Responsible NLP Checklist

Paper title: *LongSpec: Long-Context Lossless Speculative Decoding with Efficient Drafting and Verification*

Authors: *Penghui Yang, Cunxiao Du, Fengzhuo Zhang, Haonan Wang, Tianyu Pang, Chao Du, Bo An*

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?
This work focuses on lossless inference acceleration for existing LLMs. Since LongSpec is provably lossless (the output distribution is identical to that of the target model), it does not introduce additional risks beyond those already present in the underlying target LLMs.

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

N/A 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?
We use only publicly available pretraining and benchmark datasets (SlimPajama-6B, Prolong-64k, LongBench, and standard math reasoning benchmarks), which have been released and vetted by their original creators. We did not collect any new data involving individuals, and our method does not introduce or amplify offensive content since it produces outputs identical in distribution to the target model.

☑ 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 and Appendix D describe the datasets used for training (SlimPajama-6B, Prolong-64k, and our long-context SFT subset) and evaluation (five LongBench subsets and four math reasoning benchmarks), along with batch size, training epochs, and the random offset ranges used per model.

☑ 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 target/draft model configurations, training stages, and benchmarks. Appendix D reports detailed hyperparameters including hardware (8 A100 80GB), optimizer (AdamW), learning rates (5e-4 for SlimPajama, 5e-6 for long-context data), batch sizes, learning rate schedule, ZeRO configuration, beam widths for tree decoding, and inference precision (float16).

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?
Table 1 reports average acceptance length and decoding speed (tokens/s) averaged across each evaluation dataset. Table 3 reports the same metrics for math reasoning benchmarks. Appendix F (Table 5) further reports a breakdown of latency and acceptance length across different prefill-length ranges.

☒ **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?**

N/A E1. If you used AI assistants, did you include information about their use?
(left blank)