Black-Box Dataset-Level Source Ranking via Context Mixture Factorization

Milad Fotouhi
Amazon & University of Washington
mfotouhi@uw.edu

Mohammad Taha Bahadori
Amazon & University of Washington
bahadori@uw.edu

Oluwaseyi Feyisetan
Amazon
sey@amazon.com

Payman Arabshahi
University of Washington
paymana@uw.edu

Abstract

Training data attribution (TDA) for large language models is dominated by gradient- and retraining-based methods that require model internals, ruling out the most widely deployed closed-weight systems. We study a strictly black-box variant of the problem: given a set of candidate datasets, rank them by how much an instruction-tuned LLM's behavior on a query set is consistent with having previously seen each one, using only input–output access through retrieval-augmented prompting. We introduce **Context Mixture Factorization (CMF)**, which models the LLM's output distribution as a mixture of a base component and dataset-specific components and recovers per-dataset contribution weights by regularized matrix factorization via alternating projected least squares. CMF explicitly decomposes background (π_0) and dataset-specific influence and requires only $K+1$ prompts per query for K candidate datasets. As a comparison method we also formulate the **Shapley Context Method (SCM)**, a residual-similarity attribution score aggregated via Shapley values. We evaluate four instruction-tuned LLMs on BoolQ, a synthetic FakeQ variant, and a post-cutoff Olympic2024 dataset, against three baselines: a TRAK-style gradient-similarity scorer, ContextCite, and retrieval cosine similarity. CMF and SCM recover the ordering predicted by our exposure-proxy setup (BoolQ>FakeQ>Olympic2024), and a controlled fine-tuning sweep shows attribution rises monotonically with actual exposure. We are explicit that the resulting scores measure a behavioral proxy for parametric-knowledge alignment, not verified pretraining inclusion: FakeQ (created by us, seen by no model) receives non-trivial CMF mass, consistent with the metric also reflecting topical/structural similarity. CMF is most useful when interpreted alongside its π_0 decomposition and the fine-tuning monotonicity result rather than as an exposure probability in absolute terms.

1 Introduction

Training Data Attribution (TDA) quantifies the contributions of different data sources to a model's outputs (Park et al., 2023; Nguyen et al., 2023). It is essential for curating training corpora and debugging neural networks (Xia et al., 2024; Qin et al., 2025), and for assigning monetary value to proprietary training data — a prerequisite for fair compensation and data management (Ghorbani and Zou, 2019; Nohyun et al., 2022; Choe et al., 2024). Unlike retraining- or gradient-based approaches, our methods require no model internals and instead exploit in-context learning behavior, making them directly applicable to black-box LLMs.

Existing TDA methods fall mainly into two categories (Hammoudeh and Lowd, 2024; Worledge et al., 2024). Retraining approaches (Feldman and Zhang, 2020; Ghorbani and Zou, 2019) retrain the model without the target data source, which is computationally expensive. Influence-function approaches (Koh and Liang, 2017; Pruthi et al., 2020; Chen et al., 2021; Park et al., 2023) relax the need for full retraining by computing only a few gradients per data point; recent work has pushed this line to LLM scale, with closed-form approximations for LoRA-tuned models (Kwon et al., 2024), gradient-projection approaches that report large throughput gains on Llama3-8B (Choe et al., 2024), and scalable fact-tracing across 160B-token corpora (Chang et al., 2025). Despite their efficiency, influence-function methods rely on a linear approximation that can be inaccurate, and they compute dataset-level attribution as a sum of per-point scores — a procedure that does not capture holistic dataset contribution (Hammoudeh and Lowd, 2024; Park et al., 2023). Both families require access to model internals, which is infeasible for closed-weight LLMs. The related technique of machine unlearning (Ginart et al., 2019; Sekhari et al., 2021) is also expensive for obtaining

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 33781–33801
October 24-29, 2026 ©2026 Association for Computational Linguistics

contribution scores, although recent benchmarks and methods such as TOFU (Maini et al., 2024a), WMDP/RMU (Li et al., 2024), NPO (Zhang et al., 2024), and MUSE (Shi et al., 2024b) have substantially advanced LLM-specific unlearning evaluation.

