Scalable Data Attribution Reveals Fine-Tuning Unlocks LLM Pretrained Knowledge in Biomedical Question Answering

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

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

***Abstract*—Fine-tuning large language models (LLMs) on biomedical question answering tasks yields strong performance, but the relative contributions of pretrained knowledge versus fine-tuning data remain opaque. Understanding this knowledge provenance is essential for trustworthy medical AI deployment. We employ LARK (Linearized Attribution via Rapid Kernel), a gradient-based training data attribution method that approximates influence functions using identity Hessian assumption and OPORP dimensionality reduction, achieving 54,000× storage compression for 7B-parameter models. Applying this method to OLMo-3-7B fine-tuned on PubMedQA, we discover a "knowledge awakening" phenomenon: while fine-tuning data dominates early training, pretrained knowledge influence increases monotonically throughout fine-tuning (0% to 28.4% pretraining dominance). Gradient-based attribution correctly distinguishes domain-relevant from irrelevant data, with medical content showing 28.5% higher influence than entertainment content, and outperforms lexical methods (BM25) by 46.2 percentage points in attribution accuracy. Our findings suggest that fine-tuning serves as a key that progressively unlocks pretrained domain knowledge rather than simply imprinting new patterns.**

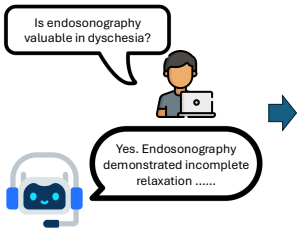
***Index Terms*—Large language model, training data attribution, influence function, biomedical question answering, fine-tuning dynamics**

I. Introduction

Large language models (LLMs) have achieved remarkable success in biomedical question answering, yet a fundamental question remains unanswered: *What knowledge source (pretraining or fine-tuning data) drives the model's predictions?* This transparency is critical for deploying trustworthy AI in healthcare, where clinical accountability demands understanding the provenance of model outputs.

When fine-tuning an LLM on biomedical QA tasks like PubMedQA, the model draws from two knowledge sources: (1) the vast pretrained corpus containing general and medical knowledge, and (2) task-specific fine-tuning examples. Current practices treat these models as black boxes, offering no visibility into which source influences specific predictions. This opacity hinders bias detection, model debugging, and regulatory compliance in medical AI systems.

Training data attribution methods aim to address this challenge by quantifying each training sample's influence on model predictions. However, existing approaches face critical limitations at LLM scale. Traditional influence functions [1] require inverting the Hessian matrix, which is computationally infeasible for billion-parameter models. Full gradient storage demands terabytes per dataset. Meanwhile, lexical methods like BM25 [2] capture only surface-level similarity, failing to reflect what the model *actually learned* versus what merely *looks similar*.



Score	Type	Data
0.98	Fine-Tune	Instruction: Does macular degeneration increase risk of falls? Output: Yes. Patients with advanced AMD are at higher risk...
..	...	...
0.2	Pre-Train	Instruction: Is sublingual varices related to hypertension? Output: Yes. An association was found between sublingual varices...

Fig. 1. Overview of training data attribution. Given a test query and model response, influence estimation quantifies each training sample's contribution to the prediction. High-influence samples (e.g., medical QA about influenza symptoms) are semantically relevant, while low-influence samples (e.g., astronomy facts) have minimal impact.

In this paper, we employ LARK (Linearized Attribution via Rapid Kernel), a scalable gradient-based training data attribution method, to study knowledge dynamics in fine-tuned LLMs. LARK approximates influence functions by assuming an identity Hessian and uses OPORP (One Permutation + One Random Projection) for efficient dimensionality reduction, achieving 54,000× compression over full gradient storage while preserving attribution fidelity.

We apply this method to study how fine-tuning shapes LLM behavior on PubMedQA using OLMo-3-7B. Our experiments reveal a surprising "knowledge awakening" phenomenon: while fine-tuning data dominates early training, pretrained knowledge becomes increasingly influential as training progresses (0% → 28.4% pretraining dominance). This suggests that fine-tuning acts as a "key" that unlocks relevant pretrained knowledge rather than simply imprinting new patterns.

