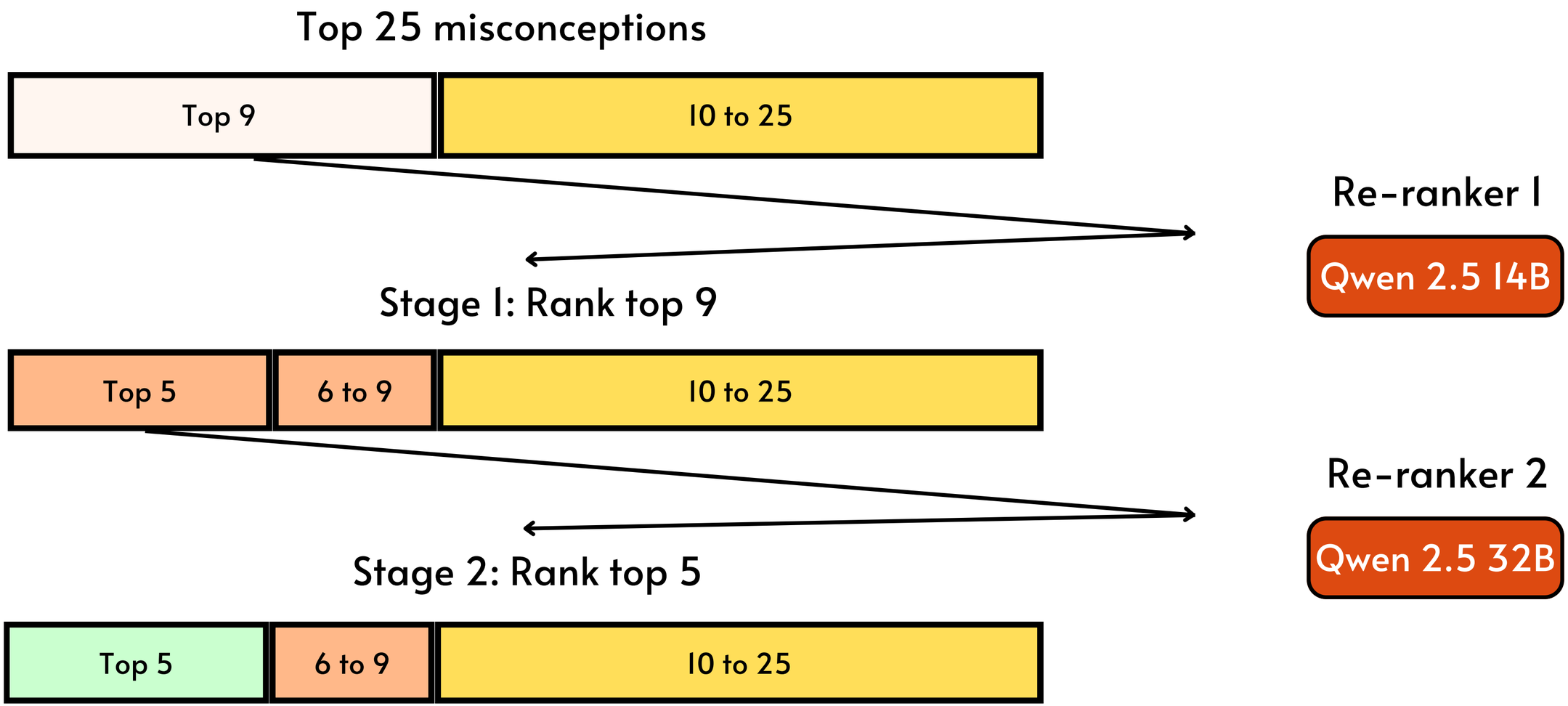
2. Ensemble embeddings



2. Ensemble embeddings

	Public score	Private Score
Before enssemble	0.49952	0.44615
After ensemble	0.5298	0.48434

2. Ensemble re-rankers



2. Ensemble re-rankers

	Public score	Private Score
Before enssemble	0.49952	0.44615
After ensemble	0.48119	0.44669

