CIS4900: Large Language Models for Detecting Student Misconceptions and Providing Feedback

Ella Luedeke

January 2025

1 Introduction

Students often have misconceptions in STEM courses due to the focus on application and real-world examples, not conceptual material. Writing exercises can pinpoint these misconceptions, but are usually deemed not worth the effort of manual grading. Developing a tool that uses Natural Language Processing and Machine Learning to do such a process autonomously can help instructors reduce misconceptions and increase learning.

Data was taken from written responses from students in an introductory physics course. The data was cleaned to be used in the machine learning algorithm using standard techniques such as tokenization, which breaks down the text into words and phrases to parse the text. In context learning was used to determine whether the student's response was a misconception and the category of the misconception. This is accomplished by training the model on established examples so that it can replicate on new data points. A report is then generated for the instructor and student to enhance learning.

2 Related Work

Research has shown that when students explain concepts to themselves, it helps them learn by identifying gaps in their knowledge and giving professors insight into these gaps. One of these studies is "Assessing Student Explanations with Large Language Models Using Fine-Tuning and Few-Shot

Learning" by Carpenter et al. [1]. The paper discusses the ExplainIt Classroom Response System, which utilizes large language models to grade student responses in STEM courses. Llama 2, GPT-3.5, and GPT-4 were used to automate the assessment of student knowledge. Two methods were used: few-shot learning and fine-tuning. In few-shot learning, the GPT-4 model was given ten labeled student responses and used them as a reference to predict how to label new responses into multiple classes (e.g., "correct," "partially correct," and "incorrect"). Fine-tuning used the FLAN-T5 model, which was trained on a data set of responses and labels and customized for the task. The fine-tuning model achieved the highest accuracy and performance, but requires a lot of training data. The few-shot performed well in handling diverse answers and without large data sets, but is expensive.

Limitations included comparing the results of fine-tuned models and the few-shot models. The fine-tuned models were trained on 90 percent of the data, while the few-shot models used ten student responses to be trained. The few-shot model has technical limitations, such as context length and cost. The fine-tuning model requires substantial labeled data.

3 Methodology

3.1 Data

Initial data analysis showed that the training data set contained three distinct misconceptions: acceleration limited to speed (alts), acceleration signs limited (asl), and quantity and rate (qar). 66 student responses were collected, then tokenized down to sentences using NLTK [1]. They were labeled by the physics instructors. The distribution of labels is fairly unbalanced, as seen in Figure 1, with "alts" containing 22, "asl" containing 24, and "qar" containing 5. While all misconceptions were experimented with, the primary focus was on the acceleration limited to speed misconception, which is when students associate acceleration with only speeding up and slowing down. The data was then split into a train and test set. The train set was sentences and the test set was sentences. With each individual experiment, the train set was then split further into a new train and test, called the validation set.

[1] https://www.nltk.org/

This done with the sklearn [2] python module, and stratified with respects to the class.

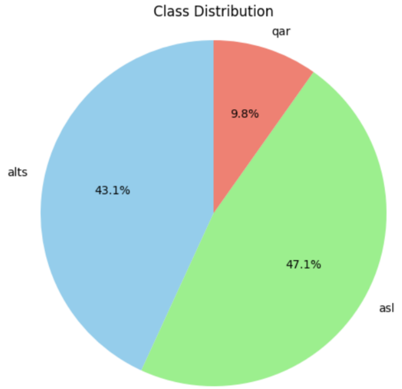

Figure 1: Pie chart demonstrating class distribution

3.2 Models

Open source models were chosen for cost effectiveness and accessibility. Ollama's selection of LLMs were analyzed: settling on Llama3.2, Phi4, Gemma, and Mistral. Through the course of the study, Llama3.2 and Phi4 were found to be the most consistent performance and were prioritized.

3.3 Experimental Set Up

Certain metrics were selected to evaluate each experiment's performance. F1, the harmonic mean between precision and recall, was the primary metric. Precision and recall were also calculated. The formulas are below.

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

[2] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html

3.3.1 Zero Shot Prompting

Zero shot prompting is a type of in-context learning used to prompt LLMs, where examples are not provided. Only context about the task at hand and background information is provided. The training set was used to test the Llama 3.2 model on zero-shot prompting, utilizing in-context learning. The model was given instructions to act as an expert physics professor and definitions of the misconceptions. The model was given a multi-classification problem where it could classify a given student sentence into one, multiple, or none of the three misconception classes.

Multi-label prompt:

> Take on the role of an expert physics professor. Your task is to classify student responses into one, multiple, or none of the following misconception categories: ...
>
> Student response: {text}
>
> Respond only with the misconception label(s).

