XXXXXX

Ye Cao
ServiceNow
Santa Clara, USA
mr.ye.cao.yc@gmail.com

Zhaozhuo Xu
Stevens Insitute of Technology
Hoboken, USA
zxu79@stevens.edu

***Abstract*—**
***Index Terms*—Large language model, data attribution**

I. Introduction

A. What is the problem?

B. Why is it interesting and important?

C. Why is it hard? (E.g., why do naive approaches fail?)

D. Why hasn't it been solved before?

E. What are the key components of my approach and results?

II. Problem Statement

A. Behavior-Aware Data Attribution for Biomedical LLMs

B. Quantitative Evaluation of Impacts from Pre-Training or Fine-Tuning Data

III. LARK: Linearized Attribution Kernel for Behavior-Aware Data Attribution

IV. Experiment

- **RQ1:**
- **RQ2:**

A. Settings

Dataset. We used the GSM8k dataset [1], a benchmark that contains a bunch of math-related word problems. Most of the problems require multi-step arithmetic reasoning. We mostly focused on the "test" split from the "main" configuration of the dataset. This provides clean, curated problems with step-by-step answers for evaluation. The GSM8k dataset is widely used for evaluating the performance of a model. This research builds the foundation of meta-questions to probe the model's understanding of compression.

Testbed. All experiments were performed on a single workstation equipped with an NVIDIA GPU. In addition, we created and ran code on Google Colab. The code dynamically selects the computation device based on availability. Typically, it uses CUDA when a GPU is present or defaults to CPU otherwise. For our results, we used a CUDA GPU, which enables efficient inference with the Qwen2.5-0.5 B-Instruct model [2]. The models and tokenizers were loaded from the Hugging Face Transformers library with torch_dtype "auto" to optimize memory usage during inference.

Evaluation Metric. We obtain the ground truth regarding the compression states of LLMs. Given a probe query, if the LLM responses with the answer that aligns with its ground truth compression states, we mark it as correct. Then we report the average accuracy over the all queries.

B. Answers to RQ1:

C. Answers to RQ2: .

V. Related Works

Recent advances in model compression have prioritized making LLMs more accessible to resource-constrained environments through techniques such as quantization [3]–[5] and sparsification [6]–[8]. However, these methods often induce performance degradation, especially on complex reasoning tasks. [9] introduced the concept of compression-aware computing, advocating for LLM systems that can recognize and adapt to their compressed states. This framework emphasizes the integration of compression techniques with model introspection to preserve performance while enhancing efficiency—particularly relevant for domains requiring scalable and sustainable AI, such as biomedical research [10].

Complementing this direction, [11] proposed a surprisingly simple yet effective strategy: using soft prompt tuning to recover performance in heavily compressed LLMs. Their findings demonstrate that learned prompts can significantly restore task performance across quantized and pruned models. More importantly, these prompts exhibit transferability across tasks, compression levels, and even model architectures—subverting the conventional belief that prompt tuning is inherently task-specific. This line of work suggests that LLMs, though altered by compression, retain latent capacities that can be accessed or enhanced through targeted prompting.

VI. Conclusion

This work initiates a study on the self-awareness of LLMs in the context of weight quantization. We propose a probing methodology that combines arithmetic reasoning with introspective querying to assess whether compressed LLMs can identify their own compression state. Our results show that several models can infer their quantized status with notable accuracy, especially when guided by in-context examples. These findings point toward the feasibility of embedding introspection capabilities in future development and have implications for building more interpretable AI systems.

REFERENCES

[1] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano *et al.*, "Training verifiers to solve math word problems," *arXiv preprint arXiv:2110.14168*, 2021.

[2] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[3] Z. Liu, J. Yuan, H. Jin, S. Zhong, Z. Xu, V. Braverman, B. Chen, and X. Hu, "KIVI: A tuning-free asymmetric 2bit quantization for KV cache," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=L057s2Rq8O

[4] T. Zhang, J. W. Yi, B. Yao, Z. Xu, and A. Shrivastava, "Nomad-attention: Efficient llm inference on cpus through multiply-add-free attention," *Advances in Neural Information Processing Systems*, 2024.

[5] T. Zhang, J. Yi, Z. Xu, and A. Shrivastava, "Kv cache is 1 bit per channel: Efficient large language model inference with coupled quantization," *Advances in Neural Information Processing Systems*, 2024.

[6] Y. Zhou, Z. Chen, Z. Xu, V. Lin, and B. Chen, "Sirius: Contextual sparsity with correction for efficient llms," *Advances in Neural Information Processing Systems*, 2024.

[7] W. Guo, J. Long, Y. Zeng, Z. Liu, X. Yang, Y. Ran, J. R. Gardner, O. Bastani, C. D. Sa, X. Yu, B. Chen, and Z. Xu, "Zeroth-order fine-tuning of llms with transferable static sparsity," in *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. [Online]. Available: https://openreview.net/forum?id=myYzr50xBh

[8] Y. Wu, W. Guo, Z. Liu, H. Ji, Z. Xu, and D. Zhang, "How large language models encode theory-of-mind: a study on sparse parameter patterns," *npj Artificial Intelligence*, vol. 1, no. 1, p. 20, Aug 2025. [Online]. Available: https://doi.org/10.1038/s44387-025-00031-9

[9] Z. Xu, "Compression-aware computing for scalable and sustainable AI," in *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, T. Walsh, J. Shah, and Z. Kolter, Eds. AAAI Press, 2025, p. 28733. [Online]. Available: https://doi.org/10.1609/aaai.v39i27.35126

[10] Y. Wu, Y. Yang, Z. Liu, Z. Li, K. Pahwa, R. Li, W. Zheng, X. Hu, and Z. Xu, "Weighted diversified sampling for efficient data-driven single-cell gene-gene interaction discovery," *arXiv preprint arXiv:2410.15616*, 2024.

[11] Z. Xu, Z. Liu, B. Chen, S. Zhong, Y. Tang, J. Wang, K. Zhou, X. Hu, and A. Shrivastava, "Soft prompt recovers compressed llms, transferably," in *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. [Online]. Available: https://openreview.net/forum?id=muBJPCIqZT