

ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN



BÁO CÁO

Đề án: Eedi - Mining Misconceptions in Mathematics

MÔN: HỌC SÂU CHO KHOA HỌC DỮ LIỆU

Lớp: 21_21

Giảng viên hướng dẫn: TS. Nguyễn Tiến Huy
TS. Lê Thanh Tùng
ThS. Nguyễn Trần Duy Minh

Nhóm thực hiện: Nhóm 1

Thông tin thành viên:

21120058 Phạm Nhật Duy
21120102 Nguyễn Trúc Nguyên
21120158 Trương Công Trung
21120279 Lê Trần Minh Khuê

THÀNH PHỐ HỒ CHÍ MINH, THÁNG 1/ 2025

Lời cảm ơn

Chúng em xin chân thành cảm ơn TS. Nguyễn Tiến Huy, TS. Lê Thanh Tùng, ThS. Nguyễn Trần Duy Minh, những giảng viên đã dày công truyền đạt kiến thức và hướng dẫn chúng em trong quá trình học và thực hiện đồ án này.

Chúng em xin gửi lời cảm ơn đến khoa Công nghệ thông tin, Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia TP HCM đã tạo điều kiện cho chúng em học tập và hoàn thành đồ án này.

Chúng em đã cố gắng vận dụng những kiến thức đã học được để hoàn thành đồ án. Nhưng do kiến thức hạn chế và chưa có nhiều kinh nghiệm nên khó tránh khỏi những thiếu sót trong quá trình nghiên cứu và trình bày.

Rất kính mong sự góp ý của quý Thầy để bài báo cáo của chúng em được hoàn thiện hơn.

Chúng em trân trọng cảm ơn sự quan tâm giúp đỡ của các Thầy trong suốt môn học này.

Xin trân trọng cảm ơn!

TP. Hồ Chí Minh, ngày 10 tháng 01 năm 2025

NHÓM SINH VIÊN THỰC HIỆN

Đại diện

Phạm Nhật Duy

Mục lục

1	Tổng quan	5
1.1	Tóm tắt đề bài	5
1.1.1	Giới thiệu bài toán	5
1.1.2	Yêu cầu thực hiện	5
1.2	Đánh giá mức độ hoàn thành	6
1.3	Ghi chú	6
1.3.1	Môi trường thực nghiệm	6
1.3.2	Tổ chức bài làm	6
2	Nội dung báo cáo	7
2.1	Khám phá dữ liệu	7
2.1.1	Tổng quan dữ liệu	7
2.1.2	Kiểm tra giá trị và phân bố dữ liệu	8
2.2	Tiền xử lý dữ liệu	15
2.2.1	Các ký tự dư thừa	15
2.2.2	Chuẩn hoá dữ liệu	16
2.3	Xây dựng mô hình	16
2.3.1	Mô hình cơ sở	16
2.3.2	Tinh chỉnh mô hình	17
2.4	Triển khai và đánh giá mô hình	17
2.4.1	Giai đoạn 1: Truy xuất	18
2.4.2	Giai đoạn 2: Xếp hạng	19
2.4.3	Đánh giá	22
	Tài liệu tham khảo	23

Danh mục bảng

1	Bảng phân công công việc	6
2	Bảng mô tả các thuộc tính của câu hỏi	7
3	Xem xét giá trị của từng cột	9
4	Danh sách tần suất xuất hiện của từng ConstructName	10
5	Danh sách tần suất xuất hiện của từng SubjectName	11
6	Danh sách các ngộ nhận phổ biến	13

Danh mục hình

1	Số dòng null trong các cột của dataframe	8
2	Xem xét số lượng câu trả lời đúng của các câu trả lời 'không có ngộ nhận'	8
3	Danh sách tần suất xuất hiện của từng CorrectAnswer	12
4	Danh sách tần suất xuất hiện của từng CorrectAnswer	14
5	Danh sách tần suất xuất hiện của từng CorrectAnswer	15
6	Flow chart minh họa quy trình dự đoán ngộ nhận toán học	18
7	Kết quả nhóm đạt được	22

Danh mục từ viết tắt

- AWQ **A**ctivated **W**eight **Q**uantization, một phương pháp lượng tử hóa trọng số giúp tối ưu hóa hiệu năng của các mô hình học sâu trong khi vẫn duy trì độ chính xác cao..
- LoRA **L**ow-**R**ank **A**daptation, một phương pháp tinh chỉnh mô hình học sâu lớn bằng cách chỉ cập nhật các ma trận hạng thấp, giúp tiết kiệm tài nguyên tính toán..
- SFT **S**upervised **F**ine-**T**uning, phương pháp tinh chỉnh mô hình ngôn ngữ lớn bằng cách sử dụng dữ liệu được gán nhãn. SFT giúp cải thiện khả năng của mô hình trong việc thực hiện các tác vụ cụ thể bằng cách học từ các ví dụ đầu vào-đầu ra được giám sát..
- trl **T**ransformer **R**einforcement **L**earning, một thư viện được thiết kế để áp dụng các phương pháp học tăng cường vào huấn luyện các mô hình Transformer..
- vLLM **v**irtualized **L**arge **L**anguage **M**odel, một kiến trúc tối ưu cho việc triển khai các mô hình ngôn ngữ lớn, cải thiện khả năng mở rộng và tốc độ xử lý thông qua ảo hóa tài nguyên..

1 Tổng quan

1.1 Tóm tắt đề bài

1.1.1 Giới thiệu bài toán

Cuộc thi từ Eidi đặt ra một thách thức hấp dẫn: phát triển một mô hình học máy xử lý ngôn ngữ tự nhiên có khả năng phân tích và dự đoán mối liên hệ giữa các quan niệm sai lầm (ngộ nhận) và các yếu tố gây nhiễu (những câu trả lời không chính xác) trong các câu hỏi trắc nghiệm.

Đặc biệt, trong các câu hỏi trắc nghiệm toán học, những lựa chọn không chính xác thường được thiết kế để đánh lừa người trả lời, phản ánh các sai lầm phổ biến trong tư duy hoặc tính toán. Ví dụ, một người trả lời tính toán sai nhưng lại tìm thấy chính kết quả sai đó trong các phương án, điều này gây ra sự nhầm lẫn. Mô hình cần phải nhận biết được những sai sót này và xác định ngộ nhận đã dẫn đến chúng.