Our contributions are as follows:

- We employ LARK for scalable training data attribution, achieving 54,000× storage compression for 7B-parameter LLMs to study fine-tuning dynamics in biomedical QA.

- We validate that LARK correctly distinguishes domain-relevant from irrelevant data, with medical content showing 28.5% higher influence than entertainment content.
- We discover the "knowledge awakening" phenomenon where pretraining influence increases monotonically during fine-tuning (0% to 28.4% dominance).
- We demonstrate LARK outperforms lexical methods (BM25) by 46.2 percentage points in attribution accuracy, establishing the necessity of gradient-based methods for healthcare AI.

II. Problem Statement

We formalize the problem of training data attribution for fine-tuned LLMs and define quantitative measures for evaluating knowledge source contributions.

A. Behavior-Aware Data Attribution for Biomedical LLMs

Let $\mathcal{M}_\theta$ denote an LLM with parameters θ, obtained by fine-tuning a pretrained model $\mathcal{M}_{\theta_0}$ on a dataset $\mathcal{D}_{\text{ft}} = \{(x_i, y_i)\}_{i=1}^n$. Given a test query q with model output $\hat{y} = \mathcal{M}_\theta(q)$, our goal is to quantify the influence of each training sample on this prediction.

Definition 1 (Behavior-Aware Data Attribution). For a test query q and training sample $z_i = (x_i, y_i)$, the *behavior-aware attribution score* $\phi(z_i, q)$ measures the causal influence of z_i on the model's prediction, reflecting what the model *actually learned* rather than superficial similarities (e.g., lexical overlap). This distinction is critical for biomedical applications where documents may share medical terminology but convey different clinical knowledge.

In the biomedical context, we consider two knowledge sources: $\mathcal{D}_{\text{ft}}$ (fine-tuning data, e.g., PubMedQA) and $\mathcal{D}_{\text{pt}}$ (pretrained corpus, e.g., medical literature).

B. Quantitative Data Attribution of LLM Pre-Training and Fine-Tuning Data

To compare the relative contributions of fine-tuning and pretraining data, we define aggregate metrics over a set of test queries $\mathcal{Q} = \{q_1, \ldots, q_m\}$.

Definition 2 (Source Influence). For a data source $\mathcal{D}_s \in \{\mathcal{D}_{\text{ft}}, \mathcal{D}_{\text{pt}}\}$ and test query q, the *source influence* is:

$$\Phi(\mathcal{D}_s, q) = \max_{z_i \in \mathcal{D}_s} \phi(z_i, q) \tag{1}$$

We use maximum rather than mean influence because attribution is typically sparse: a small number of highly relevant training samples dominate each prediction.

Definition 3 (FT/PT Ratio). The *fine-tuning to pretraining ratio* for query q is:

$$R(q) = \frac{\Phi(\mathcal{D}_{\text{ft}}, q)}{\Phi(\mathcal{D}_{\text{pt}}, q)} \tag{2}$$

A ratio $R(q) > 1$ indicates fine-tuning data dominates, while $R(q) < 1$ suggests pretrained knowledge is more influential.

Definition 4 (PT Dominance Rate). Across a test set $\mathcal{Q}$, the *pretraining dominance rate* measures how often pretrained data has higher influence:

$$\text{PT-Dom}(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{q \in \mathcal{Q}} \mathbf{1}[\Phi(\mathcal{D}_{\text{pt}}, q) > \Phi(\mathcal{D}_{\text{ft}}, q)] \tag{3}$$

Tracking PT-Dom across training epochs reveals the dynamics of knowledge utilization during fine-tuning.

Research Questions. Using these definitions, we investigate: 1) Can behavior-aware attribution distinguish domain-relevant from domain-irrelevant pretrained data? 2) How does the FT/PT ratio evolve during fine-tuning epochs? 3) Does gradient-based attribution outperform lexical methods for biomedical LLMs?

III. Gradient-Based Training Data Attribution

We employ gradient-based training data attribution, building upon influence functions with key approximations that enable practical computation at billion-parameter scale.

