

RAMP: Exploring the Feasibility of Detecting Physics Student Misconceptions in Writing Assignments Using Large Language Models

Ella Luedeke*, Natalie Spiro, Indika Kahanda†, W. Brian Lane, J. Caleb Spiers, Terrie Galanti, Upulee Kanewala

University of North Florida, Jacksonville, FL, USA

n01553443@unf.edu, n01402403@unf.edu, indika.kahanda@unf.edu, brian.lane@unf.edu, j.caleb.speirs@unf.edu, terrie.galanti@unf.edu, upulee.kanewala@unf.edu

Abstract

Students in introductory STEM courses frequently have misconceptions about the material. Writing assignments can help instructors identify these, but are often impractical and time-consuming to grade, especially in large classes. In this study, we curated student responses from an introductory physics assignment based on a misconception related to motion. We formulated the task of identifying misconceptions within a sentence as a binary classification task and developed the RAMP (Reporter of Aggregated Misconceptions in Physics) classifier using ModernBERT. Experimental results indicate that RAMP is effective for identifying student misconceptions, noticeably outperforming various prompting techniques using several LLMs and traditional machine learning classifiers. With the refinement of hyperparameters and additional data, RAMP may be improved to an acceptable level, where it can be used as the back-end of an instructor-facing tool that reports student misconceptions across writing assignments in introductory physics courses.

Introduction and Background

Student misconceptions in challenging STEM courses represent a persistent educational challenge that can significantly impede learning outcomes (Resbiantoro, Setiani, and others 2022; Chen et al. 2020; Neidorf et al. 2019; Latif and Zhai 2025; Zhou, Kim, and Ahmed 2024). In introductory Physics courses, students often enter with misconceptions, and the instructional emphasis fails to resolve those misconceptions. Although writing assignments provide valuable opportunities for instructors to identify and address these misconceptions (Zarestky et al. 2022), the practical constraints of grading, particularly in large-enrollment courses, often make them prohibitively time-consuming to implement effectively (Arthurs, Hsia, and Schweinle 2015).

Recent advances in large language models (LLMs) present promising solutions for automating student writing assessment and providing personalized feedback (Kökver, Pektaş, and Çelik 2025; Carpenter et al. 2024a; Chen and Singh 2025; Chen and Wan 2025). These technologies offer

the potential to make writing-based assessments more feasible on a scale while maintaining pedagogical value through the timely identification of misunderstandings (Bernius, Krusche, and Bruegge 2022; Ferreira Mello et al. 2025).

Savage and Rebello (Savage and Rebello 2025) developed LLM-based methods to investigate students' explanations on conceptual physics questions, which demonstrated that a commercial LLM can "fairly assess students' explanations, comparable to human graders". Carpenter et al. (Carpenter et al. 2024b) demonstrated that fine-tuning and few-shot learning of LLMs can be effective for assessing student explanations in Computing. Sonkar et al. (Sonkar et al. 2024) developed open-source LLM-based cognitive models of students with misconceptions in algebra.

Traditional Machine Learning (ML) techniques have also proven successful at discovering misconceptions and other educational uses. Gomes and Jaques (Gomes and Jaques 2023) created a Density-Based Spatial Clustering of Applications with Noise algorithm to identify algebraic misconceptions from student responses. Weegar and Idestam-Almquist (Weegar and Idestam-Almquist 2023) found success in reducing instructor workloads utilizing Random Forest, Support Vector Machines, and a Complement Naive Bayes' (CNB) classifier models to automate grading.

This study investigates the feasibility of identifying physics misconceptions by formulating this task as a supervised binary classification. We developed the RAMP (Reporter of Aggregated Misconceptions in Physics) classifier using ModernBERT (Bidirectional Encoder Representations from Transformers) with supervised fine-tuning. We compare RAMP to two baseline approaches: (a) in-context learning with various prompting techniques on LLMs, and (b) classical text classification models based on multinomial naive bayes (MNB), and Light Gradient Boosting Machine (LightGBM). By analyzing student responses from an introductory physics course labeled with a motion-related misconception type, we compare the effectiveness of these approaches in identifying misconceptions in student writing. To the best of our knowledge, this is the first study to explore the use of both fine-tuning and prompting of open-source/commercial LLMs for identifying student misconceptions in Physics.

