Responsible NLP Checklist

Paper title: *PED: Route-Decoupled Diagnostics for Persona Consistency in Spoken Agents*
Authors: *Weihao Liu, Junrui Wei, Zhao Zhang, Ju Zhang*

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?
Ethical Considerations (unnumbered section).

☑ **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?
No. We analyze model-generated text and speech only and do not collect real user conversations, personal identifiers, or recordings from real individuals; see Ethical Considerations.

☑ 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?
Yes. See Section 4.3 (Dialogue Runs and Logged Data), Section 5 (Results and Analysis), and Table 1. For each persona-system pair, we collect K=20 stateless anchor samples per route and one 25-turn stateful dialogue run under a Baseline/Stress/Recovery protocol.

☑ **C. Did you run computational experiments?**

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No. The paper describes the experimental setup and key controls (e.g., fixed decoding within each configuration, fixed speaker reference, and fixed dialogue protocol; see Sections 3.1, 4, and 4.4), but it does not report concrete decoding hyperparameter values such as temperature, top-p, or max tokens, nor any hyperparameter search.

☑ 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?
Yes. See Section 5 (Results and Analysis), including phase-wise TEC/AEC summaries, nearest-anchor accuracy and distributions, dominant-label rates, and the robustness checks with multi-seed reruns for Cascade and repeated E2E sanity-check runs.

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.

☑ **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.?
No. We report the audit setup and task in Section 5.5 (Blinded Human Audit), but we do not include the full text of the annotator instructions or screenshots in the paper.

☒ 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)?
No. We do not report recruitment or payment details in the paper. The human audit was a small-scale diagnostic sanity check on model-generated speech clips rather than a formal large-scale annotation study.

☒ 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)?
No. The human audit uses model-generated speech clips only and does not involve using or curating personal data from participants.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No. We do not report IRB/ethics-board approval. The human audit was a small-scale evaluation on model-generated speech clips only, with no real-user data collected.

☑ **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 used ChatGPT only for minor language polishing of a few paragraphs. No AI system was used to generate experimental data, run experiments, or produce or verify results.