A second, complementary line of work attributes outputs without relying on gradients. Lookup-based systems such as Infini-gram (Liu et al., 2024a) and OLMoTrace (Liu et al., 2025) surface verbatim training-data matches at trillion-token scale, providing surface-level attribution but no notion of dataset-level contribution. Closer to our setting, ContextCite (Cohen-Wang et al., 2024) attributes a single LLM generation to spans of a single retrieved context via random ablations, and MIRAGE (Qi et al., 2024) performs analogous span-level attribution in RAG using model internals. A related but distinct line of work targets dataset-level exposure detection: *LLM Dataset Inference* (Maini et al., 2024b) aggregates membership-inference features over an entire candidate corpus to test "did the model train on this dataset," Min-K% Prob (Shi et al., 2024a) is a black-box membership-inference probe, and black-box test-set contamination methods (Golchin and Surdeanu, 2024; Oren et al., 2024) detect whether specific benchmark instances leaked into pretraining. These methods are the closest existing answer to the binary inclusion question and use loss/likelihood signals rather than the behavioral output-similarity signal that CMF/SCM use; we view them as complementary measurements. Duan et al. (2024) cautions that exposure evaluations built from members and non-members drawn from different distributions can be confounded by temporal/topical shift; we revisit this caveat when discussing our FakeQ and Olympic2024 results. Our setting and unit of analysis differ from the prior context-attribution work in three ways. First, our *unit* of attribution is an entire candidate *dataset*, not a span; we aggregate across queries to produce a per-source score useful for data valuation and unlearning audits. Second, our model assumes that the LLM's output mixes a fixed background distribution with dataset-specific distributions, and CMF *explicitly recovers* the background weight π_0; span-level ablation methods provide no analogous decomposition. Third, CMF requires only $K{+}1$ prompts per query for K candidate datasets, making whole-dataset scoring tractable on closed-weight LLMs where per-query ablation budgets are limited. Benchmarks such as DATE-LM (Deng et al., 2025) have begun to standardize evaluation of LLM-scale TDA; we discuss our evaluation choices in §4.

Contributions.

- We formulate dataset-level training data attribution for closed-weight LLMs as a regularized matrix-factorization problem and introduce **CMF**, which recovers the mixture weights from $K{+}1$ prompts per query and explicitly decomposes background versus dataset-specific influence (π_0 vs. π_k). A synthetic identifiability experiment shows CMF recovers known mixture weights with L_1 error 0.10 at noise $\sigma{=}0.10$.

- We position CMF against **SCM**, a Shapley-value variant of the residual-similarity intuition that does not separately model a background component, which clarifies which empirical gains come from the mixture structure and which from the marginal-contribution framing.

- We construct **FakeQ** (synthetic perturbation of BoolQ) and **Olympic2024** (constructed from post-July-2024 data, post-cutoff for the open-weight checkpoints we use according to their published model cards) as attribution test beds spanning likely-seen, synthetic-unseen, and unseen-to-high-confidence exposure, and evaluate four instruction-tuned LLMs (GPT-4 (Achiam et al., 2023), bloomz-1b1 (Le Scao et al., 2023), Mistral-7B-Instruct (Jiang et al., 2023), Phi-3-mini (Abdin et al., 2024)) on BoolQ (Clark et al., 2019), FakeQ, and Olympic2024.

- Against a TRAK-style gradient-similarity scorer (Park et al., 2023) (an adaptation that scores candidate corpora rather than training-set examples), ContextCite, and retrieval cosine similarity, our methods recover the ordering predicted by our exposure-proxy setup; ContextCite, by contrast, inverts on Olympic2024 because it measures context reliance rather than training-data exposure. A controlled fine-tuning sweep and a multi-checkpoint unlearning case study provide further evidence that the metrics behave as intended.

2 Methodology

We have access to tuples (q, c, y) consisting of a question q, a context c, and an outcome y. The LLM outputs $M(q \mid c) = p(y \mid q, c)$; we write $c = \emptyset$ when no context is provided. Our goal is to quantify the contributions of training datasets $D_1, \ldots, D_K$ in $p(y \mid q, c)$, given a query set $Q = \{q_1, \ldots, q_m\}$. Without loss of generality, we describe the methods for binary outcomes $y \in \{0, 1\}$.

What "attribution" means here. We distinguish three quantities often conflated under "attribution": (i) *training-data exposure* (did the model see this data during pretraining), (ii) *contextual utility* (does this data change the output when supplied as context), and (iii) *retrieval relevance* (semantic similarity to the query). CMF and SCM are behavioral *proxies for (i)*; ContextCite (Cohen-Wang et al., 2024) measures (ii); the retrieval-cosine baseline measures (iii). Because CMF and SCM operate on output-similarity signals between no-context and with-context prompts, they more precisely measure *parametric-knowledge alignment between the model and a candidate dataset* — a quantity that correlates with exposure but also reflects topical/structural similarity and the model's prior ability to answer the query. We therefore do not claim that high CMF/SCM scores certify training-set inclusion; we treat the metrics as behavioral source-ranking proxies and validate the *direction* of the signal in a controlled fine-tuning sweep (§4.3). Methods that aim at the strict membership question — e.g., LLM Dataset Inference (Maini et al., 2024b) or Min-K% Prob (Shi et al., 2024a) — use loss-based signals and are complementary.

