Responsible NLP Checklist

Paper title: *HCSpec: Two-Tier Horizontal Cascade Speculative Decoding for High-Efficiency Large Language Model Inference*
Authors: *Yizhou Zhang, Siming Chen, Hao Ye, Erhu Feng*

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.

N/A A2. Did you discuss any potential risks of your work?
Our work focuses exclusively on accelerating large language models (LLMs) during inferenceit does not data collection, fairness considerations, human interaction, or any sociotechnical intervention. Consequently, there are no foreseeable ethical, safety, or societal risks associated with this research. Therefore, the Potential Risks item is N/A for this paper.

☑ 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?
All datasets used are well-established, community-vetted benchmarks widely adopted in top-tier NLP research. They have undergone rigorous curation: PII has been removed or never present; offensive or harmful content is explicitly excluded per their original documentation and preprocessing pipelines. Our work operates solely at the inference algorithm level and does not involve data collection, annotation, or generation thus posing no risk of reintroducing or amplifying such content.

☑ 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?
Appendix A (Related Datasets) reports key statistics for all datasets, including sample counts and structural properties.

☑ C. Did you run computational experiments?

☒ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 (Effectiveness of TT-HCSD) describes the experimental setup, but we did not perform hyperparameter search. Learning rates, the optimizer, and other key hyperparameters follow established defaults from prior work (i.e., EAGLE-3). Due to computational constraints, systematic tuning was not conducted.

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?
All reported results are from single runs. Due to computational constraints, we did not conduct multiple trials to compute error bars, standard deviations, or summary statistics.

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

N/A 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.?
(left blank)

N/A 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)?
(left blank)

N/A 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)?
(left blank)

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
(left blank)

☑ **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?
As stated in the ACL Policy on AI Writing Assistance (Section a), assistance limited to language polishing, such as grammar correction, rephrasing, or stylistic refinement of text originally authored by humans, does not require disclosure. We used LLMs solely for such proofreading-level editing (e.g., improving sentence fluency or concision) and did not use them for content generation, idea suggestion, technical explanation, or code writing. Therefore, disclosure is not required.