Epistemic Safety in AI Physics Tutors: Evaluating Pedagogical Structure in Generated Explanations

Abdul Mohammad
University of California, Berkeley
abdulmohammad@berkeley.edu

Abstract

AI systems often answer introductory physics questions correctly, yet students need explanations that expose the reasoning behind the answer. We frame this as an epistemic safety issue for AI tutoring: a fluent explanation can invite learner trust while omitting the conceptual bridge needed to detect misunderstanding. We study whether AI-generated explanations contain pedagogical discourse moves such as framing the physical situation, applying principles, checking conclusions, giving intuition-building examples, and marking caveats. Using 60 explanations across 20 introductory physics prompts, we compare human-written and AI-generated responses with sentence-level labels, quality ratings, and a lightweight keyword baseline. The annotation scheme reaches Cohen's $\kappa = 0.6706$ on a 72-pair overlap subset. In this pilot, AI explanations match human explanations on correctness, show lower intuition-bridging coverage (20% vs 35%), and receive lower completeness ratings (4.0 vs 5.0). The findings suggest that AI physics tutor evaluation should treat explanation structure as a deployment check as well as a pedagogical preference.

1 Introduction

Evaluation at the intersection of physics and AI has advanced quickly, especially for answer-level reasoning tasks. Educational settings place an additional demand on these systems. A useful physics tutor needs to frame concepts, make reasoning legible, and connect formal equations to learner intuition. In practice, many LLM outputs are fluent and locally plausible while still being incomplete as teaching artifacts. This creates an epistemic safety concern: students can become confident in an explanation without having enough structure to notice what they do not understand.

Consider the question: why does an object moving in a circle accelerate even if its speed is constant? A thin answer may state that centripetal acceleration points inward. A stronger teaching explanation first addresses the misconception that constant speed implies no acceleration, explains that velocity includes direction, and then gives an intuitive example such as a car turning around a curve. Only the second exposes the reasoning a beginner needs to calibrate trust.

We treat this as a specific, limited form of AI safety for educational deployment. The risk is not physical harm or explicit hallucination. It is misplaced confidence caused by incomplete reasoning. When a response sounds coherent and authoritative, students may overtrust it even if key assumptions, caveats, or conceptual bridges are missing. The paper contributes:

- a compact 5-label discourse taxonomy with usable pilot reliability ($\kappa = 0.6706$ on 72 overlap pairs),
- a paired human and AI study showing lower AI intuition-bridging coverage at explanation level (-15 pp), and

Conference on Physics and AI at Stanford University (PAI 2026).

Category	Prompts	Human	AI
Mechanics	8	8	16
Electricity & Magnetism	8	8	16
Thermo/Energy	4	4	8
Total	20	20	40

Table 1: Pilot dataset composition.

Label	Plain meaning	Example function
`FRAME`	Sets up the situation	Constant speed differs from constant velocity.
`PRINCIPLE`	States or applies a rule	Acceleration is change in velocity over time.
`VERIFY`	Checks a conclusion	Without inward force, motion would be straight.
`INTUITION`	Gives an example or analogy	A car turning around a curve changes direction.
`CAVEAT`	Marks a condition or limit	This assumes air resistance is negligible.

Table 2: Plain-English label guide used to make the discourse scheme readable beyond NLP specialists.

- a lightweight keyword baseline for sentence labeling, included to anchor the benchmark as a reproducible evaluation task.

2 Related Work

Recent benchmarks such as PhysReason and UGPhysics provide broad coverage of physics reasoning at answer level [Zhang et al., 2025, Xu et al., 2025]. The present study targets a narrower educational and safety-relevant question: whether a correct explanation contains the discourse moves that support calibrated understanding. Methodologically, the annotation protocol follows established reliability guidance in computational linguistics [Cohen, 1960, Landis and Koch, 1977, Artstein and Poesio, 2008], adapted to a compact pedagogy-oriented taxonomy. The contribution is a discourse-aware evaluation layer for introductory physics explanation quality and for educational deployment checks.

3 Dataset and Annotation Protocol

3.1 Data Scope and Sources

The study uses 20 prompts and 60 explanations (1 human + 2 AI per prompt), distributed across mechanics (8 prompts), electricity and magnetism (8), and thermo/energy (4). Human references come from public introductory physics resources licensed for reuse and were normalized into a comparable explanatory format [OpenStax, 2026].

3.2 Annotation Schema

Each sentence receives exactly one of five labels (Table 2). During calibration, principle statements and direct applications were merged into a single `PRINCIPLE` category because annotators most often confused that boundary. Explanation-level ratings are 1–5 for clarity, correctness, and completeness.

3.3 Reliability Protocol

A 72-pair overlap subset was independently labeled by two annotators for calibration. Final agreement is $\kappa = 0.6706$ and $P_o = 0.8333$, which is suitable for pilot benchmark analysis with transparent limitations.

3.4 Heuristic Baseline

As a simple baseline, a rule-based classifier assigns sentence labels from cue phrases, such as analogy markers for `INTUITION` and qualification markers for `CAVEAT`. On 381 singly labeled sentences, this baseline reaches 70.6% accuracy and 0.575 macro-F1. Most errors collapse `VERIFY` and `CAVEAT` into `PRINCIPLE`, giving a lower-bound comparison for future automated labeling work.