2.1 The Shapley Context Method (SCM)

The intuition is that if an LLM has internalized information from D_k, then adding D_k as RAG context should leave the with-context output close to the no-context output. We operationalize this via Shapley aggregation of pairwise similarities. For a subset $S \subseteq \{D_1, \ldots, D_K\}$, let

$$s_S = \mathrm{sim}\big(M(q),\, M(q \mid c_S)\big) \in [0, 1] \quad (1)$$

be the similarity between the no-context output $M(q)$ and the with-context-S output $M(q \mid c_S)$. The Shapley value of dataset D_k is

$$\phi_k = \sum_{S \subseteq \{D_1, \ldots, D_K\} \setminus \{D_k\}} C_{S,K}\big(s_{S \cup \{D_k\}} - s_S\big), \quad (2)$$

with normalization $C_{S,K} = |S|!(K - |S| - 1)!/K!$.

Interpretation. A larger *raw* ϕ_k means that, averaged across subsets, adding D_k pulls the with-context output *closer* to the no-context output — i.e., D_k aligns with what the model already produces. We read this as evidence of prior internalization. A smaller (or negative) raw ϕ_k means D_k, when added as context, pushes the with-context output *further* from the no-context output. SCM thus tracks proxy (i) above through the lens of context-induced similarity changes, rather than measuring (ii) directly.

Reporting convention. We report raw ϕ_k throughout (no min-max rescaling), so the marginal-contribution interpretation is preserved. CMF's source-normalized $\widehat{\pi}$ and SCM's raw ϕ_k are therefore on similar but not strictly identical scales; cross-method comparison in Table 1 relies on qualitative ordering, not exact numerical equivalence. SCM absorbs cross-source variability and retrieval noise into a single Shapley term; CMF (below) makes the noise structure explicit via π_0. Algorithm 1 (Appendix A) gives the full procedure.

2.2 Context Mixture Factorization (CMF)

Mixture modeling has long been used to capture latent source contributions (Nguyen et al., 2023). CMF adapts this to dataset attribution: separating a base component π_0 from dataset-specific components yields a more stable estimator under retrieval noise. We model

$$p(y \mid q) = \pi_0 \widetilde{p}_0(y \mid q) + \sum_{k=1}^{K} \pi_k \widetilde{p}_k(y \mid q), \quad (3)$$

where $\widetilde{p}_0$ is a general-purpose language model and $\widetilde{p}_k$ are latent models specialized on each D_k. We model the effect of providing the kth dataset as context as the intervention

$$p(y \mid q, c_k) = \pi_0 \widetilde{p}_0(y \mid q) + (1 - \pi_0)\widetilde{p}_k(y \mid q). \quad (4)$$

The key assumption is that the right-hand sides of Eq. (3) and (4) contain no context terms; we discuss this assumption's limits in Remark 1 and validate it empirically in Appendix G.

Remark 1 (behavioral proxy, not mechanistic decomposition). We emphasize that Eq. (3)–(4) define a *statistical surrogate* for the LLM's outputs, not a claim about its parameters or internal computation. CMF does not partition the model's weights or activations; it estimates the mixture coefficients $\boldsymbol{\pi}$ that best explain the observed input–output behavior under retrieval-augmented prompting. The recovered $\widehat{\pi}_k$ should therefore be interpreted as "the empirical weight at which behavior on queries that retrieve from D_k resembles a D_k-specialized model," not as "the fraction of the model's parameters allocated to D_k." Model (3) further does not capture cross-source interactions (e.g., a retrieved context cueing recall of unrelated pretraining facts), but allows us to recover per-dataset contributions without estimating each $\widetilde{p}_k$ explicitly.

Matrix factorization formulation. For each of the m queries we issue $K{+}1$ prompts (the no-context prompt and one per dataset), giving the linear system

$$P = \Pi \widetilde{P}, \tag{5}$$

where $P \in [0,1]^{(K+1)\times m}$ is observed and $\Pi \in [0,1]^{(K+1)\times(K+1)}$ and $\widetilde{P} \in [0,1]^{(K+1)\times m}$ are latent. The structure of Π encodes Eq. (4): row k has weight π_0 on the base and $1-\pi_0$ on $\widetilde{p}_k$. In principle one could use up to 2^K prompts (varying S); we choose $K{+}1$ for compute efficiency and rely on regularization for identifiability.