3.Data Augmentation

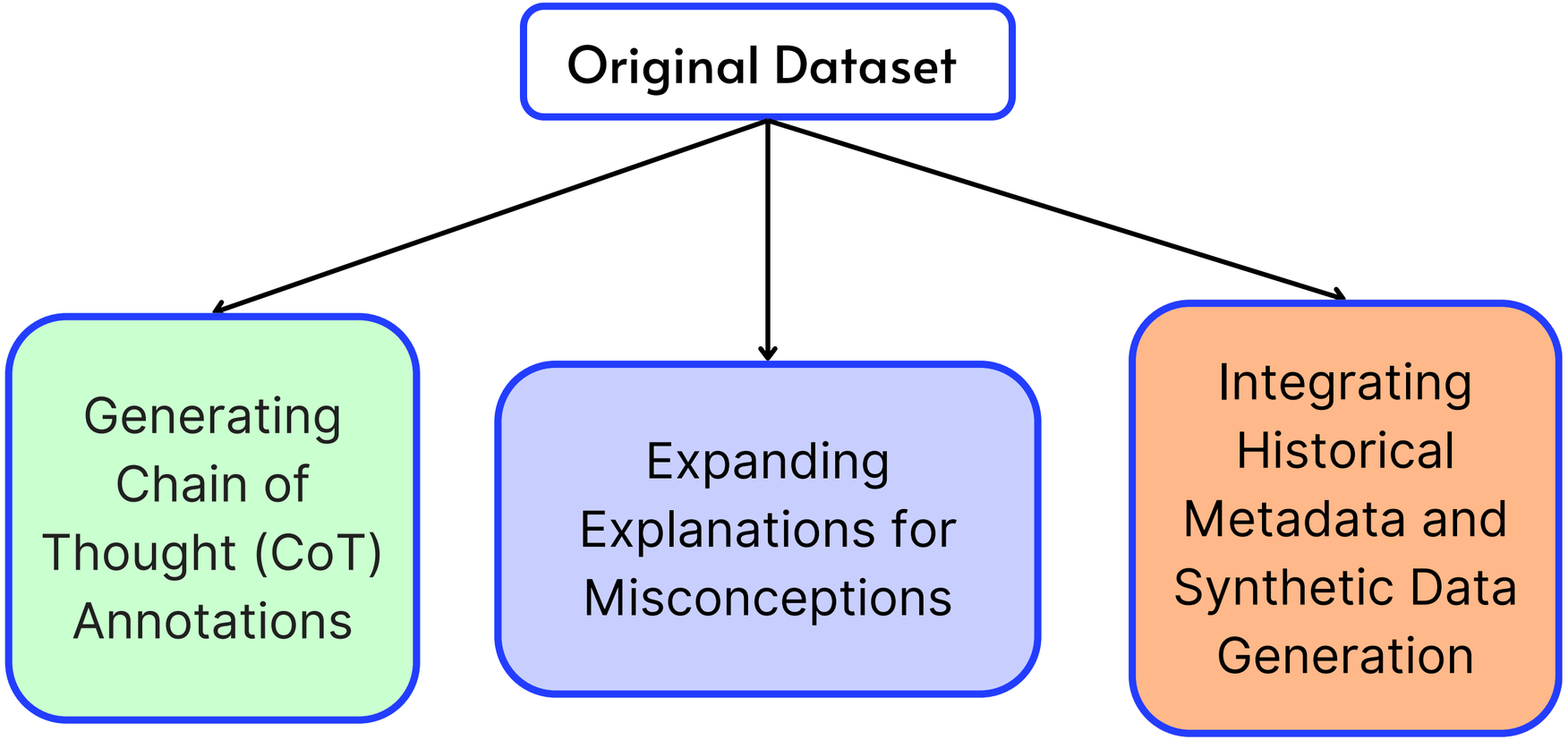
Objectives

- Resolve data imbalance in training datasets
- Improve recognition of rare misconceptions