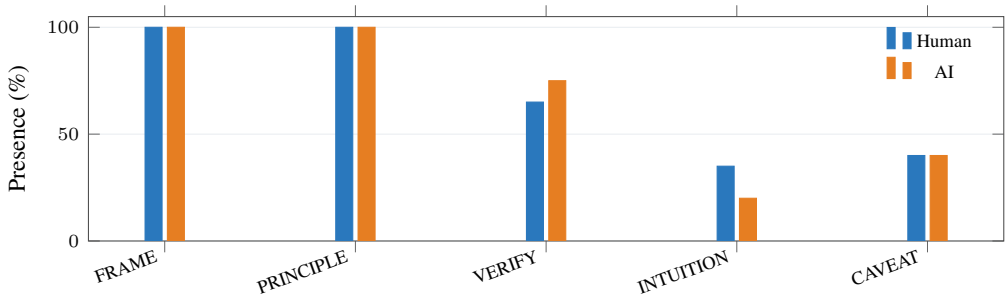

Figure 1: Explanation-level discourse move presence. AI responses show lower `INTUITION` prevalence despite matching human responses on `FRAME` and `PRINCIPLE`.

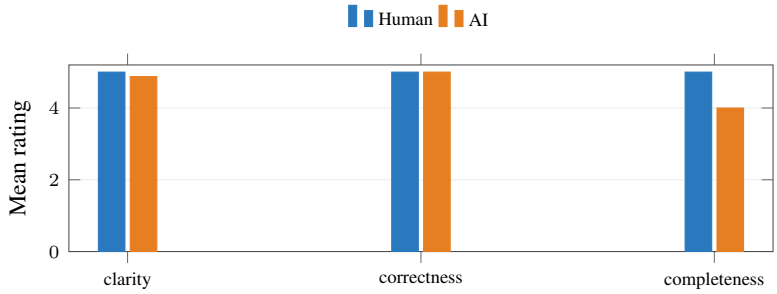

Figure 2: Holistic ratings by source. Correctness is saturated; completeness carries the strongest separation signal.

4 Results

4.1 Move Prevalence by Source

Figure 1 shows explanation-level move presence rates. Both sources are saturated on `FRAME` and `PRINCIPLE` (100%), then diverge on lower-frequency pedagogical moves. The strongest gap is `INTUITION` (35% human vs 20% AI), while AI shows higher `VERIFY` presence (75% vs 65%).

Sentence-level distributions are close on dominant structural content (`PRINCIPLE`: 55.37% human, 54.62% AI). The difference is concentrated in learner-facing bridges, with core inferential content remaining similar. This pattern is safety-relevant because introductory students often struggle to map a formal statement onto a physical situation. An explanation can be correct while still leaving that mapping implicit, creating a response that is trustworthy in surface form but weak as a guide to learner belief.

4.2 Holistic Ratings

Figure 2 and Table 3 summarize ratings from the primary annotator. Correctness is saturated (5.0 for both), while completeness shows the largest separation (5.0 human vs 4.0 AI).

4.3 Uncertainty

For key move contrasts, uncertainty remains wide at pilot scale: `INTUITION` difference is -15.0 pp (95% CI: $[-39.3, 9.3]$, $p = 0.206$) and `VERIFY` is $+10.0$ pp (95% CI: $[-14.8, 34.8]$, $p = 0.418$). Prompt-level leave-one-out checks preserve the negative INTUITION direction ($[-21.1, -10.5]$ pp), although multiple-comparison-adjusted significance is not reached. We report this as directional pilot evidence. In qualitative review, the circular-motion examples show the same pattern: AI responses identify inward acceleration, while the human explanation more directly repairs the constant-speed misconception by distinguishing speed from velocity.

Source	Clarity	Correctness	Completeness
Human ($n = 20$)	5.000	5.000	5.000
AI ($n = 40$)	4.875	5.000	4.000
AI - Human	-0.125	0.000	-1.000

Table 3: Mean holistic ratings.

5 Implications and Limitations

The practical implication is that discourse labels can become a diagnostic layer for AI tutoring systems. A tutor response could be checked for correctness and then inspected for missing pedagogical moves: no physical framing, no everyday example, no stated idealization, or no misconception repair. A revision system could then ask the model to add the missing move. In a deployed tutor, this could function as a low-cost epistemic safety guardrail. The goal is to reduce fluent explanations that invite overtrust while leaving key reasoning steps hidden.

This is a pilot study. The dataset is small, interval estimates for move prevalence are wide, and aggregate ratings rely on one primary annotator. The human explanations were normalized for comparability, which reduces stylistic noise but may compress natural human variation. Completeness ratings also show no within-group variance in this summary, so inferential claims about that metric should remain cautious. The safety framing is also limited: the study does not measure overtrust directly, does not test downstream learning, and does not evaluate high-stakes physical action. The next version should use a larger multi-rater corpus, unrevised human sources, and learner-outcome validation.

6 Conclusion

This pilot study contributes a compact protocol for evaluating the pedagogical structure of AI-generated physics explanations. The evidence suggests that educational AI systems should be evaluated for correctness and for the reasoning structure that helps learners understand, question, and calibrate trust in an explanation. In that sense, explanation structure is part of epistemic safety for AI tutors.

References

Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960.

J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

OpenStax. Openstax licensing and attribution information. `https://openstax.org/books/college-physics-2e/pages/6-2-centripetal-acceleration`, 2026.

Xin Xu, Qiyun Xu, Tong Xiao, Tianhao Chen, Yuchen Yan, Jiaxin Zhang, Shizhe Diao, Can Yang, and Yang Wang. Ugphysics: A comprehensive benchmark for undergraduate physics reasoning with large language models, 2025. arXiv:2502.00334.

Xinyu Zhang, Yuxuan Dong, Yanrui Wu, Jiaxing Huang, Chengyou Jia, Basura Fernando, Mike Zheng Shou, Lingling Zhang, and Jun Liu. Physreason: A comprehensive benchmark towards physics-based reasoning, 2025. arXiv:2502.12054.