

Responsible NLP Checklist

Paper title: *AscendKernelGen: LLM-Driven Kernel Generation for NPUs*

Authors: *Xinzi Cao, Jianyang Zhai, Pengfei Li, Zhiheng Hu, CenYan, mubingxu, Guanghuan Fang, Bin She, Jiayu Li, Yihan Su, Dongyang Tao, Feidiao Yang, Chang-Dong Wang, Yutong Lu, Weicheng Xue, Bin Zhou, Yonghong Tian*

How to read the checklist symbols:

- the authors responded 'yes'
- the authors responded 'no'
- 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 low-level NPU kernel generation and evaluation for system and hardware research purposes. It does not involve human subjects, user-facing applications, personal data, or natural language content with social impact. Therefore, we do not identify specific ethical or societal risks associated with this work.

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?

The data used in this work consists of system-level kernel code, compiler error logs, and documentation-derived or synthetic reasoning traces. It does not contain natural language data about individuals, user-generated content, or any information that names or uniquely identifies people. Therefore, no explicit anonymization or offensive content filtering procedures were required or discussed.

- 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?

We provide detailed experimental configurations in Appendix B, including training hyperparameters such as learning rates, batch sizes, and optimization settings. Appendix E describes the construction of the Ascend-CoT dataset, including approximately 7k high-quality kernel-level chain-of-thought samples. The full Ascend-CoT dataset is used for supervised fine-tuning, while evaluation is conducted exclusively on the separate NPUIKernelBench benchmark to avoid overlap between training and testing. In addition, summary statistics of the Ascend-CoT dataset are reported in Section 4.1.

C. Did you run computational experiments?

- C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?

Section 5.1 and Appendix B

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?

Yes, please see Sec 5.1. Each configuration was evaluated at least three times on real NPU hardware under identical settings, and the reported results are the mean values across runs. We make this aggregation protocol explicit in the paper. Across repeated trials, compilation and correctness outcomes were consistent, while runtime-related metrics showed only minor variation, supporting the stability and reproducibility of the reported results.

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.?

This study does not involve human participants or annotators. All datasets are constructed from publicly available technical documentation and code, and all experiments are conducted through automated model training and evaluation.

- 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)?

This study does not involve any human participants or annotators. All contributors are members of the research team, and no recruitment or payment of participants was conducted.

- 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)?

This work does not use or curate data originating from identifiable individuals. All datasets are constructed from publicly available technical documentation, official API references, and programmatic kernel implementations. Therefore, no human data or consent is involved.

- D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?

This work does not involve human subjects, user studies, or the collection of personal data. All datasets are derived from publicly available technical documentation and programmatic code artifacts. Therefore, ethics review board approval is not applicable.

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 were used solely for language polishing and improving clarity of presentation. They did not contribute to the research design, experiments, analysis, or conclusions, and therefore were not explicitly disclosed in the manuscript.