Our findings indicate that, while all methods have the potential to identify sentences containing misconceptions,

*Corresponding author: n01553443@unf.edu

†Corresponding author: indika.kahanda@unf.edu

RAMP performs the best. In-context-learned models struggle with precision, often producing false positives that still require instructor intervention, while MNB performs relatively better in terms of precision but with lower recall. Overall, there is promise that further prompt refinement and/or more training data may improve overall RAMP performance, reaching accuracy levels suitable for use at the back-end of an instructor-facing tool that reports misconceptions across writing assignments in introductory physics courses.

Methodology

We formulate the task of detecting a misconception within a student's written response as a supervised binary classification problem, where each sentence in the response is annotated with a binary label indicating the presence or absence of the misconception. This reading assignment on kinematics, shown in Figure 1, was administered in an introductory physics course in the Fall of 2023.

Data

66 consented student responses were collected and anonymized (this study was determined to be exempt from full review by the University of North Florida Institutional Review Board [Protocol #IRB-FY2024-84]). In this assignment, students manipulated position, velocity, and acceleration values in a *PhET* simulation (Wieman, Adams, and Perkins 2008), depicted in Figure 1, and then discussed their observations in relation to the reading from the textbook. Complete student responses (i.e., answers to all parts) were gathered together. Course instructors found that misconceptions were sentence specific, as each sentence tended to relate to one misconception. Therefore, the responses were first tokenized into separate sentences using NLTK (<https://www.nltk.org/>). Then, each sentence was labeled by the course instructor with respect to a misconception coded as ALTS (acceleration limited to speed), which occurs when students associate acceleration with only speeding up and slowing down, without acknowledging that changing direction also constitutes acceleration.

Two members of the research team with disciplinary backgrounds in physics reviewed the responses and helped resolve any labeling ambiguities. A sentence was labeled positive if it contained the misconception; otherwise, it was labeled negative. No misconceptions labels spanned multiple sentences. This, highly imbalanced, original dataset was composed of 753 sentences (30 positives).

Below is an example of a positive sentence (shown in red) and a negative sentence (shown in green). Note that the positive sentence directly associates constant acceleration with speed changing, while the negative example refers to the velocity changing, leaving open the possibility of a change in magnitude or direction.

“an object with constant speed moves at a constant rate, while an object with constant acceleration changes speed at a constant rate...”

“An object in constant speed has an acceleration of zero, while an object in constant acceleration has velocity that is continuously changing with time, so its velocity will not be constant...”

Models

We developed the RAMP classifier by fine-tuning ModernBERT (Warner et al. 2024). As LLM baselines, we used the following LLMs with prompting: Qwen (Bai et al. 2023) and GPT5 (OpenAI 2023). As non-LLM baselines, we developed a Multinomial Naive Bayes and LightGBM model.

Experimental Setup and Metrics

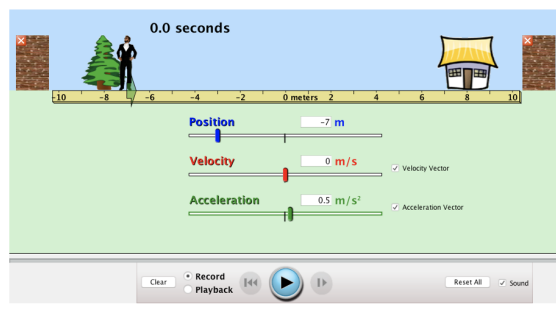
The original labeled dataset was split into a *train* (75%, 516 sentences, 22 positives) and *test* (25%, 222 sentences, 8 positives) set with stratification. This form of hold-out validation technique was preferred over k-fold cross validation to avoid overfitting. Using the train set, we confirmed that all occurrences of ALTS misconceptions were labeled to the sentences in response specifically to sub-question 8: “What is the difference between an object with constant acceleration and constant speed?”. So, we created a *filtered* train set by removing all sentences that are semantically dissimilar to that sub-question prompt using *thefuzz* (<https://github.com/seatgeek/thefuzz>) library. This filtered train set contains 89 sentences, with 22 positives still retained, making it a much more refined dataset that showed the difference between the positive and negative classes. We split the filtered train data into train (80%, 71 instances, 18 positives) and validation (20%, 18 sentences, 4 positives) sets using random stratified splitting. These two datasets were used to develop all our classifiers.

