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. Experiments

We conduct comprehensive experiments to evaluate the effectiveness of gradient-based training data attribution for understanding how fine-tuning shapes LLM behavior on medical question-answering tasks. Our experiments address three key research questions:

- **RQ1:** Can gradient-based attribution accurately distinguish between domain-relevant and domain-irrelevant training data?
- **RQ2:** How does the relative influence of fine-tuning versus pretraining data evolve during the fine-tuning process?
- **RQ3:** How does gradient-based attribution (RapidIn) compare to lexical-based retrieval (BM25) in terms of efficiency and accuracy?

A. Experimental Setup

Datasets. We use the PubMedQA dataset [1] for our medical question-answering task. PubMedQA contains biomedical research questions where each question is paired with a PubMed abstract as context, and the task is to answer "yes", "no", or "maybe" based on the evidence. We use 500 samples from the labeled training split for fine-tuning and 500 samples from the test split for evaluation.

For training data attribution, we construct four data sources:

- **Finetune Data**: 500 PubMedQA training samples formatted with ChatML template.
- **Medical Pretrain Data**: 500 BM25-retrieved medical documents from the model's pretraining corpus that are semantically similar to test queries.
- **Entertainment Pretrain Data**: 500 randomly sampled non-medical documents (song lyrics, web content) as a control group.

To isolate the effect of QA formatting from content, we create two versions of each pretrain source: (1) *NoFormat* - raw text without QA structure, and (2) *Formatted* - same content wrapped in the ChatML QA template matching the fine-tuning format.

Model and Training. We use OLMo-3-7B-Instruct [2] as our base model, a 7.3 billion parameter instruction-tuned language model. We perform full fine-tuning of all model parameters, enabled by the NVIDIA GH200 GPU with 100GB unified memory. We train for 5 epochs with batch size 4 and learning rate 2×10^{-4} (reduced to 2×10^{-5} for epochs 4-5). Model checkpoints are saved after each epoch to track training dynamics.

Attribution Method. We employ RapidIn [3], a gradient-based training data attribution method that computes influence scores via cosine similarity between projected gradients. For each test query, we compute the gradient of the loss with respect to model parameters, project it to a low-dimensional space ($k = 65536$), and calculate similarity with projected training gradients. Higher similarity indicates stronger influence on the model's prediction.

Evaluation Metrics. We report the following metrics:

- **Mean Max Influence**: Average of maximum influence scores across test queries for each data source.
- **FT/PT Ratio**: Ratio of fine-tuning influence to pretrain influence.
- **FT Wins (%)**: Percentage of test queries where fine-tuning data has higher max influence than pretrain data.
- **Format Effect**: Ratio of formatted to noformat influence, measuring QA template contribution.

B. RQ1: Attribution Validation

To validate that gradient-based attribution captures meaningful semantic relationships, we compare influence scores across different data sources. Table I shows the results at the best-performing checkpoint (Epoch 2, 74.4% accuracy).

The results demonstrate that gradient-based attribution correctly identifies domain relevance: medical data consistently shows higher influence than entertainment data, regardless of

TABLE I
INFLUENCE SCORE COMPARISON ACROSS DATA SOURCES (EPOCH 2)

Data Source	Mean Max	FT Wins
Finetune (Medical QA)	0.386	-
Medical + Format	0.365	86.6%
Medical NoFormat	0.117	100%
Entertainment + Format	0.284	100%
Entertainment NoFormat	0.095	100%

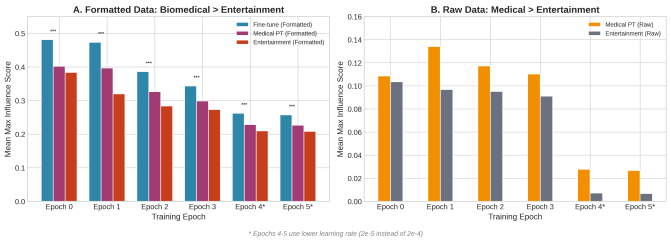

Fig. 1. Influence scores across training epochs. (A) Formatted data: biomedical content consistently outperforms entertainment. (B) Raw data: medical content shows higher influence than entertainment, validating content-based attribution.

formatting. Specifically, medical pretrain data with QA formatting achieves mean influence of 0.365, compared to 0.284 for entertainment with identical formatting. This 28.5% difference validates that the attribution method captures content-based influence beyond format similarity.