Giải pháp không chỉ dừng lại ở việc nhận diện các ngộ nhận về kiến thức toán học hiện có mà còn phải có khả năng khái quát hóa những sai lầm mới. Điều này mở ra tiềm năng hỗ trợ giáo viên trong việc nhanh chóng nhận diện và giải quyết các ngộ nhận phổ biến, từ đó cải thiện chất lượng giáo dục và tăng cường hiệu quả học tập cho học sinh.

1.1.2 Yêu cầu thực hiện

- Phân tích bộ dữ liệu để nhận diện các đặc trưng quan trọng, bao gồm mối quan hệ giữa câu hỏi, câu trả lời đúng, và các đáp án gây nhiễu (distractors), đồng thời xác định các mẫu quan niệm sai lầm thường gặp.
- Phát triển một mô hình học máy có khả năng dự đoán chính xác mối liên kết giữa các quan niệm sai lầm và các đáp án không chính xác. Mô hình cần tập trung vào việc nhận diện các đặc trưng gây nhầm lẫn và khái quát hóa các quan niệm sai lầm mới từ dữ liệu.
- Đánh giá hiệu suất của mô hình trên các chỉ số thích hợp, như độ chính xác và khả năng tổng quát hóa, nhằm đảm bảo rằng mô hình có thể áp dụng hiệu quả trên các tập dữ liệu khác nhau.
- Nghiên cứu và tối ưu hóa mô hình để đạt được kết quả cao trên bộ dữ liệu đã cung cấp, đồng thời phát hiện những điểm cải tiến tiềm năng cho ứng dụng thực tế.

1.2 Đánh giá mức độ hoàn thành

STT	Nội dung	Thành viên thực hiện	Hoàn thành
1	Tìm hiểu các mô hình ngôn ngữ lớn và nghiên cứu các phương pháp phù hợp cho bài toán	Cả nhóm	100%
2	Phân tích và tiền xử lý dữ liệu để chuẩn bị đầu vào cho mô hình	Minh Khuê	100%
3	Xây dựng mô hình cơ sở để thử nghiệm ban đầu với bộ dữ liệu	Công Trung	100%
4	Tinh chỉnh mô hình dựa trên các kỹ thuật cải tiến và tối ưu hóa hiệu suất	Nhật Duy	100%
5	Chuẩn bị tài liệu trình bày bao gồm nội dung slide và nội dung thuyết trình	Trúc Nguyên	100%
6	Trình bày báo cáo kết quả và mô hình	Trúc Nguyên	100%
7	Viết báo cáo chi tiết về phương pháp, thực nghiệm và kết quả đạt được	Nhật Duy	100%

Bảng 1: Bảng phân công công việc

1.3 Ghi chú

1.3.1 Môi trường thực nghiệm

Các thuật toán được chạy và đo các số liệu trên nền tảng Kaggle, một môi trường trực tuyến phổ biến cho việc thực nghiệm và triển khai các mô hình học máy. Nền tảng này cung cấp tài nguyên phần cứng và phần mềm như sau:

- CPU: 2.3 GHz Intel Xeon Processor
- GPU: NVIDIA T4
- RAM: 13 GB
- Lưu trữ: 20 GB Disk Space
- Hệ điều hành: Ubuntu 20.04 (Linux-based)

Chương trình được thực thi trong môi trường Jupyter Notebook tích hợp sẵn trên Kaggle, hỗ trợ Python và các thư viện liên quan như TensorFlow và Scikit-learn.

1.3.2 Tổ chức bài làm

Nhóm bố trí bài làm trong folder "21120058_21120102_21120158_21120279" như sau:

- Source

- 1. data discovery.ipynb: File notebook tìm hiểu tổng quan, phân phối và phân tích dữ liệu.
 - 2. data preprocessing.ipynb: File notebook tiền xử lý cơ bản dữ liệu.
 - 3. train qwen2.5-14b.ipynb: File notebook tinh chỉnh mô hình qwen2.5-14b dựa trên dữ liệu.
 - 4. experimental pipeline.ipynb: File notebook triển khai mô hình để phân tích và xác định kết quả.
- Report.pdf: File trình bày quá trình thực hiện và kết quả đồ án. (file này)
 - Slide.pdf: File trình bày nội dung báo cáo, dùng để minh họa các ý chính trong video thuyết trình.
 - Video.txt: File chứa link Youtube video trình bày nội dung đồ án.

2 Nội dung báo cáo

2.1 Khám phá dữ liệu

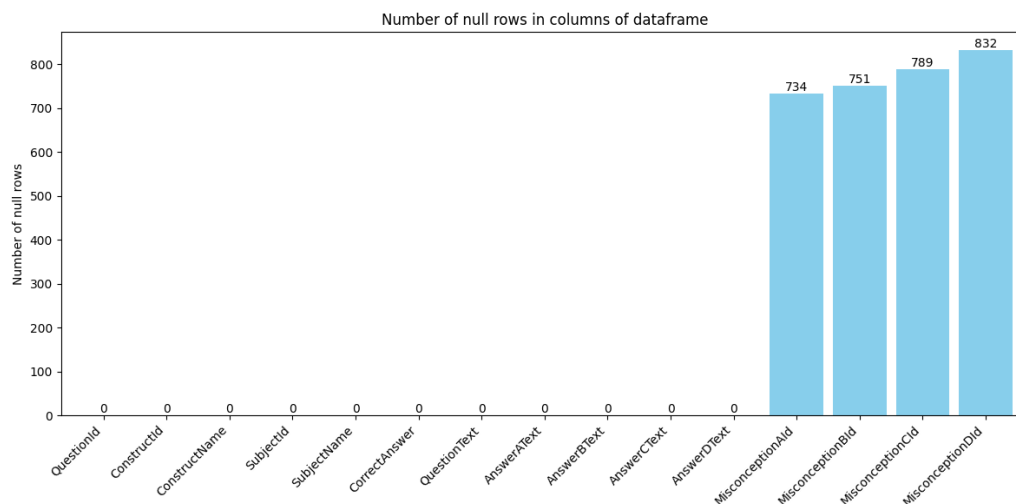
2.1.1 Tổng quan dữ liệu

Bộ dữ liệu hiện có gồm 1869 dòng, với tổng số 15 cột chứa thông tin về các câu hỏi trắc nghiệm và khái niệm liên quan đến câu hỏi đó, cùng với các câu trả lời. Tất cả dữ liệu được viết bằng tiếng Anh. Cụ thể từng cột như sau:

