Responsible NLP Checklist

Paper title: *LearnerCoMPASS: Intelligent Tutoring System with Dynamic Cognitive Diagnosis and Multi-Model Path Planning*
Authors: *Ziji Sheng, Guiyao Tie, Weidong Wang, Pan Zhou, Daizong Liu*

How to read the checklist symbols:

☑ the authors responded 'yes'

☒ the authors responded 'no'

N/A the authors indicated that the question does not apply to their work

☐ the authors did not respond to the checkbox question

For background on the checklist and guidance provided to the authors, see the Responsible NLP Checklist page at ACL Rolling Review.

☑ A. Questions mandatory for all submissions.

☑ A1. Did you describe the limitations of your work?
This paper has a Limitations section.

☑ A2. Did you discuss any potential risks of your work?
We explicitly discuss the risks associated with LLM hallucinations in high-stakes STEM educational domains in Section 1 (Introduction) and Section 3.1. Furthermore, Section 4.4.2 ("Impact on Reliability and Hallucination Mitigation") and Table 4 provide a detailed breakdown of hallucination risks and our mitigation strategies. We also address Content Safety and the filtering of potential biases or unsafe content in Appendix A.8.3 ("Intended Use and Compliance").

☑ B. Did you use or create scientific artifacts? (e.g. code, datasets, models)

☑ B4. Did you discuss the steps taken to check whether the data that was collected/used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect/anonymize it?
Appendix A.8.3 discusses privacy preservation, confirming that utilized datasets are anonymized and contain no PII. It also mentions content safety mechanisms to verify and filter potential biases or unsafe content. Appendix A.5.1 confirms no PII was collected from human evaluators.

☑ B6. Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?
Section 4.1 reports statistics such as the number of samples (200 resumes, 317 learning goals). Section 4.4.3 mentions the sample size (1,000) for hallucination analysis. Appendix A.8.1 estimates the total computational volume (approx. 85 million tokens).

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 describes the experimental setup and baselines. Appendix A.2 and Appendix A.3 detail the specific prompts and agent configurations (acting as hyperparameters for the agentic framework). Appendix A.8.1 mentions that the computational budget included hyperparameter tuning.

The Responsible NLP Checklist used at ACL Rolling Review is adopted from NAACL 2022, with the addition of ACL 2023 question on AI writing assistance and further refinements based on ARR practice. ACL 2026 used a subset of ARR checklist form.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 & Table 10 (in Appendix A.7) reports the Mean and Standard Deviation (Std. Dev) for human evaluation scores. Figure 4 uses violin plots to show the distribution, medians, and quartiles of the results, providing transparent descriptive statistics beyond single runs.

☑ **D. Did you use human annotators (e.g., crowdworkers) or research with human subjects?**

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix A.6.1 details the evaluation protocol provided to participants. Appendix A.6.2 and Table 9 present the full 5-point Likert Scale questionnaire and the specific criteria/rubric used for the evaluation.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Appendix A.6.1 ("Methodology" and "Data Usage and Consent") reports that "subject matter experts and students" were recruited and that "Participation was voluntary and compensated."

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?
Appendix A.6.1 ("Data Usage and Consent") explicitly states: "Prior to the evaluation, all participants provided informed consent." It also lists the specific information informed to them regarding anonymity and data usage.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The paper does not explicitly mention IRB approval. However, Appendix A.6.1 and Appendix A.8.3 outline ethical measures taken, including informed consent, anonymity, and strict data usage for research purposes only.

☑ **E. Did you use AI assistants (e.g., ChatGPT, Copilot) in your research, coding, or writing?**

☑ E1. If you used AI assistants, did you include information about their use?
We utilized AI assistants (e.g., ChatGPT-4o) specifically for grammatical error correction and polishing the writing style. All scientific claims, experimental designs, and logical arguments are the original work of the authors. And Section 4.1 describes the use of an "AI-Generated Skill Metadata Dataset". Section 4.2 details the use of "LLM-based Semantic Evaluation (using GPT-5.2... as the judge)". Appendix A.8.1 ("Model Deployments") lists all specific AI models (e.g., GPT-4o, Claude-3.7, Deepseek-R1) used as components or baselines in the research.