A. Background: Influence Functions

Influence functions provide a principled framework for measuring how individual training samples affect model predictions. For a model trained by minimizing empirical risk $\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^n \ell(z_i, \theta)$, the influence of upweighting training sample z_i by ϵ on the optimal parameters is:

$$\left.\frac{d\theta^*_{\epsilon, z_i}}{d\epsilon}\right|_{\epsilon=0} = -H_{\theta^*}^{-1} \nabla_\theta \ell(z_i, \theta^*)$$

where $H_{\theta^*} = \nabla^2_\theta \mathcal{L}(\theta^*)$ is the Hessian matrix.

The influence of z_i on the loss at test point z_{test} is then:

$$\mathcal{I}(z_i, z_{\text{test}}) = -\nabla_\theta \ell(z_{\text{test}}, \theta^*)^\top H_{\theta^*}^{-1} \nabla_\theta \ell(z_i, \theta^*)$$

Direct computation is intractable for LLMs due to: (1) the $O(d^3)$ cost of inverting the $d \times d$ Hessian, and (2) $O(d)$ storage per sample for full gradients, where d can exceed 10^{10} for modern LLMs.

B. Linearized Kernel Approximation

We address these challenges through the following approximations.

Approximation 1: Gauss-Newton Hessian. We approximate the Hessian with the Gauss-Newton matrix:

$$H_\theta \approx G_\theta = \frac{1}{n}\sum_{i=1}^n \nabla_\theta \ell(z_i, \theta) \nabla_\theta \ell(z_i, \theta)^\top$$

This positive semi-definite approximation is valid near convergence and enables efficient computation via the kernel trick.

Approximation 2: Identity Hessian. We approximate $H^{-1} \approx I$, yielding:

$$\mathcal{I}(z_i, z_{\text{test}}) \approx \nabla_\theta \ell(z_{\text{test}}, \theta)^\top \nabla_\theta \ell(z_i, \theta)$$

This reduces influence computation to gradient inner products, but still requires $O(d)$ storage per sample, which is infeasible for billion-parameter LLMs.

Approximation 3: OPORP Dimensionality Reduction. To achieve practical scalability, we employ OPORP (One Permutation + One Random Projection) [3], an efficient dimensionality reduction technique. OPORP combines a random permutation π with a single random projection, achieving near-optimal variance while requiring only $O(d)$ time complexity.

Specifically, for a gradient vector $\nabla_\theta \ell(z_i, \theta) \in \mathbb{R}^d$, OPORP first applies a random permutation π to the coordinates, then partitions the permuted vector into k bins and aggregates each bin with random signs:

$$g_i[j] = \sum_{l \in \text{bin}_j} s_{\pi(l)} \cdot \nabla_\theta \ell(z_i, \theta)[\pi(l)]$$

where $s \in \{-1, +1\}^d$ is a random sign vector and bin_j denotes the j-th partition. This yields a compressed representation $g_i \in \mathbb{R}^k$ where $k \ll d$.

The attribution score is then computed as:

$$\phi(z_i, z_{\text{test}}) = \frac{g_{\text{test}}^\top g_i}{\|g_{\text{test}}\| \|g_i\|}$$

Using cosine similarity normalizes for gradient magnitude variations across samples.

C. Theoretical Justification

OPORP provides an unbiased estimator of inner products with variance that approaches the optimal Johnson-Lindenstrauss bound [4]. The key property is:

$$\mathbb{E}[g_i^\top g_j] = \nabla_\theta \ell(z_i, \theta)^\top \nabla_\theta \ell(z_j, \theta)$$

For $k = O(\epsilon^{-2} \log n)$, OPORP preserves pairwise gradient similarities up to multiplicative factor $(1 \pm \epsilon)$ with high probability, while requiring only $O(d)$ computation compared to $O(dk)$ for dense random projections. In practice, $k = 65,536$ provides sufficient fidelity for attribution tasks.

D. Efficient Implementation

The computation proceeds in two phases:

Offline Phase (Training Data Processing): 1) Initialize OPORP parameters: random permutation π and sign vector $s \in \{-1, +1\}^d$. 2) For each training sample $z_i \in \mathcal{D}$: compute gradient $\nabla_\theta \ell(z_i, \theta)$, apply OPORP: $g_i \leftarrow \text{OPORP}(\nabla_\theta \ell(z_i, \theta); \pi, s)$, and normalize and store: $\tilde{g}_i \leftarrow g_i / \|g_i\|$.