STT	Thuộc tính	Mô tả
1	QuestionId	Mã câu hỏi (int)
2	ConstructId	Mã khái niệm (int)
3	ConstructName	Khái niệm - Mức độ kiến thức chi tiết nhất liên quan tới câu hỏi (str)
4	CorrectAnswer	Đáp án đúng (A, B, C, D) (char)
5	SubjectId	Mã môn học (int)
6	SubjectName	Tên môn học (str)
7	QuestionText	Câu hỏi dạng văn bản (str)
8	AnswerAText	Đáp án A dạng văn bản (str)
9	AnswerBText	Đáp án B dạng văn bản (str)
10	AnswerCText	Đáp án C dạng văn bản (str)
11	AnswerDText	Đáp án D dạng văn bản (str)
12	MisconceptionAId	Mã ngộ nhận tương ứng với đáp án A - nếu có (int)
13	MisconceptionBId	Mã ngộ nhận tương ứng với đáp án B - nếu có (int)
14	MisconceptionCId	Mã ngộ nhận tương ứng với đáp án C - nếu có (int)
15	MisconceptionDId	Mã ngộ nhận tương ứng với đáp án D - nếu có (int)

Bảng 2: Bảng mô tả các thuộc tính của câu hỏi

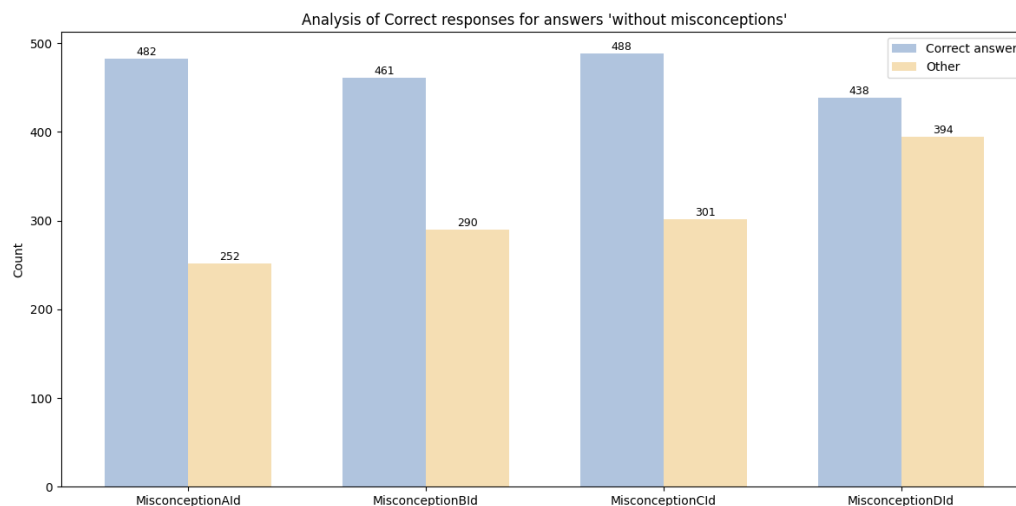
2.1.2 Kiểm tra giá trị và phân bố dữ liệu



Hình 1: Số dòng null trong các cột của dataframe

Ta nhận thấy rằng chỉ có các cột Misconception bị NaN → Các giá trị null này có thể là không có ngộ nhận, hoặc đây là đáp án đúng, hoặc là lỗi cơ bản, không liên quan đến bất kỳ ngộ nhận hay hiểu lầm nào về khái niệm.

Tiếp theo, tiến hành kiểm tra xem bao nhiêu **misconception** NaN là đáp án đúng, và phần còn lại là bao nhiêu.



Hình 2: Xem xét số lượng câu trả lời đúng của các câu trả lời 'không có ngộ nhận'

Sau khi xem xét, nhóm quyết định điền -1 cho nhóm các dữ liệu NaN. Nhóm sẽ quy ước rằng -1 là “id” của các câu trả lời đúng/ không có ngộ nhận.

Tiếp theo, chúng ta sẽ tiến hành phân tích chi tiết từng cột trong dữ liệu, nhằm hiểu rõ hơn về cấu trúc và các đặc điểm của các biến. Phân tích này sẽ giúp làm sáng tỏ mỗi

quan hệ giữa các yếu tố và hỗ trợ trong việc đưa ra những kết luận chính xác hơn về dữ liệu.

Chỉ số	Số giá trị phân biệt	Giá trị xuất hiện nhiều nhất	Số lần xuất hiện nhiều nhất
ConstructName	757	Calculate the square of a number	14
SubjectName	163	Linear Equations	53
CorrectAnswer	4	C	488
QuestionText	1857	Which of the following pairs of function machines are correct?	4
AnswerAText	1219	Only Tom	93
AnswerBText	1230	Only Katie	109
AnswerCText	1222	Both Tom and Katie	158
AnswerDText	1184	Neither is correct	187

Bảng 3: Xem xét giá trị của từng cột

Nhận thấy rằng không có cột nào chỉ gồm một giá trị. Các khái niệm nền tảng đa dạng với 757 khái niệm riêng biệt. Ngoài ra, có tới 163 môn học/ phân môn khác nhau. Số lượng câu hỏi và câu trả lời rất đa dạng; tuy nhiên, có một số câu trả lời xuất hiện với tần suất khá cao, cần được xem xét kỹ lưỡng hơn.

Cột ConstructName

Index	ConstructName	Frequency
0	Calculate the square of a number	14
1	Solve two-step linear equations, with the variable on one side, with all positive integers	13
2	Factorise a quadratic expression in the form $x^2 + bx + c$	13
3	Use the order of operations to carry out calculations involving addition, subtraction, multiplication, and/or division	12
4	Identify the order of rotational symmetry of a shape	12
5	Multiply a single term over a bracket where the term on the outside is a number and the inside contains a linear expression	11
6	Calculate the range from a list of data	11
7	Solve one-step linear inequalities in one variable where the variable appears on one side of the equation	9
8	Recognise cube numbers	9
9	Solve quadratic equations using the quadratic formula where the coefficient of x^2 is not 1	9
10	Interpret a bar chart	9
11	Use a linear sequence expressed as a pattern to make a prediction about another term in the sequence other than the next one	9
12	Interpret continuous data using a line graph	9
13	Simplify an algebraic fraction by factorising both the numerator and denominator	9
14	Simplify algebraic expressions to maintain equivalence by collecting like terms involving just one linear variable	8
15	Find missing angles using angles on a straight line	8
16	Factorise a single bracket containing a non-linear expression by taking out a single algebraic common factor (e.g. a)	8
17	Write algebraic expressions with correct algebraic convention	8
18	Express one quantity as a percentage of another mentally	8
19	For a given algebraic input, find the output of a function machine	8
...

