Neuron-Aware Data Selection In Instruction Tuning For Large Language Models

Xin Chen♤♡☆, Junchao Wu♤, Shu Yang♤, Runzhe Zhan♤, Zeyu Wu♤, Min Yang♡◊, Shujian Huang☆, Lidia S. Chao♤, Derek F. Wong♤

♤ NLP2CT Lab, Department of Computer and Information Science, University of Macau; ♡ Artificial Intelligence Research Institute, Shenzhen University of Advanced Technology
◊ Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences; ♤ Provable Responsible AI and Data Analytics Lab, KAUST
☆ National Key Laboratory for Novel Software Technology, Nanjing University

ICLR

Motivation & Problem

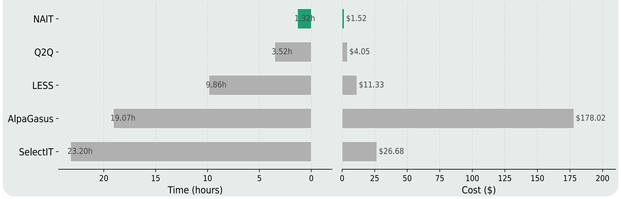
- Excessive IT data degrades LLM performance, while carefully selected small subsets can significantly enhance capabilities
- Existing IT selection methods rely on external models, surface-level features, or costly APIs, limiting scalability and targeted application
- Current methods lack interpretability and the ability to enhance specific target domain capabilities

Key Contributions

- First IT data selection framework based on neuron activation patterns, introducing a new paradigm for targeted development of model capabilities
- Reveals the correlation between neuron activation features and fundamental model capabilities, and their cross-task transferability
- Open-sources a cross-task neuron feature library and the Alpaca-NAIT dataset

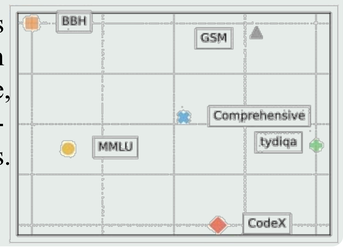
Cost Efficiency

- NAIT: 1.32h / $1.52 only, up to 19× cheaper than AlpaGasus

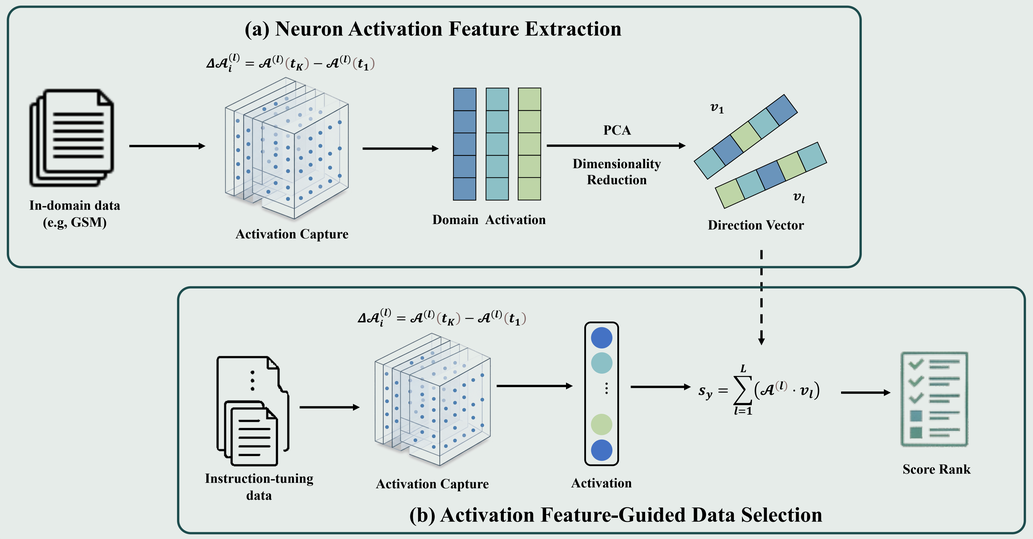


Direction of Activation Features

- Activation features form clusters in representation space, consistent with task-specific mechanisms.