We implemented RAMP with ModernBERT-base (answerdotai/ModernBERT-base) via HuggingFace (<https://huggingface.co/>). We applied SMOTE (synthetic minority oversampling technique) (Chawla et al. 2002) to expand the dataset to 50/50 class distribution (i.e., 53 positives/negatives each), and used that to fine-tune with AdamW optimizer for 10 epochs (with early stopping). We used batch size=16 and weight decay=0.01. While none of the hyperparameters were tuned, the decision to use SMOTE was determined by validation results of the fine-tuned model.

As for LLMs with In-context Learning, we used Qwen2.5-8B (temperature: 0.7) on Ollama (<https://ollama.com/>), and GPT5 (temperature: 1.0) via the OpenAI API (<https://openai.com/api/>). While with each LLM, we evaluated *zero-shot*, *few-shot*, and chain-of-thoughts (CoT) prompts (Kojima et al. 2022), we only report the results for the best-prompting strategy for Qwen and GPT. We utilized the validation data to determine the optimal number of examples to use with few-shot and CoT prompts (data not shown). Initially, we experimented with several other open-source models: Llama3.2, Mistral, GPT-OSS, and Phi4. However, based on their inferior performance on the validation data, we chose only Qwen for final evaluation.

Read the web page: Calculating with constant acceleration. To give you some practice on working with graphs showing constant acceleration, go to the PhET simulation Moving Man to perform a few tasks. By typing numbers into the starting value spaces, set the man to stand near the tree. Give him a velocity of 0 m/s and an acceleration of 0.5 m/s². Click the boxes to show the velocity and acceleration vectors. Your screen should look like the figure below.

Start the man in motion until he hits the wall, then hit pause to stop recording. Use the playback feature to answer these questions.



Answer these questions:

1. What happened to the blue position slider as the man moved across the screen?
2. What happened to the red velocity slider as the man moved across the screen?

Now look at your position-time, velocity-time and acceleration-time graphs.

3. What shape is your position graph? Does this make sense? Why?
4. What shape is your velocity graph? Does this make sense? Why?
5. What is the slope of your velocity graph? Does this make sense? Why?
6. What does the slope of the velocity graph represent?
7. Describe what happens to the man when he is accelerating?
8. What is the difference between an object with constant acceleration and an object with constant speed?

Figure 1: The kinematics reading assignment administered in an introductory college Physics course.

We used scikit-learn (<https://scikit-learn.org>) to implement both the Multinomial Naive Bayes and LightGBM classifiers. TF-IDF Vectorizer was used for pre-processing. None of the hyperparameters were tuned, but the same SMOTE process used by RAMP was applied. All classifiers were evaluated using Fmax (maximum possible F1 score), Precision at Fmax, and Recall at Fmax.

The code for RAMP, Multinomial Naive Bayes, and LightGBM classifiers was developed with the assistance of Gemini. Qwen, RAMP, Multinomial Naive Bayes, and LightGBM classifiers were trained on Google Colab for Education (free for educators and students) using a T4 GPU with 12.7 GB System RAM, 15 GB GPU RAM, and 112.6 GB Disk space. The total runtime for all Qwen prompts is approximately 48 minutes. The approximate runtimes for RAMP, Multinomial Naive Bayes, and LightGBM classifiers were 5 mins, \approx 1 minute, and \approx 1 minute. None of the above experiments incurred any monetary costs. GPT5 was run via the OpenAI API calls on Google Colab, which incurred a cost of \$5.32.