Bảng 4: Danh sách tần suất xuất hiện của từng ConstructName

Có thể thấy, top 20 khái niệm nền tảng phổ biến nhất có mặt trong bộ dữ liệu là những khái niệm liên quan đến các chủ đề **đại số** và **hình học cơ bản**.

Cột SubjectName

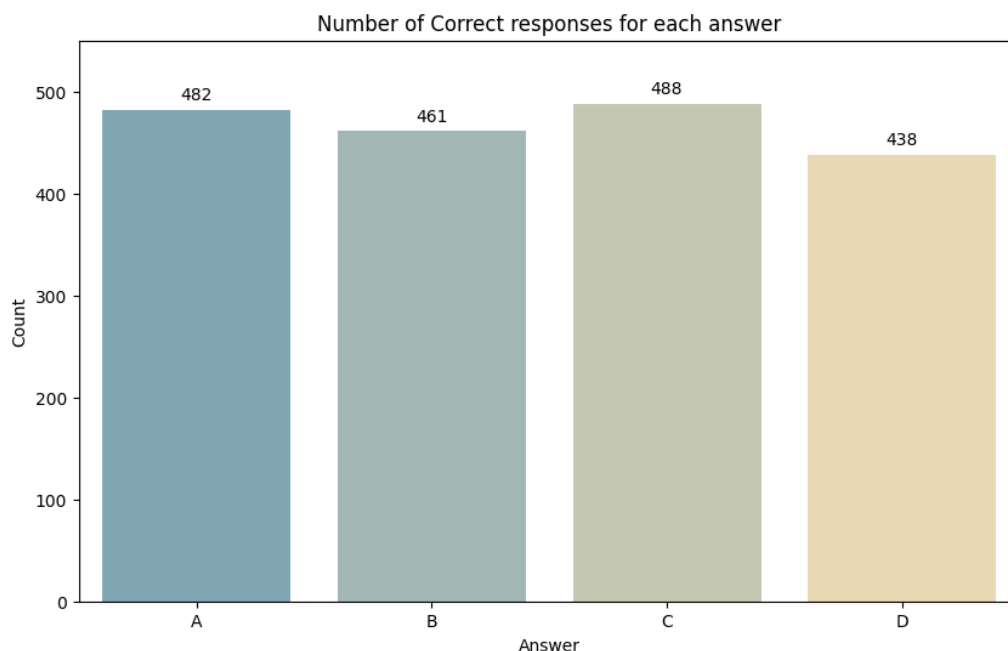
Index	SubjectName	Frequency
0	Linear Equations	53
1	Linear Sequences (nth term)	44
2	BIDMAS	37
3	Quadratic Equations	36
4	Area of Simple Shapes	36
5	Writing Expressions	31
6	Place Value	31
7	Substitution into Formula	30
8	Multiplying and Dividing with Decimals	30
9	Function Machines	30
10	Reflection	28
11	Adding and Subtracting Fractions	28
12	Time	28
13	Mental Multiplication and Division	28
14	Basic Angle Facts (straight line, opposite, around a point, etc)	27
15	Averages (mean, median, mode) from a List of Data	26
16	Mental Addition and Subtraction	25
17	Real Life Graphs	25
18	Percentages of an Amount	24
19	Squares, Cubes, etc	24
...

Bảng 5: Danh sách tần suất xuất hiện của từng SubjectName

Dữ liệu bao gồm nhiều nhóm chủ đề khác nhau: **đại số, hình học, số học, thống kê, và ứng dụng thực tế.**

- **Tần suất cao (30–50+):** Chủ đề nền tảng thường là những kiến thức cơ bản và quan trọng.
- **Tần suất trung bình (10–29):** Các chủ đề hỗ trợ hoặc mở rộng (ví dụ như: “Real Life Graphs”, “Factorising into a Double Bracket”) thường xuất hiện đủ để cung cấp ngữ cảnh nhưng không phải trọng tâm chính.
- **Tần suất thấp (<10):** Các môn học chuyên sâu, mang tính nâng cao hoặc ít gặp trong thực tế (ví dụ như: “Surface Area of Prisms”, “Algebraic Proof”).

Cột CorrectAnswer



Hình 3: Danh sách tần suất xuất hiện của từng CorrectAnswer

Số lần các đáp án là câu trả lời đúng phân bố đồng đều.

Các ngộ nhận phổ biến

MisconceptionId	MisconceptionName	Frequency
1214	When solving an equation, uses the same operation rather than the inverse.	54
1379	Rounds down instead of up	43
2316	Mixes up squaring and multiplying by 2 or doubling	38
1507	Carries out operations from left to right regardless of priority order	36
1990	Fails to reflect across mirror line	33
1880	Mixes up greater than and less than symbols	32
1597	Believes multiplying two negatives gives a negative answer	27
2392	Rounds to the wrong degree of accuracy (rounds too much)	27
220	Only multiplies the first term in the expansion of a bracket	22
1248	Rounds to the wrong degree of accuracy (rounds too little)	22
77	Does not follow the arrows through a function machine, changes the order of the operations asked.	22
1072	Multiplies by the index	20
2481	When multiplying decimals, divides by the wrong power of 10 when reinserting the decimal	19
2359	Believes division is commutative	19
2355	Does not know how to find order of rotational symmetry	19

MisconceptionId	MisconceptionName	Frequency
974	Believes multiplying a positive by a negative gives a positive answer	18
2271	When solving a problem that requires an inverse operation (e.g. missing number problems), does the original operation	18
1988	Rounds up instead of down	17
217	When adding fractions, adds the numerators and denominators	16
31	Does not understand bar modelling in algebra	16
1198	Does not recognise that a linear sequence must increase or decrease by same amount	16
108	Uses only the first two terms of a sequence to work out a term-to-term rule	16
113	Does not recognise difference of two squares	15
161	Answers as if there are 100 minutes in an hour when changing from hours to minutes	14
172	When subtracting fractions, subtracts the numerators and denominators	14
1527	Does not know how to find missing lengths in a composite shape	14
557	Assumes a fact without considering enough examples	14
1510	Believes subtraction is commutative	14
2143	Adds instead of multiplying when expanding bracket	14
1383	When two digits sum to 10 or more during an addition problem, does not add one to the preceding digit	13