Alternating projected least squares. We estimate $\boldsymbol{\pi}$ by jointly minimizing reconstruction error with two regularizers — entropy on $\boldsymbol{\pi}$ and Frobenius shrinkage of $\widetilde{P}$ toward $1/2$:

$$\begin{aligned} \widehat{\boldsymbol{\pi}} = \arg\min_{\boldsymbol{\pi},\widetilde{P}} & \left\| P - \Pi\widetilde{P} \right\|_F^2 \\ & - \lambda_\pi H(\boldsymbol{\pi}) + \lambda_{\widetilde{P}} \left\| \widetilde{P} - \tfrac{1}{2} \right\|_F^2, \\ \text{s.t.} \quad & \boldsymbol{\pi} \succeq \mathbf{0},\ \mathbf{1}^\top \boldsymbol{\pi} = 1,\ 0 \preceq \widetilde{P} \preceq 1, \end{aligned} \tag{6}$$

where $H(\cdot)$ denotes Shannon entropy. The entropy term encodes a null hypothesis of equal source contributions; the Frobenius term encodes that latent probabilities should be uninformative absent evidence to the contrary. Without these regularizers the problem admits many solutions; a sensitivity sweep (Appendix H) shows that $\lambda_\pi = \lambda_{\widetilde{P}} = 10^{-1}$ gives stable solutions across random initializations.

Eq. (6) is biconvex — fixing either $\boldsymbol{\pi}$ or $\widetilde{P}$ leaves the other a convex problem (Gorski et al., 2007) — so we solve it by alternating projected least squares. We initialize $\widetilde{P}$ near $1/2$ and $\boldsymbol{\pi}$ near $1/(K+1)$. Confidence intervals for both SCM and CMF are obtained by bootstrapping over the query set (Tibshirani and Efron, 1993). Algorithm 2 (Appendix A) gives the full procedure.

When does CMF recover the true mixture? CMF is identifiable up to regularization shrinkage when (a) the latent distributions $\widetilde{p}_k$ are distinguishable on Q, (b) the retrieval c_k actually surfaces D_k-relevant content, and (c) regularization breaks the rotational symmetry of $\Pi\widetilde{P}$ without dominating the fit. Conditions (a)–(b) are testable empirically (rank/condition number of $\widetilde{P}$; per-source retrieval quality); (c) is addressed by the sensitivity sweep in Appendix H. The synthetic recovery experiment in Appendix G verifies that CMF recovers known $\boldsymbol{\pi}^*$ to L_1 error ≤ 0.10 at noise $\sigma \leq 0.10$, degrading smoothly rather than catastrophically.

3 Implementation

3.1 Prompt Engineering

We use binary Yes/No QA tasks for tractable scalar similarity. After testing several prompt variants (Appendix B), we settled on the boolean instruction *"Given the context below, answer the question that follows with only 'Yes', 'No', or 'I don't know' if the context is insufficient. {question}? The answer to this question is"*. This prompt produces clean single-token responses for GPT-4, bloomz-1b1, and Mistral 7B. Phi-3-mini often emits additional text, so we apply a zero-shot classifier (1,000-character window) to map free-form output to {Yes, No, IDK} before computing similarity. Appendix I shows this choice shifts absolute Phi-3 attribution by at most one CI width without changing qualitative rankings.

3.2 Retrieval-Augmented Generation

Since entire datasets do not fit in an LLM context window, we use RAG (Lewis et al., 2020) to surface dataset-relevant passages per query. We split each corpus into chunks with the LangChain *RecursiveCharacterTextSplitter* (chunk size 512 tokens, no overlap), embed chunks with *thenlper/gte-small* (384-d embeddings), and index them with FAISS (Johnson et al., 2019). At query time we embed the query, retrieve the top k=3 nearest chunks per source, and truncate the concatenated context to 2,000 characters. Appendix E reports the effect of context provision on raw task accuracy.