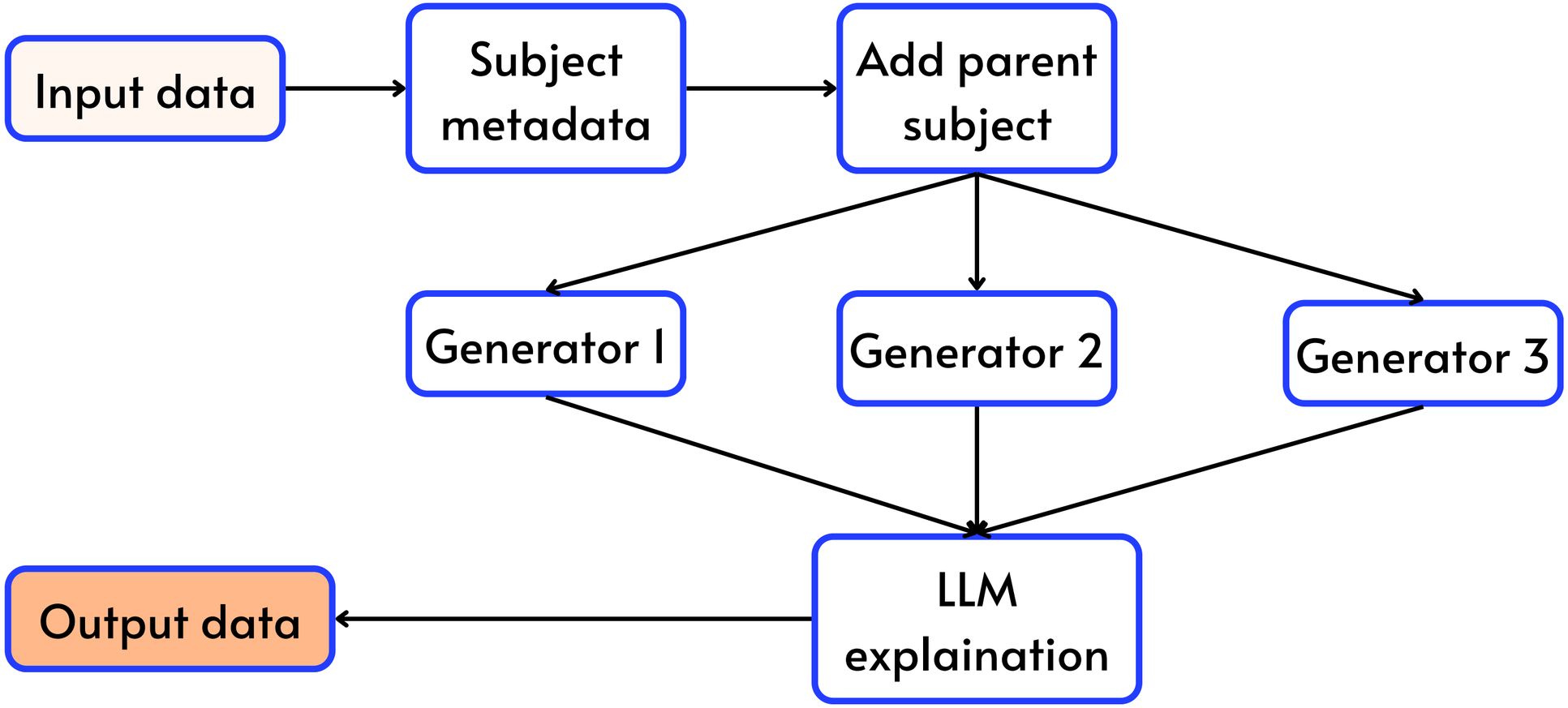
↓

Generate data to **fine-tune model**, enhancing accuracy and generalization

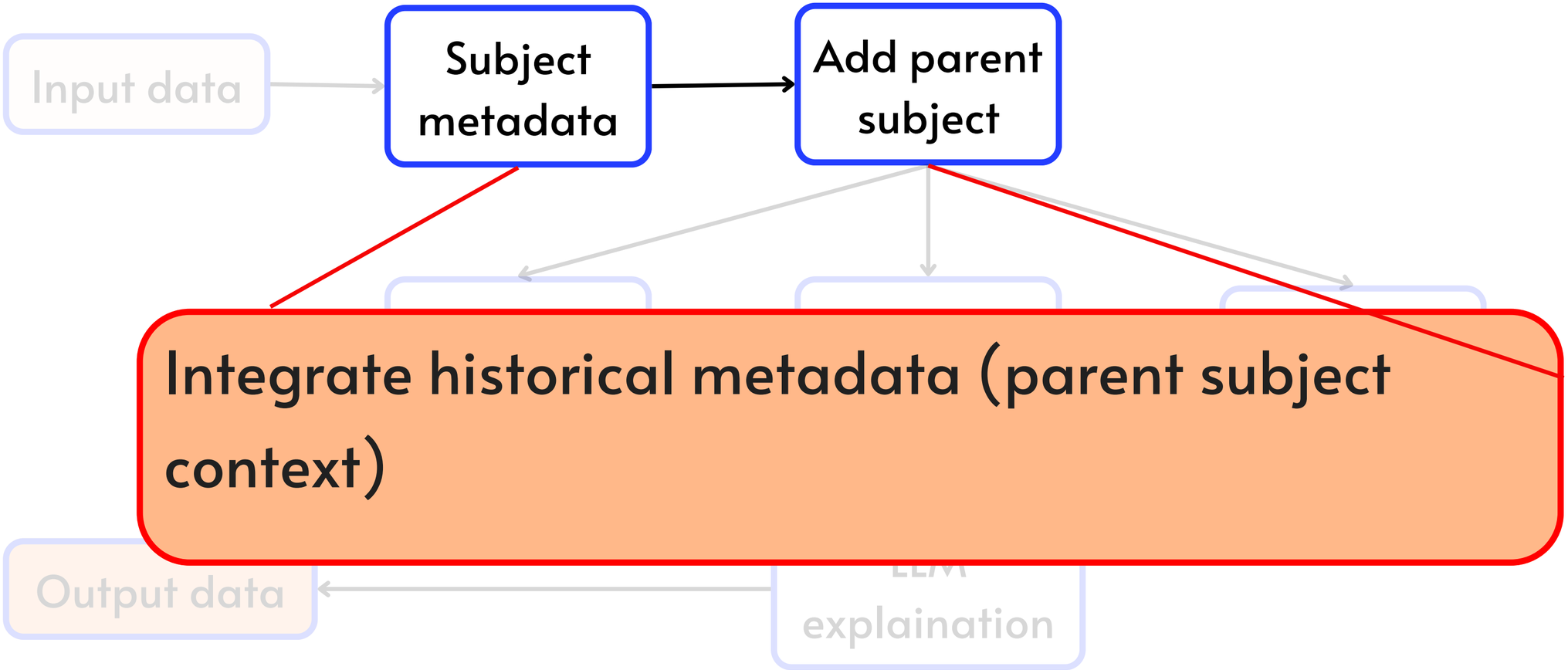
Approaches


Original Dataset
Generating Chain of Thought (CoT) Annotations
Expanding Explanations for Misconceptions
Integrating Historical Metadata and Synthetic Data Generation