Methodology — NAIT

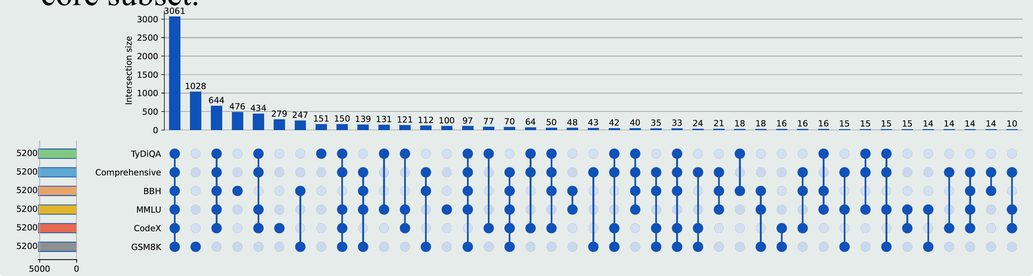


Method Comparison

Method	Externally-Independent	Gradient-Free	Targeted Ability	Cost-Effective	Interpretability
LIMA (Zhou et al., 2023)	✔	✔	✗	✗	✗
Instruction Mining (Cao et al., 2023)	✔	✔	✗	✗	○
AlpaGasus (Chen et al., 2024)	✗	✔	✗	✗	✗
SelectIT (Liu et al., 2024)	✔	✔	✗	✗	○
LESS (Xia et al., 2024)	✔	✗	✔	✗	○
NAIT (Ours)	✔	✔	✔	✔	✔

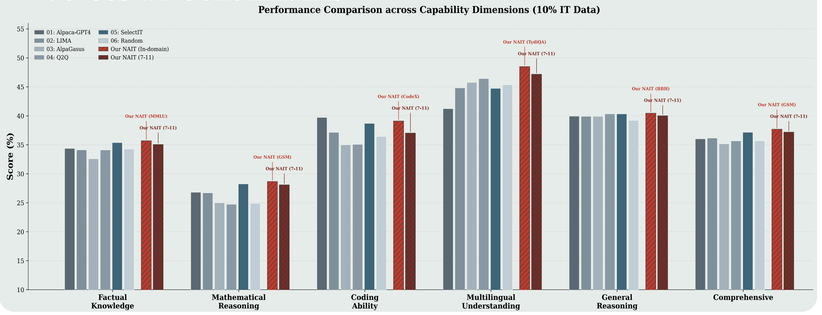
Data Distribution

- 58.87% of selected data overlaps across tasks, forming a stable general core subset.



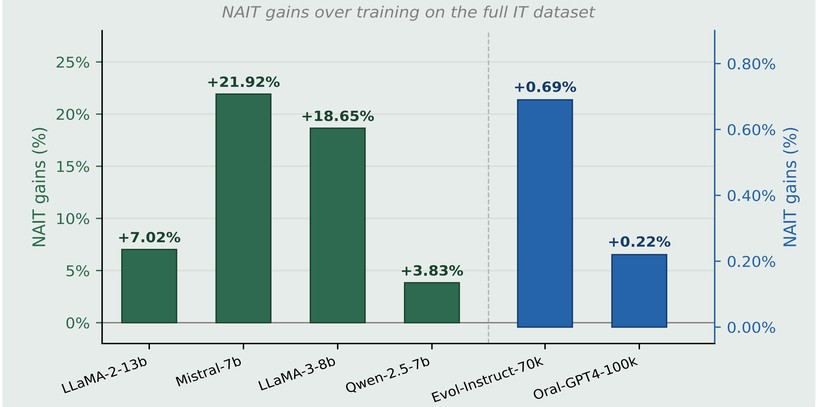
Main Result

- NAIT achieves the best or near-best performance across all benchmarks



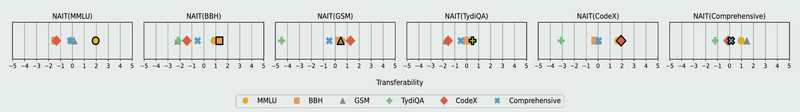
Robustness

- NAIT demonstrates strong generalizability across both models and datasets



Interpretability & Transferability

- GSM and CodeX features show strong cross-task positive transfer; TydiQA features show weaker transferability.



Scan for more!!