3.3 Provenance and reproducibility

All reported values are computed from completed experimental runs. Bootstrap CIs use B=1,000 resamples of the query set (fixed seed). Exact model checkpoints, decoding settings, similarity function, hardware (a single AWS EC2 g6.16xlarge, 4×L4), software versions, and total compute (∼80 GPU-hours for CMF, ∼360 for SCM, ∼60 for TRAK-style scoring, plus the GPT-4 API budget detailed in Appendix L) are listed in that appendix. To support independent reimplementation, Appendix L documents the complete experimental configuration — prompt templates, retrieval settings, model checkpoints and decoding parameters, the similarity function, optimization hyperparameters, and per-method compute — and the Olympic2024 construction procedure is specified in full in the appendix.

4 Experiments

4.1 Setup

Datasets. We evaluate on three datasets that span a continuum of *expected* exposure to the LLMs we test: (1) the BoolQ Q&A dataset (Clark et al., 2019) of (question, context, binary answer) tuples, which we expect instruction-tuned LLMs to have seen during pretraining; (2) **FakeQ**, constructed by altering the queries and contexts in BoolQ so the surface text is unseen even though the topical distribution is preserved; and (3) **Olympic2024**, a newly constructed dataset (1,200 context snippets, 3,600 Yes/No QA pairs split equally into train/val/test) based on the Paris 2024 Olympics, which took place 26 July–11 August 2024 (Appendix D). For the open-weight models in our study — bloomz-1b1, Phi-3-mini-4k-instruct, and Mistral-7B-Instruct-v0.2 — Olympic2024 falls strictly after each model card's published training cutoff, so Olympic2024 is unseen by those checkpoints to the extent the model cards are accurate. For GPT-4 we use the `gpt-4-turbo-2024-04-09` API snapshot, which predates the Paris 2024 Olympics; Olympic2024 is therefore very unlikely to be included in the snapshot's training data, but because OpenAI does not disclose the corpus we avoid treating this as verified ground truth. We use five additional corpora from the Wikipedia "Field of Science" subset on HuggingFace (Chemistry, Natural Science, History and Archaeology, Biology, Law) as background-control sources; each contains more than one million samples.

Models and baselines. We evaluate four instruction-tuned LLMs spanning closed and open weights: **GPT-4** (Achiam et al., 2023) (closed weights, accessed via API; `gpt-4-turbo-2024-04-09` snapshot), **bloomz-1b1** (Le Scao et al., 2023) (1.1B parameters, open weights), **Mistral-7B-Instruct-v0.2** (Jiang et al., 2023) (7.3B, open weights), and **Phi-3-mini-4k-instruct** (Abdin et al., 2024) (3.8B, open weights). Throughout this paper, references to "Bloomz" or "BLOOMZ" denote the `bigscience/bloomz-1b1` checkpoint; we did not use the larger Bloomz-176B variant. All models are prompted with the boolean-output template described in §3.1; per-LLM accuracy with and without RAG context is reported in Appendix Table 2. We compare three baselines against our two proposed methods (CMF, SCM):

- **ContextCite** (Cohen-Wang et al., 2024): a span-level context-reliance measure used here as a contrast rather than a direct competitor (it answers "which spans of this context did the model use" rather than "which dataset did the model see"). We use the authors' default of 64 random ablations per query (scikit-learn LASSO, α=0.01; ablation units = retrieved chunks; per-chunk attributions summed per source, then per-query normalized). For open-weight LLMs we use the response probability from local logits; for GPT-4 we substitute the top-token log-probability from the chat API on a 500-query subsample per dataset (Appendix L).
- **Retrieval-cosine**: the average cosine similarity between the query embedding and the top-k retrieved chunk embeddings per source — a sanity baseline that measures retrieval relevance only.
- **TRAK-style gradient-similarity scoring** (Park et al., 2023): rather than running TRAK in its standard configuration (training-set examples, ensembling over models trained on training-set subsets), we adapt the projected-gradient feature mechanism to score *candidate corpora* from a single final checkpoint, and call this "TRAK-style gradient-similarity scoring" to avoid conflating it with the LDS-evaluated method of Park et al. (2023). We use the public TRAK projection mechanism (default projection dimension) on bloomz-1b1 and Phi-3-mini (§4.5).

We use m=2,000 queries per dataset, with K_{src}=6 candidate sources total (5 Wikipedia background controls plus the target dataset for that evaluation). For BoolQ and FakeQ this sample is drawn from each dataset's validation split. Olympic2024's validation and test splits jointly contain 2,400 QA pairs (1,200 each); we sample m=2,000 from their union (held out from the fine-tuning split used in §4.3). Under this convention CMF issues $K_{\text{src}}+1 = 7$ prompts per query (one with no context plus one per candidate source) and SCM evaluates up to $2^{K_{\text{src}}} = 64$ unique context subsets per query. All reported values are means over the query set with 95% bootstrap confidence intervals (B=1,000 resamples, fixed seed).