Method 3: Leveraging Metadata and Generating Synthetic Data

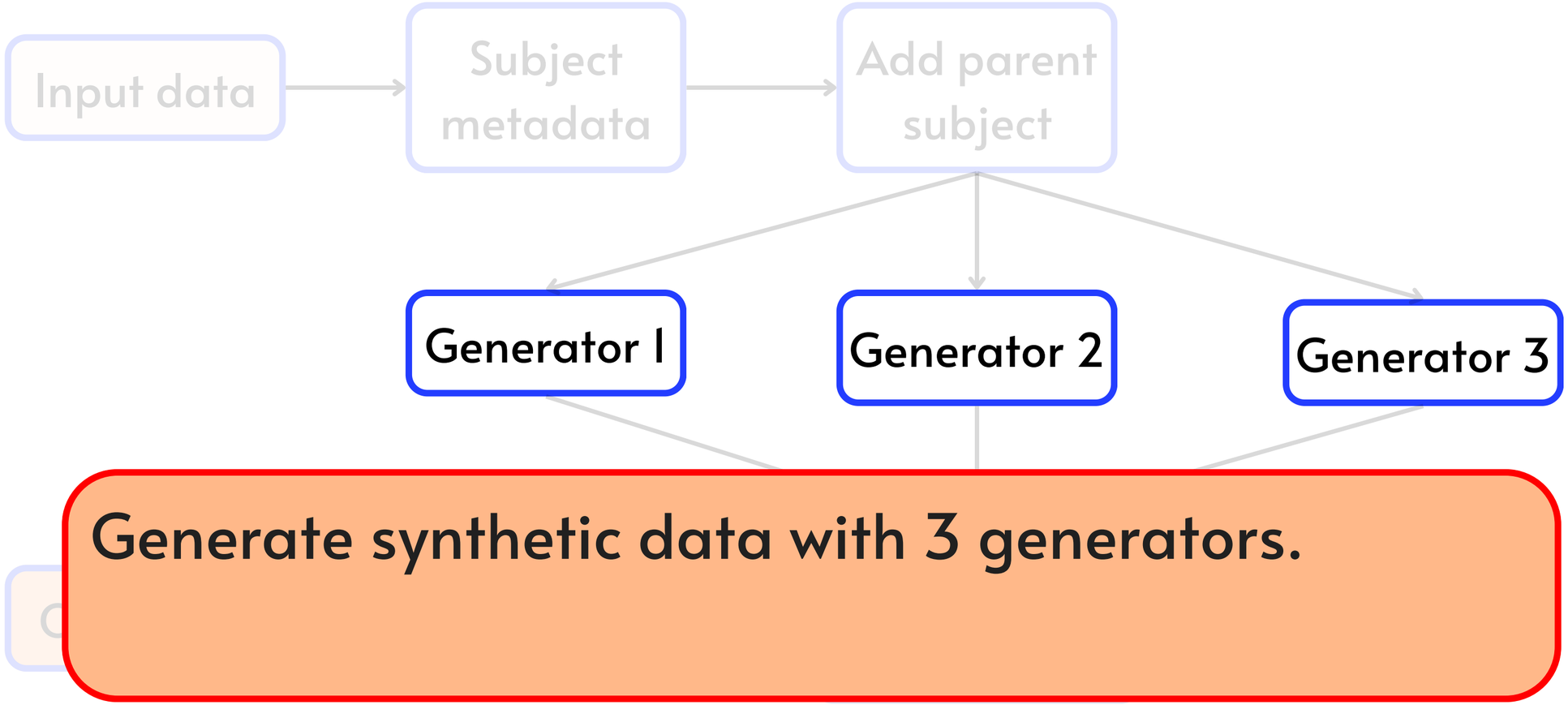


Method 3: Leveraging Metadata and Generating Synthetic Data

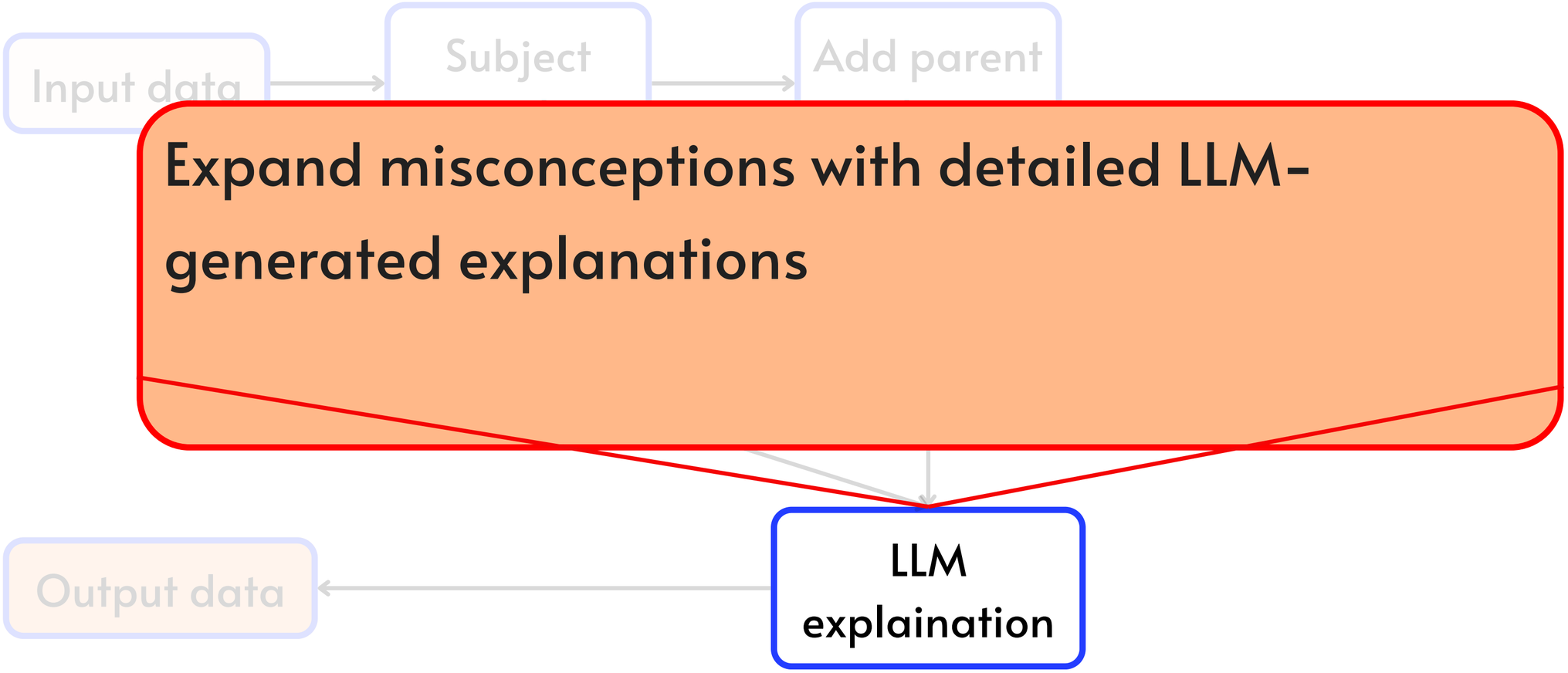


Integrate historical metadata (parent subject context)

Method 3: Leveraging Metadata and Generating Synthetic Data



Method 3: Leveraging Metadata and Generating Synthetic Data



Method 3: Leveraging Metadata and Generating Synthetic Data

The athor: ***"Using misconception augmentation significantly boosted retriever performance by about 2-4%."***

However:

- Each step takes over 5 hours.
- The overall generation process takes approximately 50 hours.

➡ Fail to generate the whole data using Kaggle environment

05
Conclusion

1. Leaderboard score: Public: 0.49952 Private: 0.44615

2. Reflection:

Strength	Weakness
• The embeddings of LLM represent the semantics of text more effectively. • Two-stage retrieval create more accurate prediction.	• The query speed is still limited and ineffective for large datasets. • The correctness of the prediction is also limited

3. Future works
   - Experiment with various data augmentation methods.
   - Fine-tune the language model on both existing and synthetic datasets to improve inference accuracy.
   - Find ways to optimize memory usage on Kaggle and experiment with models with more parameters.
   - Test more effective prompting strategies, such as Chain-of-Thought.

Thank you for listening!