Responsible NLP Checklist

Paper title: *FastKV: Decoupling of Context Reduction and KV Cache Compression for Prefill-Decoding Acceleration*

Authors: *Dongwon Jo, Jiwon Song, Yulhwa Kim, Jae-Joon Kim*

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 improving the computational efficiency of LLM inference and does not introduce new capabilities or application domains. The method operates purely at the architectural level (KV cache compression and token selection during prefill), and does not raise potential risks beyond those already associated with large language models in general.

☑ 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 publicly available benchmarks (LongBench, RULER, Needle-in-a-Haystack) that do not contain personally identifying information or offensive 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?
Dataset details including the number of samples, evaluation metrics, task types, and language information are reported in Appendix C.2

☑ C. Did you run computational experiments?

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Baseline configurations are described in Section 5.1. Additional hyperparameter sweeps are provided in Appendix D.5.

☑ 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 averaged over fixed evaluation sets as defined by each benchmark (LongBench, RULER, Needle-in-a-Haystack). Runtime measurements in Appendix D.4 (Table 8) include mean and standard deviation across repeated 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?**

☒ 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.?
No human annotators or participants were involved in this study.

☒ 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)?
No human annotators or participants were involved in this study.

☒ 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)?
Only publicly available benchmark datasets were used.

☒ N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No human subjects research was conducted. Ethics review board approval was not required.

☑ **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?
AI writing assistants were used for grammar checking.