Results

As evident from Figure 2, our RAMP classifier comfortably outperformed all baselines with an Fmax value of 0.67, which is a result of obtaining an excellent recall and a reasonable precision. GPT and LightGBM produced better recall values than RAMP, but lower precision, indicating a tendency to classify accurate sentences as containing ALTS more often. On the other hand, the Multinomial Naive Bayes achieved reasonable precision and recall, leading to an Fmax of 0.5. Finally, the open-source model Qwen, unsurprisingly, achieved mediocre performance.

Upon further investigation, we observed that RAMP misclassified 1 out of 8 true positives and 6 out of 205 true negatives in the test set. Our expert curators re-evaluated the RAMP's false positives with the highest predicted confidence scores. According to them, some of these sentences were initially assigned a negative label due to the ambiguity

of the language used by the student, but could have easily received a positive label. For example, in the sentence “an object with constant acceleration means it’s increasing constantly and an object at constant speed means it’s going at the same speed constantly”, it is unclear whether “it’s” refers to the object or the object’s speed, so it was labeled negative. This provides further confidence that reported performance is an underestimation of RAMP’s true classification power.

The words with higher *attention* in correctly classified positive examples seem to be strongly related to the core concepts of the positive class: (a) “constant acceleration”: is consistently and highly attended to, (b) “gaining speed”, “constantly gaining speed”, “speed is changing”, “increase in rate”, “move faster”, “velocity is continuously changing”: are phrases and words that describe the effect of constant acceleration on speed or velocity are also frequently attended to, and (c) “speed”, “velocity”: often receive significant attention.

Conclusions and Future Work

In this study, we explored the feasibility of detecting physics student misconceptions in written responses. We developed the RAMP classifier using written responses to a reading assignment from an introductory physics course that were annotated with a misconception related to motion. RAMP is powered by ModernBERT and is reasonably effective at distinguishing between accurate sentences and ones with misconceptions, as evidenced by its performance against several baselines.

However, there is room for improvement to make it accurate enough to serve as the back-end of an instructor-facing educational assessment tool for classifying and reporting student misconceptions across writing assignments in introductory physics courses. One such improvement is domain adaptation through unsupervised fine-tuning on domain-specific text corpora (e.g., physics textbooks). Comparing ModernBERT to other models pre-trained with science-specific text, such as SciBERT, will be worthwhile. Using

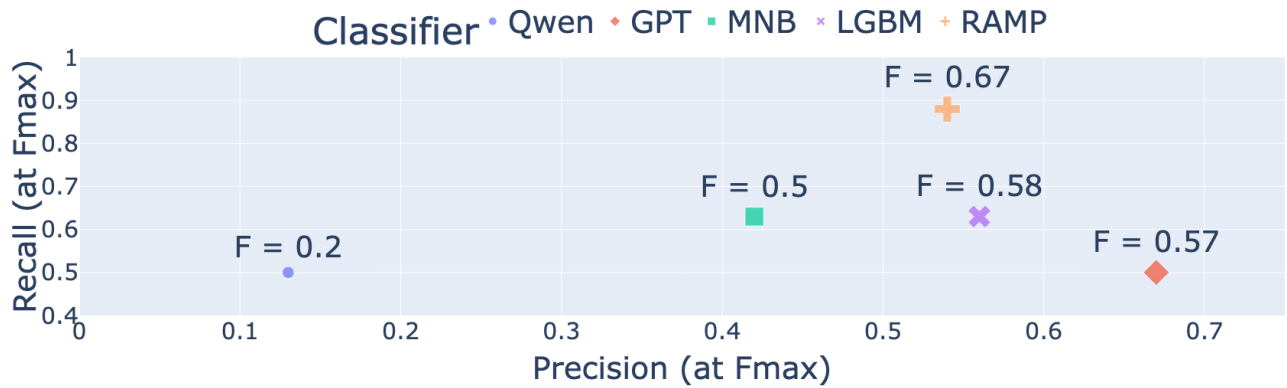


Figure 2: Overall Performance Comparison. Qwen: Qwen2.5(CoT), GPT: GPT5(CoT), MNB: Multinomial Naive Bayes, LGBM: LightGBM, and RAMP: our classifier powered by ModernBERT. Data labels indicate Fmax values.

student response data from assignments of varying types containing a range of misconception types would likely lead to better and robust models. Finally, we would like to develop a prototype web app for practical instructor classroom use, allowing for efficient assessment of conceptual understanding and feedback.