A simplified binary version focused on alts:

> You are an expert physics professor... Determine whether the response contains the "alts" misconception. Output ONLY 0 (no misconception) or 1 (misconception).

3.3.2 Few Shot Prompting

As mentioned in the data section, validation set was created from about 20% of training data. Few shot prompting was then ran. Few shot prompting is very similar to zero shot in that it receives the same information: context, task, and a description of the misconception. The difference is now example student responses that both contain and do not contain this misconception are given to the model and it is asked to identify possible misconceptions based on that information. The classification was reduced to a binary problem for training, focusing on a specific misconception at a time. Experiments varied the number of examples (k) and observed the impact on F1 score. Sentence similiarity filtering using `thefuzz` [2] was also used to tailor the response set to the specific question that corresponded to the misconception, which is explored more later.

Few Shot Prompt

> You are an expert physics professor looking for misconceptions in student responses. -'alts' = acceleration limited to speed, which is when students associate acceleration with only speeding up and slowing down.
>
> Here are examples that contain the misconception. positive examples
>
> Here are examples that do not contain the misconception. negative examples
>
> Student response: text
>
> Task: Determine whether the student response contains the misconception using the examples given.
>
> Response: Output ONLY a 0 if it does not contain the misconception or a 1 if it does.

3.3.3 Chain of Thought

Chain of thought (CoT) prompting is a different type of prompting, designed to tailor to LLMs architecture by asking them to reason through their answer. These intermediary steps allow a more complex response. It can be done in zero shot or few shot, similar to the methods previously explored. In this study, both were utilized. Zero shot chain of thought was implemented with the prompt "let's think step by step." It resulted with raising the F1 to be about 0.3, but lowered recall.

Zero Shot CoT Prompt

> Take on the role of an expert physics professor. Your task is to determine whether student responses contain the following misconception: 'alts' = acceleration limited to speed, which is when students associate acceleration with only speeding up and slowing down.
>
> Student Response: "text"

Let's think step by step.

Reasoning: (Explain step-by-step analysis here, limit to one sentence with no numbers)

Result: (Output ONLY 0 for does not contain or 1 does contain on a new line.)

Few Shot CoT Prompt

Take on the role of an expert physics professor. Your task is to determine whether student responses contain the following misconception:

'alts' = acceleration limited to speed, which is when students associate acceleration with only speeding up and slowing down.

Below are examples of how to reason through responses:

Example 1: Student Response: "An object with constant acceleration will constantly get faster, where an object with constant speed will neither accelerate nor decelerate."

Reasoning: The student equates acceleration with speed increase only, showing a misunderstanding that constant acceleration always means "getting faster." Result: 1

—

Example 2: Student Response: "This makes sense because if the acceleration is held constant then the speed must also be increasing."

Reasoning: The student recognizes the relationship between constant acceleration and increasing speed without misrepresenting acceleration as only speed change. Result: 0

—

Now analyze the new response below:

Student Response: "text"

Let's think step by step.

Reasoning: (Explain step-by-step analysis here, limit to one sentence with no numbers)

Result: (Output ONLY 0 for does not contain or 1 does contain on a new line.)

3.3.4 Recognizing Textual Entailment

Another type of prompting was derived from a task in natural language processing called recognizing textual entailment (RTE) [3]. Given the perfect answer or gold-standard sentence created by the instructor, the models were then asked to find contradictions in the student answers.

RTE Prompt

> You are an expert physics professor grading student responses.
>
> Questions: Describe what happens to the man when he is accelerating? What is the difference between an object with constant acceleration and an object with constant speed?
>
> Perfect Answers: He moves to the right and gets faster to the right. This means he covers more distance in a given time interval An object with constant speed covers the same distance in a given time interval. Assuming the motion is one-dimensional, constant speed means 0 acceleration. Constant acceleration, however, means the velocity changes by the same amount during each time interval, and the distance covered in each time interval increases
>
> Student Response: text Task: Determine whether the student response contradicts the perfect answers.
>
> Response: Output ONLY a 0 if it does not contradict or a 1 if it does.

3.3.5 Filtered Data Set

Towards the end of the study, the earlier experiment of filtering the data set was revisited. The train set was filtered utilizing thefuzz [2] token set ratio method with the sentence that corresponding most to the ALTS class, "What is the difference between an object with constant acceleration and an object with constant speed?" Only sentences with a similarity score greater than 60 were kept, resulting in a new dataset with 22 positives and 67 negatives.

The new dataset was then split using sklearn's [3] train test split function into a stratified train/test set with respect to the 'ALTS' class. Three different methods of prompting: zero shot, few shot, and RTE, were then run again with the filtered data.