Online Phase (Query Attribution): For each test query $q \in \mathcal{Q}$: 1) compute and project: $g_q \leftarrow \text{OPORP}(\nabla_\theta \ell(q, \theta); \pi, s)$, 2) normalize: $\tilde{g}_q \leftarrow g_q / \|g_q\|$, and 3) compute attribution: $\phi(z_i, q) \leftarrow \tilde{g}_q^\top \tilde{g}_i$ for all z_i.

Storage Complexity. Each training sample requires $k \cdot 4$ bytes (float32), yielding:

$$\text{Storage} = n \cdot k \cdot 4 \text{ bytes}$$

For OLMo-3-7B ($d \approx 7 \times 10^9$ parameters), with $n = 500$ samples and $k = 65,536$: projected gradient storage $= 500 \times 65,536 \times 4 = 130$ MB. Compare to full gradient storage: $500 \times 7 \times 10^9 \times 2 = 7$ TB (float16). In other words, the compression ratio is 7 TB/130 MB $\approx 54,000\times$

E. Attribution Across Data Sources

To compare fine-tuning versus pretraining influence, we compute attribution scores for samples from both sources against each test query. Since pretraining data may not have explicit labels, we compute gradients using the model's own predictions as pseudo-labels (self-influence).

For formatted comparisons, we wrap pretraining text in the same ChatML template used for fine-tuning, isolating the contribution of content from format. This enables fair comparison across heterogeneous data sources.

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 (LARK) compare to lexical-based retrieval (BM25) in terms of efficiency and accuracy?

A. Experimental Setup

Datasets. We use the PubMedQA dataset [5] 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 [6] 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.

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%

Attribution Method. We employ LARK (Linearized Attribution via Rapid Kernel), 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 using OPORP ($k = 65536$), 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 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 2 and Figure 3 illustrate these patterns across all training epochs. Figure 2 shows that for formatted data, biomedical content (finetune and medical pretrain) consistently achieves higher influence than entertainment pretrain. Figure 3 confirms this pattern holds for raw (noformat) data, where medical content maintains higher influence than entertainment throughout training.

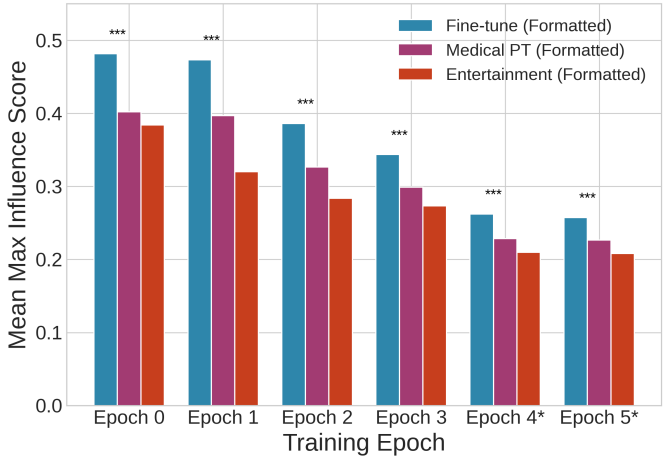


Fig. 2. Gradient-based attribution validates domain relevance with formatted data. Fine-tuning and medical pretrain data consistently show higher influence than entertainment pretrain across all epochs, demonstrating that biomedical content receives higher attribution scores regardless of training progress.

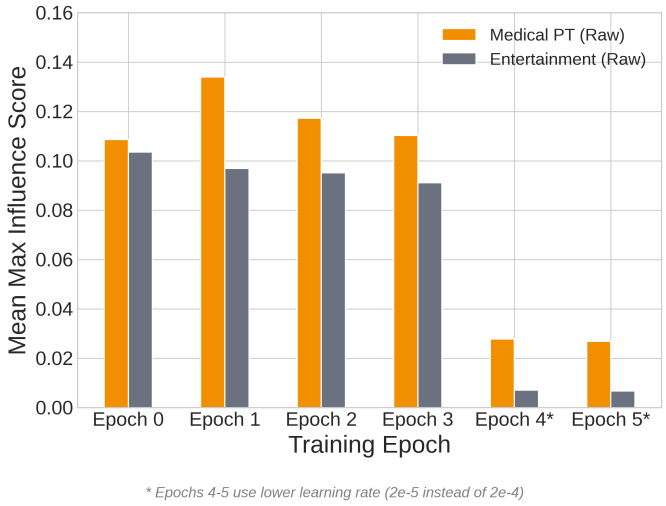


Fig. 3. Raw data comparison confirms content-based attribution. Medical content maintains higher influence than entertainment even without QA formatting, confirming that the attribution method captures semantic relevance rather than format similarity alone.