4.2 Main results

Table 1 reports attribution to the target source on BoolQ, FakeQ, and Olympic2024 for all four LLMs and all five methods. Under our exposure-proxy setup the predicted ordering is BoolQ>FakeQ>Olympic2024, reflecting that BoolQ is likely-seen, FakeQ shares topical distribution but not surface text with BoolQ, and Olympic2024 is post-cutoff for the open-weight models according to their published model cards. The measured ordering recovered by CMF, SCM, and TRAK matches this prediction.

ContextCite measures context reliance, not training-data exposure. Reading Table 1 rowwise reveals a structural difference between ContextCite and our methods. SCM, CMF, and TRAK follow the predicted exposure ordering BoolQ>FakeQ>Olympic2024. ContextCite shows the *opposite* ordering: attribution *increases* as the dataset becomes less likely to have been seen during pretraining (italicized in Table 1). This is consistent with what ContextCite is designed to measure — how much the model's generation depends on the supplied context — because a model with no parametric knowledge of a topic must rely on the context. We therefore view ContextCite as orthogonal rather than competitive: it answers "which spans of this context did the model use?" whereas CMF/SCM answer "which dataset is this output most reliant on having seen?" For training-data attribution specifically, ContextCite's inverse ranking on Olympic2024 makes it unsuitable as a drop-in replacement.

Retrieval similarity alone is not informative. The Retrieval-cosine row is high (0.70–0.82) across all three datasets, with no meaningful separation between exposure levels. The gap between BoolQ and Olympic2024 attribution observed for SCM, CMF, and TRAK-style scoring cannot therefore be explained by retrieval relevance alone.

What the BoolQ>FakeQ>Olympic2024 ordering does and does not show. The headline ordering is consistent with our exposure-proxy setup, but is not a strict inclusion verdict. FakeQ is unseen by every evaluated model, yet CMF assigns it $\widehat{\pi} \approx 0.48$–$0.53$ (Panel B) — the metric correctly registers that FakeQ shares topical/structural distribution with BoolQ. The ordering should be read as tracking the continuum from likely-seen-and-topically-aligned (BoolQ), through similar-but-unseen (FakeQ), to post-cutoff (Olympic2024), not as a membership decision. Loss-based methods that target strict inclusion (LLM Dataset Inference, Maini et al., 2024b; Min-K% Prob, Shi et al., 2024a) are complementary; the topical/temporal-shift confound is shared with membership-inference benchmarks more broadly (Duan et al., 2024). See Limitations for further discussion.

CMF–SCM comparison. CMF gives larger target-source values and tighter CIs than SCM on the same evaluation; because the two use different normalizations the magnitude gap should not be over-interpreted. The gap shrinks as exposure decreases (largest on BoolQ, smallest on Olympic2024), consistent with CMF absorbing background variance into the explicit π_0 term while SCM folds it into the marginal residual.

Per-source decomposition. CMF's full raw mixture weights $\boldsymbol{\pi}$ on the BoolQ evaluation are reported in Table 3 (Appendix F; columns sum to 1.00 by Eq. 6). The background term π_{Base} is highest for GPT-4 (0.08), consistent with GPT-4 having the strongest no-context accuracy on BoolQ (0.73, Appendix Table 2). The Wikipedia-domain controls receive low attribution (0.05–0.09), supporting the interpretation that CMF separates task-relevant from background-relevant signal. SCM and TRAK per-source tables, and per-source CMF results for FakeQ and Olympic2024, are also in Appendix F; on Olympic2024 the controls absorb much more mixture weight (Table 9), consistent with no candidate source matching the model's parametric knowledge on a post-cutoff dataset.