4 Results and Discussion

Overall, few shot did not necessarily increase performance but generally outperformed zero shot. The issue with all prompting was consistently low precision but higher recall. CoT prompting gave extremely varied and inconsistent results. RTE also had limited success, likely due to the nuance of each answer not exactly matching the gold standard sentence. Regardless of the prompting method, filtering the data set with regard to the 'alts' class largely improved performance. The engineering of the prompt also mattered with simplifying output expectations to strictly 0 and 1 led to more consistent behavior. Due to the large increases with filtering, a data set that corresponds to one specific question would likely dramatically increase performance. However, the models did tend to overfit the filtered experiments with 1s, likely due to the decreased size of the data set.

4.1 Zero-Shot Results

The F1 scores for each misconception are listed in Table 1.

4.1.1 F1 Scores per Misconception

Misconception	F1 Score
alts	0.031
asl	0.092
qar	0.027

Table 1: F1 scores for each misconception.

[3] `https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html`

4.1.2 Binary Classification for *alts*

Metric	Value
Accuracy	0.1204
Precision	0.0594
Recall	1.0000
F1 Score	0.1121
LLM Identified	101
Total Misconceptions	6

Table 2: Binary classification results for *alts*.

4.2 Few-Shot Prompting Results

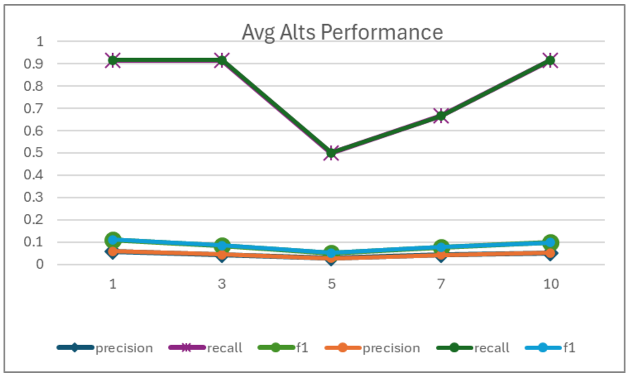

Figure 2: Average performance on **alts** with few-shot prompting.

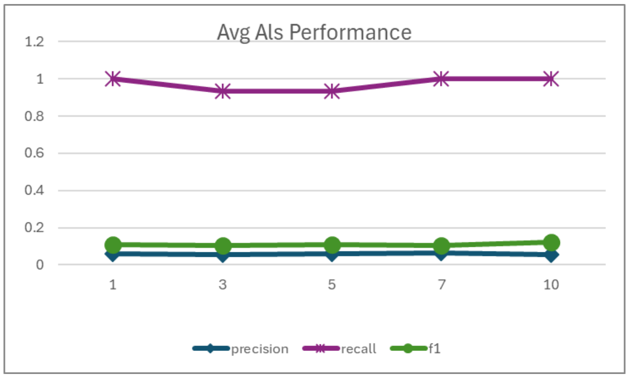

Figure 3: Average performance on **asl** with few-shot prompting.

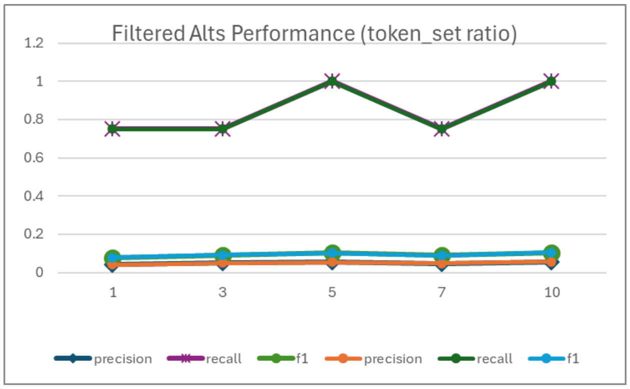

Figure 4: Token-based filtering results for **alts** classification.

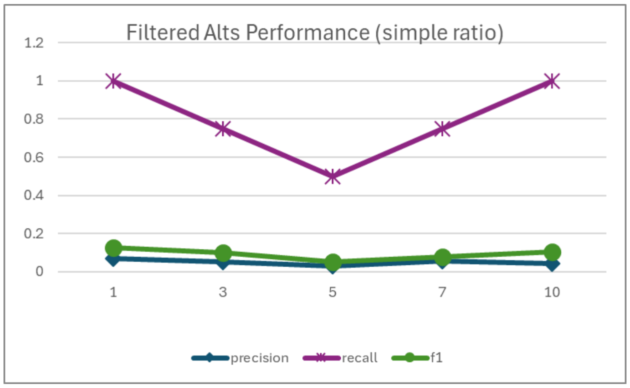

Figure 5: Simplified filtering results for **alts** classification.