Limitations

While our study offers insights into utilizing LLMs for identifying misconceptions in student writing, it also has several limitations. This study relies on data from a single assignment and misconception type restricting the ability to draw broader conclusions. This narrow focus limits the generalization of our findings to other physics concepts or STEM disciplines, which may present different linguistic patterns and conceptual challenges.

The study focused on binary classification of sentences with misconceptions, neglecting a more nuanced assessment of conceptual understanding that might consider degrees of misconception or partial understanding. This simplification may not fully capture the complexity of student reasoning and the nuance between the response containing a misconception or not. This may have also been why RAMP performed better on sentences that directly addressed the question.

Similarly, the reliance on relatively simple prompts, as opposed to more advanced ones, such as tree-of-thought, and the utilization of only a handful of LLMs add to this limitation. Due to the existence of a very large number of LLMs and their exponential growth in recent years, our findings are likely not representative. The rapidly evolving landscape of LLMs means that newer models may offer improved performance.

Our approach only relied on text-based analysis, without incorporating other modalities of student work such as diagrams, equations, or problem-solving procedures that might provide additional context for understanding misconceptions. The dataset size was relatively small, which likely contributed to the modest Fmax of 0.67. While this perfor-

mance suggests potential for the approach, it remains below what would be considered reliable for independent classroom implementation without heavy instructor oversight. Another significant limitation is that the outputs of LLMs with prompting are non-deterministic. Although multiple runs for each experiment were performed to calculate means across trials, if not fully addressed, this behavior may hurt the intuition gained by instructors. It is also possible that SMOTE introduced biases or did not capture the diversity of student writing, potentially limiting the effectiveness of this approach.

Ethical and Broader Impact

The use of LLMs for identifying misconceptions could significantly reduce the instructor’s workload, allowing for a greater use of writing assignments and providing student feedback in a timely manner, even in large classes. This democratizes access to writing-based assessments across different class sizes and resource levels, potentially improving educational outcomes for a broader range of students.

Although our work has the potential to transform educational assessment practices, it also raises important concerns about fairness, privacy, and the appropriate role of technology in education. Since some studies have shown issues of bias and fairness with current large language models, there is a possibility that our current models may unfairly assess student responses from minority groups. Similarly, since LLMs are typically developed using primarily English data, there is a possibility that our models may unfairly assess responses from nonnative English-speaking students. In other words, model performance disparities could disadvantage certain student populations if the system misinterprets writing styles or expressions commonly used by specific demographic groups.

Since these systems require access to student work, another major concern is student privacy and data security. Clear policies regarding data retention, usage rights, and appropriate consent procedures are essential before implementing such technology on a large scale. There is also the

risk of over-reliance on automated assessment because human instructors bring contextual understanding and adaptability that current AI systems lack. It is essential to ensure that these tools are used only to supplement instructor-student interactions.

Finally, transparency about the use of AI in assessment is ethically necessary. Students should understand when their work is being evaluated by automated systems, how these systems function, and what recourse exists for contesting potentially erroneous classifications.

Acknowledgments

This work was supported by the 2024 Innovation in AI Grant, UNF. We would like to thank Ryan Nugent for preparing the original curated data for models.