Table 1: Attribution to the target source on BoolQ (Panel A), FakeQ (Panel B), and Olympic2024 (Panel C). Values are mean±95% bootstrap CI half-width over m=2,000 queries. Methods use different normalizations and so the columns are on similar but not strictly identical scales: **CMF** reports the source-normalized target weight $\widehat{\pi}_{\text{target}} = \pi_{\text{target}}/(1 - \pi_{\text{Base}})$ (raw mixture weights summing to one across all candidate sources are in Appendix F); **SCM** reports raw Shapley values ϕ_{target} from Eq. (2) (no min-max rescaling); **ContextCite** reports per-query-normalized source attributions averaged over queries; **Retrieval-cosine** is a raw mean cosine similarity; **TRAK-style** (gradient-similarity scoring) reports per-source aggregated projected-gradient scores from a single final checkpoint, not the standard LDS-evaluated TRAK. **Bold** marks the largest CMF/SCM-family value per cell; *italic* marks ContextCite's inverted ordering in Panel C. **Takeaway:** CMF, SCM, and TRAK all follow the predicted exposure ordering BoolQ>FakeQ>Olympic2024 on every model, with CMF producing the sharpest separation between panels (e.g., 0.63–0.67 on BoolQ vs. 0.15–0.20 on Olympic2024); ContextCite inverts the ordering because it measures context reliance rather than exposure, and retrieval-cosine barely separates the panels at all.

Method	bloomz-1b1	GPT-4	Mistral 7B	Phi-3-mini
Panel A: BoolQ (likely-seen evaluation)				
Retrieval-cos	0.79±0.01	0.82±0.01	0.80±0.01	0.78±0.01
ContextCite	0.51±0.03	0.47±0.03	0.49±0.03	0.46±0.03
SCM	0.48±0.04	0.59±0.03	0.57±0.04	0.50±0.04
CMF	**0.63±0.02**	**0.67±0.02**	**0.65±0.02**	**0.62±0.02**
TRAK	0.61±0.02	–	–	0.58±0.02
Panel B: FakeQ (synthetic-unseen evaluation)				
Retrieval-cos	0.71±0.01	0.74±0.01	0.72±0.01	0.70±0.01
ContextCite	0.58±0.03	0.55±0.03	0.56±0.03	0.54±0.03
SCM	0.32±0.04	0.36±0.04	0.33±0.04	0.28±0.04
CMF	**0.49±0.03**	**0.53±0.03**	**0.51±0.03**	**0.48±0.03**
TRAK	0.39±0.02	–	–	0.35±0.02
Panel C: Olympic2024 (post-cutoff evaluation)				
Retrieval-cos	0.76±0.01	0.78±0.01	0.75±0.01	0.73±0.01
ContextCite	*0.71±0.03*	*0.68±0.03*	*0.70±0.03*	*0.66±0.03*
SCM	0.08±0.03	0.11±0.03	0.09±0.03	0.07±0.03
CMF	**0.16±0.02**	**0.20±0.02**	**0.18±0.02**	**0.15±0.02**
TRAK	0.12±0.02	–	–	0.09±0.02

Identifiability and regularization. The synthetic recovery experiment (Appendix G) shows CMF recovers known $\boldsymbol{\pi}^*$ to L_1 error 0.10 at σ=0.10 vs. 0.22 for SCM; the regularization sweep (Appendix H) confirms $\lambda_\pi = \lambda_{\widetilde{P}} \geq 10^{-1}$ (our default) is stable across initializations. Appendix G.1 analyzes how recovery is expected to degrade at larger source counts K and for mutually similar sources, and gives practical guidance for reading CMF rankings at scale.

4.3 Validation through fine-tuning

To probe whether the attribution metrics track actual exposure, we fine-tune bloomz-1b1 and Phi-3-mini on Olympic2024 with three learning rates $(1\times10^{-5}, 5\times10^{-5}, 1\times10^{-4})$ and compute SCM and CMF attribution at each checkpoint. Figure 1 shows both metrics increase monotonically with the fine-tuning learning rate, consistent with the metrics tracking the model's growing reliance on the dataset; CMF's values are higher than SCM's at every learning rate.

4.4 Case study: ranking unlearning methods

As a downstream use case (not a core claim of this paper), we apply CMF as an evaluation metric for three LLM unlearning methods — Gradient Ascent (GA) (Golatkar et al., 2020; Liu et al., 2024b), Fine-tuning with Random Labels (RL) (Golatkar et al., 2020), and Unlearning with Adversarial Samples (AS) (Cha et al., 2024) — by measuring residual CMF attribution to Olympic2024 after each method is applied to checkpoints fine-tuned at three learning rates $(1\times10^{-5}, 5\times10^{-5}, 1\times10^{-4})$. Across both LLMs (bloomz-1b1 and Phi-3) and all three checkpoints, the ranking **RL** $\preceq$ **GA** $\prec$ **AS** in residual attribution is stable, and post-unlearning attribution scales with the pre-unlearning attribution: no method fully reverses the highest-LR exposure. Full tables (CMF and SCM, both LLMs) are in Appendix K. We treat this as a sensitivity demonstration; the main CMF validation is the fine-tuning sweep in §4.3.