4.2.1 Model Comparison (Few-Shot, $k = 3$)

Model	F1 Score	Accuracy	Precision	Recall
Phi4	0.2222	0.7407	0.1290	0.8000
Mistral	0.1124	0.2685	0.0595	1.0000

Table 3: Few-shot model comparison results ($k = 3$).

4.3 Chain-of-Thought (CoT) Results

4.3.1 Zero-Shot CoT

Model	F1 Score	Accuracy	Precision	Recall
LLaMA 3.2	0.1290	0.7500	0.0800	0.3333

Table 4: Zero-shot CoT results.

4.3.2 Few-Shot CoT

Model	F1 Score	Accuracy	Precision	Recall
Phi4	0.1739	0.8241	0.1176	0.3333
LLaMA 3.2	0.2222	0.9352	0.3333	0.1667

Table 5: Few-shot CoT results.

4.4 RTE Results

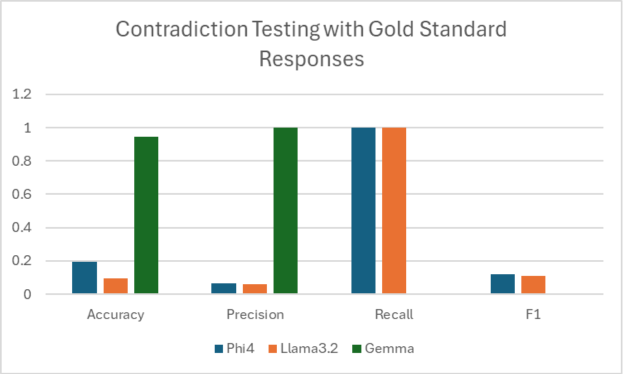

Figure 6: Contradiction detection performance.

4.5 Finalized Test Set

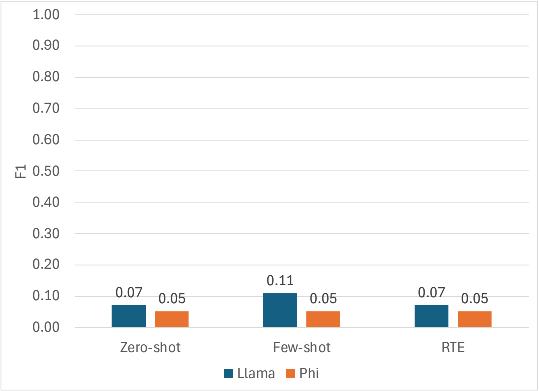

Figure 7: Original data with different types of prompting on the test set.

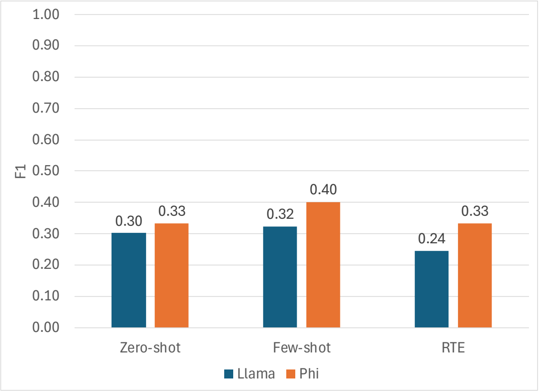

Figure 8: Filtered data with different types of prompting on the test set

5 Conclusion and Future Work

Through the course of the study, the ability for LLMs to detect misconceptions in physics students' written assignments were thoroughly tested. Various types of prompting and models were used. No clear method or model emerged as the most effective. Filtering the data, however, proved extremely valuable with all models and types of prompting. The initial results show signs of promise in the future, but the work is far from over. Future works could experiment with different models such as as Mistral or Gemini. Chain of thought could be explored more thoroughly to see if consistent results could be acquired. Additionally, fine-tuning could be used on the pre-trained models for potential increases.

References

[1] Dan Carpenter, Wookhee Min, Seung Lee, Gamze Ozogul, Xiaoying Zheng, and James Lester. Assessing student explanations with large language models using fine-tuning and few-shot learning. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 403–413, 2024.

[2] SeatGeek Cohen and contributors. Thefuzz: Fuzzy string matching in python. `https://github.com/seatgeek/thefuzz`, 2021. Python package version 0.19.0.

[3] I Made Suwija Putra, Daniel Siahaan, and Ahmad Saikhu. Recognizing textual entailment: A review of resources, approaches, applications, and challenges. *ICT Express*, 10(1):132–155, 2024.