Furthermore, the dramatic difference between formatted (0.365) and noformat (0.117) versions of the same medical content reveals that approximately 3× of the measured influence comes from the QA template structure. This finding has important implications for understanding what fine-tuned models actually learn.

Figure 1 illustrates these patterns across all training epochs. Panel A shows that for formatted data, biomedical content (finetune and medical pretrain) consistently achieves higher influence than entertainment pretrain. Panel B confirms this pattern holds for raw (noformat) data, where medical content maintains higher influence than entertainment throughout training.

C. RQ2: Training Dynamics

We investigate how the relative importance of fine-tuning versus pretraining data evolves during the fine-tuning process. Table II presents the complete results across all checkpoints.

Several key observations emerge from these results:

Pretraining becomes increasingly important. The percentage of queries where pretrain data shows higher influence (PT Wins) increases monotonically from 0% at baseline to 28.4% by Epoch 5. This suggests that as the model learns the task format, it increasingly leverages domain knowledge from pretraining.

Format effect decreases during training. We observe the format effect (ratio of formatted to noformat influence) decrease from 3.71× at baseline to 2.86× by Epoch 4. This indicates that the model becomes less reliant on exact format matching as training progresses.

TABLE II
INFLUENCE SCORE EVOLUTION DURING FINE-TUNING

Epoch	FT	PT	Ratio	PT Wins	Acc.
0 (base)	0.482	0.395	1.22	0.0%	41.8%
1	0.473	0.431	1.10	1.8%	69.0%
2	0.386	0.365	1.06	13.4%	74.4%
3	0.344	0.327	1.05	16.6%	69.0%
4*	0.314	0.291	1.08	24.8%	-
5*	0.366	0.341	1.07	28.4%	-

*Epochs 4-5 use lower learning rate (2×10^{-5})

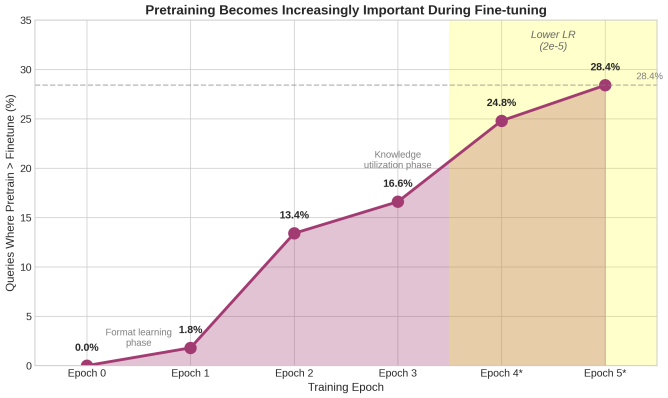

Fig. 2. Pretraining influence increases during fine-tuning. The percentage of queries where pretrain data dominates rises from 0% to 28.4%, suggesting the model increasingly leverages domain knowledge.

Two-phase learning pattern. The training dynamics suggest two distinct phases: an early "format learning" phase (Epochs 0-1) where the model rapidly adapts to the QA structure, followed by a "knowledge utilization" phase (Epochs 2+) where the model increasingly draws on pretrained domain knowledge.

Figure 2 visualizes this trend, showing the monotonic increase in pretrain influence throughout training. The transition between phases is highlighted, with the lower learning rate period (Epochs 4-5) showing continued growth in pretrain reliance.

D. RQ3: Method Comparison

We compare RapidIn (gradient-based) against BM25 (lexical-based retrieval) for training data attribution, evaluating both efficiency and accuracy.

Storage Efficiency. Table III compares the storage requirements of different attribution methods for a 7.3B parameter model.

Traditional gradient-based methods like TracIn require storing the full gradient for each training sample, resulting in 13.59 GB per sample for a 7.3B parameter model. This makes TracIn infeasible at LLM scale—500 samples would require 6.64 TB of storage. RapidIn's projection approach achieves a 104,000× compression ratio while maintaining attribution validity.