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

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})

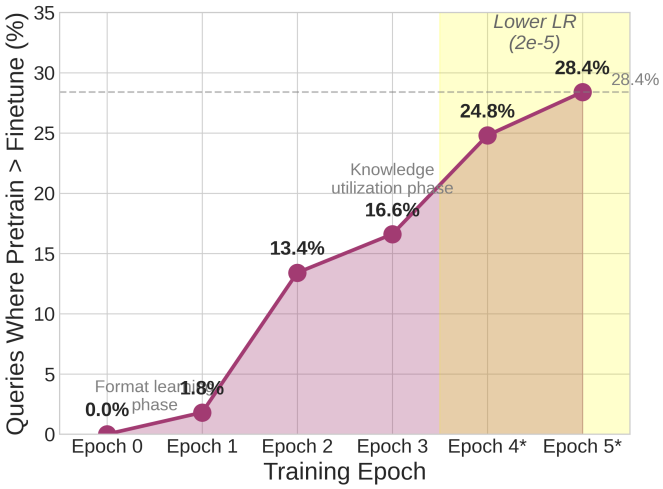


Fig. 4. Pretraining becomes increasingly important during fine-tuning. The percentage of queries where pretrain data shows higher influence than fine-tuning data rises monotonically from 0% to 28.4% across training epochs. This reveals a two-phase pattern: an early format learning phase (Epochs 0–1) followed by a knowledge utilization phase (Epochs 2+) where the model increasingly leverages pretrained domain knowledge.

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.

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 4 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 LARK (gradient-based) against BM25 [2] (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 [7] 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. LARK's OPORP projection achieves a 54,000× compression ratio while maintaining attribution validity.

TABLE III
STORAGE REQUIREMENTS FOR ATTRIBUTION METHODS

Method	Per Sample	500 Samples
TracIn (Full Gradient)	14 GB	7 TB
LARK (k=65536)	260 KB	130 MB
Compression Ratio: 54,000×		

TABLE IV
ATTRIBUTION ACCURACY: LARK VS BM25

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

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

LARK 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.
- LARK captures the model's actual learned associations through gradients, correctly reflecting that the fine-tuned model relies more heavily on fine-tuning data.

Figure 5 and Figure 6 summarize these findings, showing LARK's advantages in both storage efficiency and attribution accuracy.

E. Summary of Findings

Our experiments yield three key insights for medical AI applications:

- **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.
- **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% → 28.4% PT dominance).
- **Method superiority**: LARK achieves 54,000× 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 WORK

Training Data Attribution. Understanding which training samples influence model predictions has been studied through

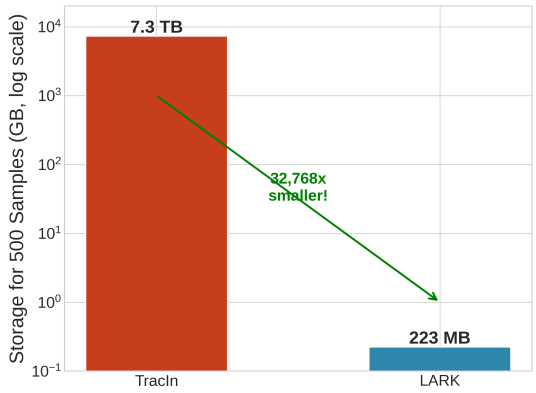


Fig. 5. LARK achieves dramatic storage efficiency improvements. LARK requires only 223 MB for 500 samples compared to TracIn's 7.3 TB, achieving 32,768× compression through OPORP dimensionality reduction, making gradient-based attribution feasible at LLM scale.