Bảng 6: Danh sách các ngộ nhận phổ biến

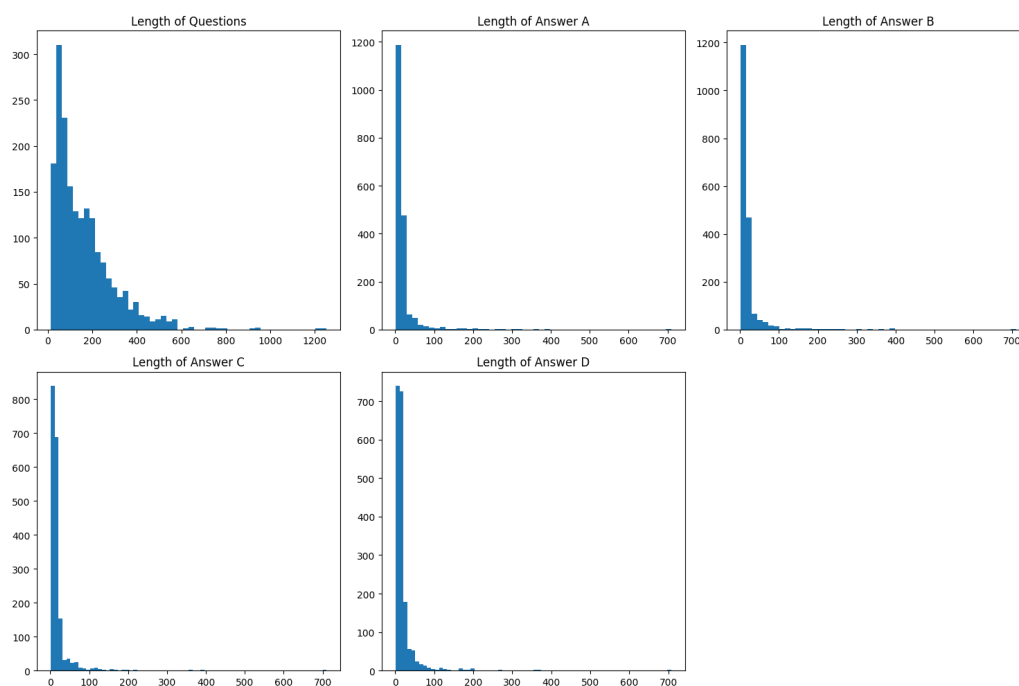
Hầu hết các ngộ nhận phổ biến (gặp ở nhiều câu hỏi) tập trung vào nhóm các quy tắc tính toán cơ bản:

- Thứ tự ưu tiên của phép toán, và nhầm lẫn các phép toán:
 - “When solving an equation, uses the same operation rather than the inverse” (54 lần).
 - “Carries out operations from left to right regardless of priority order” (36 lần).
 - “Mixes up squaring and multiplying by 2” (38 lần).
 - “Only multiplies the first term in the expansion of a bracket” (22 lần).
- Bản chất phép toán với số âm và phân số:
 - “Believes multiplying two negatives gives a negative answer” (27 lần).
 - “When adding fractions, adds the numerators and denominators” (16 lần).
- Làm tròn số:

- “Rounds down instead of up” (43 lần).
- “Rounds to the wrong degree of accuracy” (“too much” 27 + “too little” 22 lần).
- Nhận diện và sử dụng ký hiệu toán học:
 - “Fails to reflect across mirror line” (33 lần).
 - “Mixes up greater than and less than symbols” (32 lần).

Ngoài ra, còn có các ngộ nhận khác về tính toán số thực, hoặc các nhóm ngộ nhận khác nhưng chỉ đặc thù cho một số nhóm kiến thức/ lĩnh vực nâng cao hơn nên có tần suất xuất hiện thấp hơn.

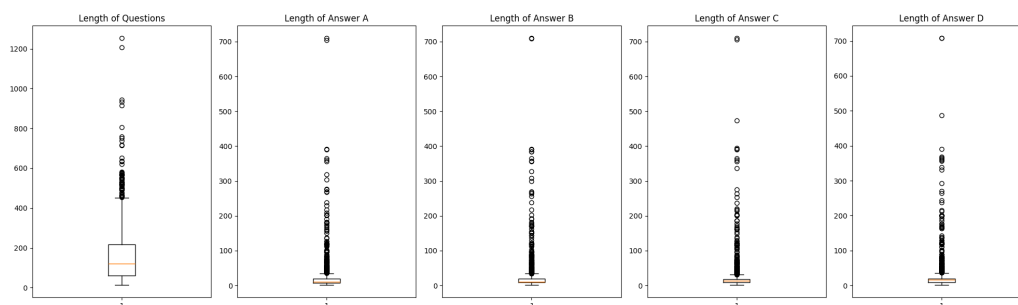
Phân tích độ dài câu hỏi và các câu trả lời



Hình 4: Danh sách tần suất xuất hiện của từng CorrectAnswer

- **Độ dài câu hỏi:** Phân bố khá rộng, có cả câu hỏi ngắn và dài, nhưng tập trung chủ yếu ở mức ngắn (dưới 200 ký tự) và trung bình (từ 200 đến 600 ký tự).
- **Độ dài của các câu trả lời:** Có phần phân tán hơn nhiều. Tập trung nhiều nhất vẫn là nhóm câu trả lời ngắn, nhưng vẫn có một số câu trả lời khá dài hoặc dài so với phần còn lại.

Nhận xét phân phối dữ liệu



Hình 5: Danh sách tần suất xuất hiện của từng CorrectAnswer

- Bộ dữ liệu có giá trị lớn và chứa nhiều thông tin hữu ích cho bài toán xác định mối quan hệ giữa các ngộ nhận và việc chọn đáp án sai.
- Bộ dữ liệu đa dạng về các phân môn và các phần kiến thức nền tảng trong toán học.
- Tuy nhiên, dữ liệu hiện tại hầu hết chỉ tập trung vào các bài toán với độ khó không quá cao. Hầu hết tập trung vào các phần kiến thức chung, chưa chuyên môn hoá. Vì vậy, mô hình sau khi huấn luyện có thể sẽ chạy tốt với các nhóm câu hỏi mức dễ, trung bình, khá, nhưng chưa thể giải quyết tốt được nhóm câu hỏi mức độ khó, hoặc thuộc các môn học nâng cao.

