

Responsible NLP Checklist

Paper title: *InstructDiff: Domain-Adaptive Data Selection via Contrastive Entropy for Efficient LLM Fine-Tuning*

Authors: *Junyou Su, He Zhu, Xiao Luo, Liyu Zhang, Hong-Yu Zhou, Yun Chen, Peng Li, Yang Liu, Guanhua Chen*

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?

The potential risks of the work are discussed in the "Limitations" section. This section details three key limitations: domain-specific tuning requirements for hyperparameters (and), dependency on warmup set size for stable entropy difference estimates, and model family specificity that affects selection consistency across different model architectures.

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 (NuminaMath, Alpaca, MedCAQA, BigCode) are publicly available research resources that have been vetted by their original creators for personally identifying information and offensive content. We did not perform additional data collection or annotation. These datasets are widely used in the research community and comply with standard data ethics practices.

- 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 and Appendix E. We report detailed statistics including diversity metrics, complexity metrics, PPL, entropy, and train/test splits.

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 C. We provide complete hyperparameter settings including learning rates, batch sizes, world model training configurations, and sampling strategies.

- 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,

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.

etc. or just a single run?

Section 4.3 and Section 5, We do many ablation studies and analysis

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.?

The paper focuses on methodological innovation of the InstructDiff framework for LLM data selection, without involving human participants or annotators. Thus, there are no instructions to participants to report.

- 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)?

The paper focuses on proposing the InstructDiff framework for LLM data selection, using only publicly available datasets and no human participants. Thus, there is no participant recruitment or payment to report.

- 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)?

The paper focuses on the methodological innovation of the InstructDiff framework for LLM data selection, using only publicly available research datasets. It does not discuss data consent from individuals whose information is included in the datasets, as consent acquisition is the responsibility of the original dataset creators.

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

The work only uses publicly available research datasets without conducting any new data collection. Thus, 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?

We used Claude/GPT-4 for literature review, code debugging, and initial draft proofreading. All core ideas, experiments, and final writing are original author contributions.