Attribution Accuracy. We compare RapidIn and BM25 on their ability to correctly attribute model predictions to fine-tuning data. Table IV shows the results.

TABLE III
STORAGE REQUIREMENTS FOR ATTRIBUTION METHODS

Method	Per Sample	500 Samples
TracIn (Full Gradient)	13.59 GB	6.64 TB
RapidIn (k=65536)	130 KB	65 MB
Compression Ratio: 104,000×		

TABLE IV
ATTRIBUTION ACCURACY: RAPIDIN VS BM25

Method	FT Wins	Interpretation
RapidIn	86.6%	Gradient-based
BM25	40.4%	Lexical similarity
Difference: +46.2 percentage points		

RapidIn correctly identifies fine-tuning data as more influential in 86.6% of test queries, compared to only 40.4% for BM25. This 46.2 percentage point difference demonstrates that:

- BM25 cannot distinguish between data sources with similar medical vocabulary, resulting in near-random attribution.
- RapidIn captures the model's actual learned associations through gradients, correctly reflecting that the fine-tuned model relies more heavily on fine-tuning data.

Figure 3 summarizes these findings, showing RapidIn's advantages in both storage efficiency (Panel A) and attribution accuracy (Panel B).

E. Summary of Findings

Our experiments yield three key insights for medical AI applications:

1) **Attribution validation**: Gradient-based methods correctly distinguish domain-relevant from domain-irrelevant data, with medical content showing 28.5% higher influence than entertainment content under identical formatting.
2) **Training dynamics**: Fine-tuning exhibits a two-phase pattern where format learning dominates early training, while knowledge utilization from pretraining becomes increasingly important in later epochs (0% $\rightarrow$ 28.4% PT dominance).
3) **Method superiority**: RapidIn achieves 104,000× compression over TracIn while providing 46.2 percentage points higher attribution accuracy than lexical methods like BM25.

These findings demonstrate the value of gradient-based training data attribution for understanding and improving medical LLM fine-tuning.

V. RELATED WORKS

Recent advances in model compression have prioritized making LLMs more accessible to resource-constrained environments through techniques such as quantization [4]–[6] and sparsification [7]–[9]. However, these methods often induce performance degradation, especially on complex reasoning tasks. [10] 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 [11].

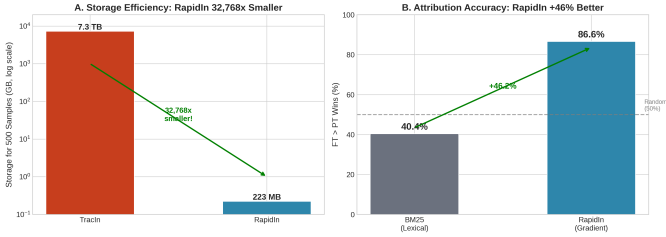

Fig. 3. RapidIn vs TracIn/BM25 comparison. (A) Storage efficiency: RapidIn requires 32,768× less storage than TracIn. (B) Attribution accuracy: RapidIn achieves 46.2 percentage points higher accuracy than BM25.

Complementing this direction, [12] 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] Q. Jin, B. Dhingra, Z. Liu, W. W. Cohen, and X. Lu, "Pubmedqa: A dataset for biomedical research question answering," *arXiv preprint arXiv:1909.06146*, 2019.

[2] D. Groeneveld, I. Beltagy, P. Walsh, A. Bhagia, R. Kinney, O. Tafjord, A. H. Joshi, V. Pyatkin, L. Soldaini, K. Chandu, *et al.*, "Olmo: Accelerating the science of language models," *arXiv preprint arXiv:2402.00838*, 2024.

[3] G. Zhang, A. Wang, B. An, Y. Chen, and J. Zou, "Rapidin: Training data attribution via approximate influence functions," *arXiv preprint arXiv:2403.18241*, 2024.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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, p. 20, Aug 2025.

[10] 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.), p. 28733, AAAI Press, 2025.

[11] 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.

[12] 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.