2.2 Tiền xử lý dữ liệu

2.2.1 Các ký tự dư thừa

Nhóm nhận thấy trong dữ liệu tồn tại các phần không cần thiết như những ký tự xuống dòng, khoảng trắng thừa, link,... nên sẽ tiến hành xử lý những ký tự này.

```
def clean_with_latex_preservation(x):
    x = re.sub("http\\w+", " ", x)
    x = re.sub(r"[\n\t\r]+", " ", x)
    x = re.sub(r"[ ]{2,}", " ", x)
    x = re.sub(r"(?!\\)\\\\s+", r"\\", x)
    x = x.strip()
    return x
```

Sau khi xử lý, các ký tự không cần thiết đã được loại bỏ. Tuy nhiên, nhóm nhận thấy các phần markdown đang có định dạng không tốt, ví dụ như `![text]()` cần xử lý loại bỏ.

```
def extract_markdown_text(x):
    processed_text = re.sub(r'!\[([^\]]*)\]\([^\)]*\)', r'\1',
```



```

        , x)

    return processed_text

```

2.2.2 Chuẩn hoá dữ liệu

Để chuẩn bị đầu vào cho mô hình ngôn ngữ lớn cần định dạng truy vấn với đầy đủ nội dung của dữ liệu. Đồng thời trong quá trình truy xuất dữ liệu, nhóm đã có xử lý với những trường hợp không có giá trị (NaN) bằng cách xử lý với những thông tin thay thế như sau:

```

question_text = row.get("QuestionText", "No question text
    provided")
subject_name = row.get("SubjectName", "Unknown subject")
construct_name = row.get("ConstructName", "Unknown construct")

correct_answer = row.get("CorrectAnswer", "Unknown")
assert wrong_choice != correct_answer
correct_answer_text = row.get(f"Answer{correct_answer}Text", "No
    correct answer text available")
wrong_answer_text = row.get(f"Answer{wrong_choice}Text", "No
    wrong answer text available")

```

Cấu trúc câu truy vấn hay prompt cho mô hình sẽ có đầy đủ thông tin gồm nội dung câu hỏi, câu trả lời đúng, câu trả lời sai mà học sinh đã chọn. Từ đó, mô hình có thể phân tích những ngộ nhận mà học sinh mắc phải.

```

TEMPLATE_INPUT = '{QUESTION}\nCorrect answer: {CORRECT_ANSWER}\n
    nStudent wrong answer: {STUDENT_WRONG_ANSWER}'

```

2.3 Xây dựng mô hình

2.3.1 Mô hình cơ sở

Trong quá trình xây dựng và áp dụng giải pháp để xác định ngộ nhận trong các bài toán toán học, việc lựa chọn mô hình ngôn ngữ lớn là yếu tố then chốt nhằm đảm bảo hiệu quả và độ chính xác. Sau khi cân nhắc nhiều yếu tố, nhóm đã lựa chọn sử dụng mô hình **Qwen2.5-14B-Instruct-AWQ**.

Mô hình Qwen2.5-14B được đánh giá cao nhờ khả năng xử lý ngôn ngữ tự nhiên ở mức độ nâng cao, đặc biệt trong các bài toán yêu cầu phân tích và suy luận sâu. Với kích thước tham số lớn, mô hình có thể hiểu ngữ cảnh và phân tích các mẫu phức tạp, phù hợp với các tình huống toán học đa dạng. Đồng thời, Qwen2.5-14B đã được tối ưu

hóa để giải quyết các vấn đề liên quan đến suy luận logic, một yếu tố quan trọng trong việc nhận diện sai lầm. Tính linh hoạt của mô hình còn cho phép tùy chỉnh và tinh chỉnh theo nhu cầu cụ thể, đảm bảo việc triển khai diễn ra thuận lợi và hiệu quả.

2.3.2 Tinh chỉnh mô hình

Để nâng cao hiệu suất của Qwen2.5-14B khi áp dụng vào bài toán xác định ngộ nhận toán học, nhóm đã quyết định sử dụng bộ dữ liệu Eedi trong giai đoạn tiền huấn luyện.

Quy trình này bao gồm hai thành phần chính: huấn luyện có giám sát (SFT) và ứng dụng kỹ thuật AWQ.

Huấn luyện có giám sát (Supervised Fine-Tuning - SFT)

Đầu tiên, dữ liệu sẽ được định dạng theo cấu trúc: [Chủ đề môn học] → [Câu hỏi] → [Phản hồi của học sinh] → [Mô tả ngộ nhận]. Cấu trúc này giúp mô hình học cách nhận diện ngộ nhận dựa trên ngữ cảnh.

Từ đó áp dụng huấn luyện `SFTTrainer` từ thư viện `trl` (Transformer Reinforcement Learning) được cung cấp bởi Hugging Face.

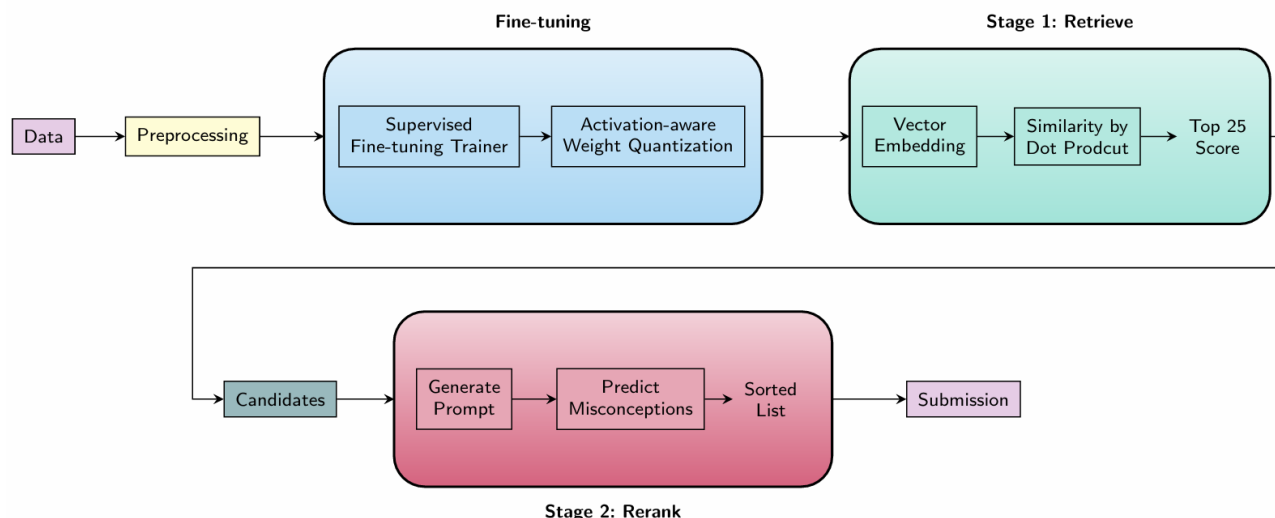
```
trainer = SFTTrainer(  
    base_model ,  
    train_dataset=dataset.select(range(680)),  
    eval_dataset=dataset.select(range(34)),  
    args=training_args ,  
    peft_config=peft_config ,  
    formatting_func=formatting_prompts_func ,  
    data_collator=collator  
)
```

Áp dụng kỹ thuật AWQ (Activation-aware Weight Quantization)

AWQ giúp biểu diễn trọng số của mô hình bằng các số nguyên 4-bit hoặc 8-bit thay vì sử dụng các số thực 16-bit hoặc 32-bit truyền thống. Điều này giảm đáng kể việc sử dụng bộ nhớ, đồng thời duy trì độ chính xác nhờ khả năng tương thích với động học kích hoạt của mô hình. AWQ đã cho phép mô hình Qwen2.5-14B chạy hiệu quả trên các GPU có bộ nhớ giới hạn, mang lại thời gian suy luận nhanh hơn và tiết kiệm tài nguyên.

