Responsible NLP Checklist

Paper title: *Jakiro: Boosting Speculative Decoding via Decoupled MoE*
Authors: *Haiduo Huang, Fuwei Yang, Zhenhua Liu, Pengju Ren*

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?
Section Limitations. Jakiro is a lossless inference-acceleration technique that preserves the base LLM's output distribution, so it neither amplifies nor introduces new risks beyond those of the underlying model. Potential misuse (e.g., faster generation of harmful content) inherits the risk profile of the base LLM and is discussed in the Limitations section.

☑ **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?
N/A. We do not collect or curate new data. All used datasets (MT-bench, HumanEval, GSM8K, Alpaca, CNN/DM, Natural Questions) and training corpus (ShareGPT, UltraChat-200K) are publicly released benchmarks processed and released by their original authors, who handled PII/offensive-content filtering upstream.

☑ 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 (Experimental Setup) reports training-data size (ShareGPT 68K / ShareGPT+UltraChat-200K ~530K) and lists all six evaluation benchmarks. Per-benchmark acceptance length (tau) and speedup (S) are reported in Tables in Section 4 and Appendix A.4.

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

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4.1 and Appendix A.4 describe the setup (models, GPUs: A100-40G / MI250-64G, training hyperparameters). Appendix A.4.7 (Hyperparameter Sensitivity) reports the sweep; Section 4.4 (Ablation Study, Table tab:aba_combined) reports N-K search and best-found values (N=K=2).

☑ 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 speedup and acceptance-length numbers (Sections 4.2-4.4, Appendix A.4) are mean values averaged over four inference runs, as stated in the figure captions (e.g., Fig. fig:MTbench_L_Comparison_T=1). Temperature T=0 and T=1 results are reported separately.

☒ **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.?
N/A. No human subjects or annotators were involved.

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)?
N/A. No human subjects or annotators were involved.

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)?
N/A. We use publicly released datasets; we did not collect data from individuals.

N/A D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
N/A. No data collection from human subjects was performed.

☑ **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?
AI assistants (ChatGPT/Claude) were used for minor language polishing and LaTeX formatting assistance only. All technical content, experimental design, code, and results are the authors' original work. No AI was used to generate experimental data or scientific claims.