References

- Arthurs, L.; Hsia, J. F.; and Schweinle, W. 2015. The oceanography concept inventory: A semicustomizable assessment for measuring student understanding of oceanography. *Journal of Geoscience Education* 63(4):310–322.
- Bai, J.; Bai, S.; Chu, Y.; Cui, Z.; Dang, K.; Deng, X.; Fan, Y.; Ge, W.; Han, X.; Huang, F.; et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Bernius, J. P.; Krusche, S.; and Bruegge, B. 2022. Machine learning based feedback on textual student answers in large courses. *Computers and Education: Artificial Intelligence* 3:100081.
- Carpenter, D.; Min, W.; Lee, S.; Ozogul, G.; Zheng, X.; and Lester, J. 2024a. 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)*, 403–413.
- Carpenter, D.; Min, W.; Lee, S.; Ozogul, G.; Zheng, X.; and Lester, J. 2024b. Assessing student explanations with large language models using fine-tuning and few-shot learning. In Kochmar, E.; Bexte, M.; Burstein, J.; Horbach, A.; Laarmann-Quante, R.; Tack, A.; Yaneva, V.; and Yuan, Z., eds., *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, 403–413. Mexico City, Mexico: Association for Computational Linguistics.
- Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16:321–357.
- Chen, Z., and Singh, C. 2025. Opportunities and challenges in harnessing digital technology for effective teaching and learning. *Trends in Higher Education* 4(1):6.
- Chen, Z., and Wan, T. 2025. Grading explanations of problem-solving process and generating feedback using large language models at human-level accuracy. *Physical Review Physics Education Research* 21(1):010126.
- Chen, C.; Sonnert, G.; Sadler, P. M.; Sasselov, D.; and Fredricks, C. 2020. The impact of student misconceptions on student persistence in a mooc. *Journal of Research in Science Teaching* 57(6):879–910.
- Ferreira Mello, R.; Pereira Junior, C.; Rodrigues, L.; Pereira, F. D.; Cabral, L.; Costa, N.; Ramalho, G.; and Gasevic, D. 2025. Automatic short answer grading in the llm era: Does gpt-4 with prompt engineering beat traditional models? In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, 93–103.
- Gomes, J. C., and Jaques, P. A. 2023. A machine learning approach for the identification of learners' misconceptions in algebraic problem-solving. In *2023 IEEE International Conference on Advanced Learning Technologies (ICALT)*, 221–225.
- Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems* 35:22199–22213.
- Kökver, Y.; Pektaş, H. M.; and Çelik, H. 2025. Artificial intelligence applications in education: Natural language processing in detecting misconceptions. *Education and Information Technologies* 30(3):3035–3066.
- Latif, E., and Zhai, X. 2025. Integrating generative ai into stem education: Enhancing conceptual understanding, addressing misconceptions, and assessing student acceptance. *Disciplinary and Interdisciplinary Science Education Research* 7(11).
- Neidorf, T.; Arora, A.; Erberber, E.; Tsokodayi, Y.; and Mai, T. 2019. Review of research into misconceptions and misunderstandings in physics and mathematics. In Neidorf, T.; Arora, A.; Erberber, E.; Tsokodayi, Y.; and Mai, T., eds., *Student Misconceptions and Errors in Physics and Mathematics*. Springer. 11–20.
- OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Resbiantoro, G.; Setiani, R.; et al. 2022. A review of misconception in physics: The diagnosis, causes, and remediation. *Journal of Turkish Science Education* 19(2):403–427.
- Savage, S., and Rebello, N. S. 2025. Using an llm to investigate students' explanations on conceptual physics questions. *arXiv preprint arXiv:2508.14823*.
- Sonkar, S.; Chen, X.; Liu, N.; Baraniuk, R. G.; and Sachan, M. 2024. Llm-based cognitive models of students with misconceptions. *arXiv preprint arXiv:2410.12294*.
- Warner, B.; Chaffin, A.; Clavié, B.; Weller, O.; Hallström, O.; Taghadouini, S.; Gallagher, A.; Biswas, R.; Ladhak, F.; Aarsen, T.; Cooper, N.; Adams, G.; Howard, J.; and Poli, I. 2024. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference.
- Weegar, R., and Idestam-Almquist, P. 2023. Reducing workload in short answer grading using machine learning. *International Journal of Artificial Intelligence in Education* 34.
- Wieman, C. E.; Adams, W. K.; and Perkins, K. K. 2008. Phet: Simulations that enhance learning. *Science* 322(5902):682–683.

Zarestky, J.; Bigler, M.; Brazile, M.; Lopes, T.; and Bangerth, W. 2022. Reflective writing supports metacognition and self-regulation in graduate computational science and engineering. *Computers and Education Open* 3:100085.

Zhou, L.; Kim, S.-M.; and Ahmed, N. 2024. Artificial intelligence applications in education: Natural language processing in detecting misconceptions. *Education and Information Technologies*.