2.4 Triển khai và đánh giá mô hình

Để tối ưu hóa hiệu quả nhận diện và phân tích ngộ nhận toán học, nhóm đã tiếp cận bài toán qua hai giai đoạn: **Truy xuất (Retrieve)** và **Xếp hạng (Rerank)**. Từ đó



Hình 6: Flow chart minh họa quy trình dự đoán ngộ nhận toán học

có thể tận dụng sức mạnh của mô hình Qwen và kỹ thuật tiên tiến để đạt hiệu suất cao trong nhận diện ngộ nhận.

2.4.1 Giai đoạn 1: Truy xuất

Đầu tiên, tạo các truy vấn chi tiết từ dữ liệu đầu vào cho mô hình bằng cách kết hợp giữa mô tả nhiệm vụ (`task_description`) và thông tin cụ thể từ dữ liệu đầu vào (`query`). Đồng thời nêu rõ ràng nhiệm vụ phân tích ngộ nhận, cách tiếp cận này giúp đảm bảo rằng mô hình nhận được đầy đủ thông tin ngữ cảnh, góp phần nâng cao hiệu suất mô hình.

```

def get_detailed_instruct(task_description: str, query: str) -> str:
    return f'<instruct>{task_description}\n<query>{query}'

task = "Given a math multiple-choice problem with a student's wrong answer, retrieve the math misconceptions"
queries = [
    get_detailed_instruct(task, q) for q in df_input['Prompt']
]

```

Sau khi xử lý dữ liệu, đầu vào vào được chuyển đổi thành vector embedding sử dụng mô hình ngôn ngữ lớn. Quá trình này bao gồm token hóa văn bản, tính toán trạng thái ẩn (hidden states) và trích xuất vector đại diện từ trạng thái cuối cùng của mô hình. Vector embedding này chính là kết quả tóm tắt ngữ nghĩa của văn bản.

```
batch_embeddings = last_token_pool(outputs.last_hidden_state,
                                   batch_dict["attention_mask"])
batch_embeddings = F.normalize(batch_embeddings, p=2, dim=1)
```

Để giảm tải bộ nhớ khi huấn luyện và suy luận trên các mô hình ngôn ngữ lớn, nhóm đã áp dụng kỹ thuật LoRA để điều chỉnh mô hình một cách hiệu quả thông qua:

- Giảm số lượng tham số cần cập nhật, từ đó giảm tải bộ nhớ và chi phí tính toán.
- LoRA chỉ thêm các tham số phụ trợ để huấn luyện, đảm bảo tính linh hoạt và dễ dàng áp dụng cho các mô hình đã được huấn luyện trước.

Nhóm sử dụng những vector embedding đại diện ngữ nghĩa để xác định mối liên hệ giữa sai lầm trong câu trả lời (truy vấn) và các ngộ nhận đã được lưu trữ (tài liệu). Mức độ tương đồng ngữ nghĩa giữa các cặp truy vấn - ngộ nhận được tính bằng tích vô hướng giữa vector nhúng của truy vấn (`query_embeddings`) và các vector nhúng của tài liệu (`doc_embeddings`).

```
scores = query_embeddings @ doc_embeddings.T
```

Sau khi tính toán điểm tương đồng, các ngộ nhận trong tài liệu được sắp xếp theo thứ tự giảm dần của điểm số cho mỗi truy vấn. Điểm càng lớn thể hiện sự tương đồng ngữ nghĩa càng cao, nhóm chỉ lấy danh sách **25 ngộ nhận có điểm tương đồng cao nhất** cho mỗi truy vấn. Việc giới hạn 25 ngộ nhận phù hợp nhất giúp tập trung vào những kết quả có giá trị nhất, đồng thời giảm bớt độ phức tạp trong xử lý và phân tích tiếp theo.

2.4.2 Giai đoạn 2: Xếp hạng

Để tăng độ chính xác cho kết quả cuối cùng, nhóm đã thiết kế kỹ lưỡng một prompt cung cấp đầy đủ ngữ cảnh, bao gồm các thông tin:

- Mô tả ngữ cảnh: Tên khái niệm (`ConstructName`) và môn học liên quan (`SubjectName`).
- Chi tiết câu hỏi: Nội dung câu hỏi (`Question`), đáp án đúng (`CorrectAnswer`), và đáp án sai (`IncorrectAnswer`).
- Danh sách ngộ nhận: Một tập các ngộ nhận tiềm năng (`Retrieval`) được lấy từ cơ sở dữ liệu, sắp xếp dựa trên điểm tương đồng với truy vấn.

```
PROMPT = """Here is a question about {ConstructName}({
    SubjectName}).
Question: {Question}
```

```
Correct Answer: {CorrectAnswer}
Incorrect Answer: {IncorrectAnswer}

You are a Mathematics teacher. Your task is to reason and
    identify the misconception behind the Incorrect Answer with
    the Question.
Answer concisely what misconception it is to lead to getting the
    Incorrect Answer.
Pick the correct misconception number from the below:

{Retrieval}
"""
```

Nhóm đã lựa chọn vLLM làm nền tảng để triển khai mô hình Qwen2.5-14B nhờ vào khả năng tối ưu hóa mạnh mẽ trong việc suy luận trên các mô hình ngôn ngữ lớn. vLLM hỗ trợ song song hóa GPU và quản lý bộ nhớ hiệu quả, giúp vận hành các mô hình có yêu cầu tài nguyên cao như Qwen2.5-14B một cách trơn tru.