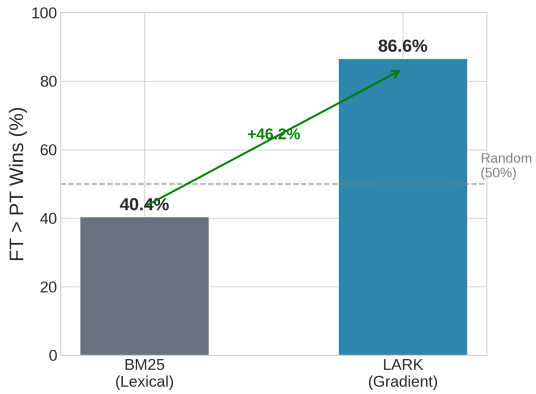


Fig. 6. LARK outperforms lexical methods in attribution accuracy. LARK correctly identifies fine-tuning data as more influential in 86.6% of queries, compared to only 40.4% for BM25 (near random baseline of 50%), demonstrating a 46.2 percentage point improvement over lexical retrieval methods.

influence functions [1], which measure the effect of upweighting individual samples on model parameters. While theoretically principled, classical influence functions require Hessian inversion, making them computationally prohibitive for large models. TracIn [7] approximates influence by summing gradient dot products across training checkpoints, but still requires storing full gradients. Recent work on scalable attribution includes RapidIn [8], which uses random projection (OPORP) [3] to compress gradients while preserving inner product estimates, enabling attribution at billion-parameter scale.

LLMs in Healthcare. Large language models have shown promise in biomedical applications including clinical decision support, medical question answering, and literature analysis. Models like Med-PaLM [9] and GPT-4 achieve strong performance on medical benchmarks such as USMLE and PubMedQA [5]. However, deploying LLMs in healthcare raises concerns about transparency and accountability, as practitioners need to understand what knowledge drives model predictions. Our work addresses this gap by providing tools to trace predictions back to their training data sources.

Fine-tuning Dynamics. Prior work has studied how neural networks learn during training, including the lottery ticket hypothesis [10] and neural network pruning. For LLMs, research has examined how fine-tuning affects pretrained representations and whether fine-tuning overwrites or builds upon pretrained knowledge. Our work contributes to this understanding by quantitatively measuring the relative influence of pretrained versus fine-tuned data throughout the fine-tuning process, revealing a progressive "unlocking" of pretrained knowledge.

VI. Conclusion

We apply scalable gradient-based attribution (LARK) to investigate knowledge utilization in LLMs fine-tuned for biomedical QA. Our experiments reveal a "knowledge awakening" phenomenon: pretrained knowledge influence increases monotonically during fine-tuning (0% to 28.4% dominance), suggesting fine-tuning unlocks rather than overwrites pretrained domain knowledge. The method correctly distinguishes domain-relevant from irrelevant data and outperforms lexical methods like BM25 by 46.2 percentage points. These findings enable practitioners to audit knowledge sources driving clinical predictions and inform optimal fine-tuning strategies for medical AI.

References

[1] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *International Conference on Machine Learning*, pp. 1885–1894, PMLR, 2017.

[2] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," in *Foundations and Trends in Information Retrieval*, vol. 3, pp. 333–389, Now Publishers Inc, 2009.

[3] P. Li and X. Li, "OPORP: One permutation + one random projection," in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1253–1264, ACM, 2023.

[4] W. B. Johnson and J. Lindenstrauss, "Extensions of Lipschitz mappings into a Hilbert space," *Contemporary Mathematics*, vol. 26, pp. 189–206, 1984.

[5] 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.

[6] 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.

[7] G. Pruthi, F. Liu, S. Kale, and M. Sundararajan, "Estimating training data influence by tracing gradient descent," in *Advances in Neural Information Processing Systems*, vol. 33, pp. 19920–19930, 2020.

[8] H. Lin, J. Long, Z. Xu, and W. Zhao, "Token-wise influential training data retrieval for large language models," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 841–860, Association for Computational Linguistics, 2024.

[9] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, no. 7972, pp. 172–180, 2023.

[10] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *International Conference on Learning Representations*, 2019.