Quy trình xếp hạng và dự đoán ngộ nhận

- **Xếp hạng ngộ nhận qua từng vòng lặp:**

- Tạo danh sách ngộ nhận tiềm năng (`c_indices`) dựa trên điểm tương đồng từ các vòng trước đó.
- Ghép danh sách này vào prompt để mô hình ngôn ngữ phân tích và lựa chọn.

- **Dùng LLM để chọn ngộ nhận:**

- Áp dụng prompt cho từng câu hỏi trong tập dữ liệu.
- Sử dụng mô hình để dự đoán ngộ nhận phù hợp nhất thông qua các tham số:
 - * Chế độ lấy mẫu (`top_k=1`, `temperature=0`) đảm bảo kết quả có độ chính xác cao.
 - * **Logits processor:** Giới hạn lựa chọn các ngộ nhận trong danh sách đã cung cấp (`MultipleChoiceLogitsProcessor`).

- **Lưu kết quả và tối ưu qua các vòng lặp:**

- Mỗi vòng, LLM chọn ngộ nhận tốt nhất từ danh sách.
- Các ngộ nhận được cập nhật, giảm thiểu nhiễu và tăng độ chính xác.

Kết quả phân tích

Here is a question about: *Use the order of operations to carry out calculations involving powers (BIDMAS).*

Question:

$$3 \times 2 + 4 - 5$$

Where do the brackets need to go to make the answer equal 13?

Correct Answer:

$$3 \times (2 + 4) - 5$$

Incorrect Answer:

$$3 \times 2 + (4 - 5)$$

Task: You are a Mathematics teacher. Your task is to reason and identify the misconception behind the Incorrect Answer with the Question.

Answer concisely what misconception it is to lead to getting the Incorrect Answer.

Pick the correct misconception number from the below:

1. Performs subtraction right to left if priority order means doing a calculation to the right first
2. Believes order of operations does not affect the answer to a calculation
3. Performs addition ahead of any other operation
4. Answers order of operations questions with brackets as if the brackets are not there
5. Performs addition ahead of subtraction
6. Does not interpret the correct order of operations from a worded problem
7. Confuses the order of operations, believes subtraction comes before multiplication
8. Believes that the order of a worded calculation should be changed to follow BIDMAS
9. Has not realised that the answer may be changed by the insertion of brackets

Kết quả trên cho thấy hệ thống đã hoạt động tốt trong việc cung cấp ngữ cảnh và hỗ trợ phân tích lỗi sai. Một số điểm nổi bật như:

- **Tính chính xác:** Prompt được thiết kế rõ ràng đã giúp mô hình nhận diện chính xác các ngộ nhận tiềm năng.
- **Khả năng suy luận:** Hệ thống không chỉ xác định lỗi sai mà còn đưa ra danh sách các ngộ nhận liên quan để hỗ trợ giáo viên đánh giá.
- **Ứng dụng thực tiễn:** Các thông tin được trình bày có cấu trúc và rõ ràng, phù hợp để áp dụng trong giáo dục, đặc biệt trong việc phân tích lỗi học sinh.

2.4.3 Đánh giá

Quy trình đánh giá mô hình sử dụng các độ đo **Average Precision** (AP) và **Mean Average Precision** (MAP) để đo lường độ chính xác của dự đoán.

- **Average Precision (AP):** AP được tính cho một cặp nhãn thực tế và dự đoán, giới hạn trong *top-k* kết quả. Công thức tính như sau:

$$AP = \frac{1}{\min(\text{len}(\text{actual}), k)} \sum_{i=1}^k \frac{\text{num_hits}}{i}$$

Trong đó:

- **num_hits:** Số lượng dự đoán đúng.
- **i:** Vị trí của dự đoán trong danh sách.

- **Mean Average Precision (MAP):** MAP là trung bình của AP trên toàn bộ danh sách nhãn thực tế và dự đoán:

$$MAP = \frac{1}{N} \sum_{i=1}^N AP_i$$

Trong đó N là tổng số cặp nhãn thực tế và dự đoán.

Bên cạnh việc thiết kế và triển khai hệ thống, nhóm đã đạt được kết quả đáng khích lệ trên hệ thống chấm điểm của cuộc thi. Điểm số do hệ thống chấm của cuộc thi ghi nhận là minh chứng cho sự chính xác và khả năng ứng dụng thực tế của giải pháp.









Leaderboard Raw Data Refresh

Q HCMUS

Public Private

The private leaderboard is calculated with approximately 72% of the test data. This competition has completed. This leaderboard reflects the final standings.

Prize Winners

#	△	Team	Members	Score	Entries	Last	Solution
114	▲ 15	HCMUS - Senior Students	  	 0.47056	346	1mo	
152	▲ 21	HCMUS-2425-1	   	0.45878	27	1mo	

Hình 7: Kết quả nhóm đạt được

Tài liệu tham khảo

Tiếng Anh

- [1] Aguilera, Frank Morales. *SFTTrainer: A Comprehensive Exploration of Its Concept, Advantages, Limitations, History, and Applications*. 2024. URL: <https://medium.com/thedeephub/sfttrainer-a-comprehensive-exploration-of-its-concept-advantages-limitations-history-and-19ab0926e74e>.
- [2] Albayrak, Emre. *Qwen Instruct AWQ - vLLM Tutorial*. 2024. URL: <https://www.kaggle.com/code/cemalemrealbayrak/qwen-2-5-14b-instruct-awq-vllm>.
- [3] Dawood, Muhammad. *Understanding the competition : EDA and Overview*. 2024. URL: <https://www.kaggle.com/code/muhammaddawood42/understanding-the-competition-eda-and-overview>.
- [4] Jiang, Zhuoxuan. *LLMs can Find Mathematical Reasoning Mistakes by Pedagogical Chain-of-Thought*. 2024. URL: <https://arxiv.org/html/2405.06705v1>.
- [5] Kiran, Jagat. *Qwen2.5-32B-AWQ for identifying likely misconceptions in reranking*. 2024. URL: <https://www.kaggle.com/code/jagatkiran/qwen14b-retrieval-qwen32b-logits-processor-zoo>.
- [6] Prata, Marília. *Misconceptions Distractors in Maths*. 2024. URL: <https://www.kaggle.com/code/mpwolke/misconceptions-distractors-in-maths>.
- [7] Sadihin, Bryan Constantine. *Mining Misconceptions in Mathematics*. 2024. URL: <https://openreview.net/pdf?id=CE85qdNSlp>.
- [8] Vo, Anh. *Retrieval with Qwen-14B-AWQ*. 2024. URL: <https://www.kaggle.com/code/anhvth226/eedi-11-21-14b>.