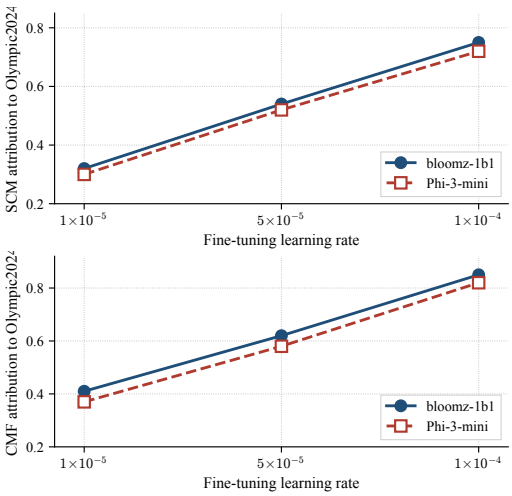


Figure 1: Attribution to Olympic2024 vs. fine-tuning learning rate, for SCM (top) and CMF (bottom). Only bloomz-1b1 and Phi-3-mini are shown; we did not fine-tune GPT-4 (no fine-tuning access via API) or Mistral 7B (omitted to keep the sweep within compute budget). Both metrics increase monotonically; CMF's values are higher across the sweep. Numerical values match the Post-FT columns of Appendix Tables 14–17.

A Goodhart-style caveat for unlearning audits. Because CMF scores are behavioral, an unlearning method optimized *against* a behavioral metric could suppress measured attribution without removing the underlying knowledge. Our case study avoids this coupling by evaluating three generic methods (GA, RL, AS) that were not designed with CMF in mind, but the risk applies to any single-metric evaluation. We therefore recommend that CMF-based unlearning audits be paired with complementary signals — e.g., loss-based membership inference (Shi et al., 2024a) — so that agreement across mechanistically distinct metrics, rather than any one score, supports a claim of successful unlearning.

4.5 Runtime and TRAK applicability

CMF is significantly faster than SCM because it requires fewer prompts per query: $K_{\text{src}}+1 = 7$ for CMF versus up to $2^{K_{\text{src}}} = 64$ unique context subsets for SCM (with caching across the Shapley computations). On an AWS EC2 g6.16xlarge instance, CMF takes 77–94 minutes per LLM (2,000 queries) versus 352–384 minutes for SCM; ContextCite (64 ablations per query) is comparable in compute to SCM (291–328 minutes). **TRAK applicability:** GPT-4 exposes no weights or gradients; Mistral 7B is open-weight, but full-parameter gradient features exceeded our GPU memory and runtime budget at 7B scale — a practical limitation, not a fundamental one (Chang et al., 2025). A convergence diagnostic showing CMF attribution stabilizes by m=2,000 is in Appendix J. SCM's wider CIs trace to dispersion in the per-pair residual $r_{k,S}$, which varies across LLMs from -0.32 on Bloomz to -0.03 on GPT-4 (Appendix M); CMF absorbs this dispersion into the explicit π_0 term.

Additional robustness checks. Appendix N adds a no-context/with-context agreement baseline and open-weight membership-inference baselines (Min-K% Prob, Min-K%++, LLM Dataset Inference). Agreement recovers the same coarse BoolQ>FakeQ>Olympic2024 ordering but is a single dataset-level statistic, without CMF's decomposition or controlled-exposure response (§4.3); Min-K% and dataset inference qualitatively agree with CMF but yield lower seen-vs-post-cutoff AUC (0.61–0.71 vs. 0.83–0.86) and do not apply to GPT-4. A paired test (Appendix Table 24) shows CMF's target-vs-control separation is significant on BoolQ and FakeQ for all models, and on Olympic2024 for GPT-4/Mistral but not Bloomz/Phi-3 — consistent with the softened wording on Table 9.

5 Conclusion

We presented CMF, a black-box mixture-factorization method for dataset-level source ranking, alongside SCM, a Shapley-based residual comparison. Across four LLMs and three datasets spanning likely-seen, synthetic-unseen, and post-cutoff exposure, both methods rank sources in the order predicted by our exposure proxies, and CMF additionally recovers a base-model contribution π_0 that correlates with no-context accuracy. Synthetic identifiability, controlled fine-tuning, and multi-checkpoint unlearning support the behavioral source-ranking interpretation.

Limitations

Our work faces several limitations that we summarize here to inform future research and responsible interpretation of our results.

Scale. Our experiments use m=2,000 queries per dataset, sufficient to drive 95% bootstrap CIs of ± 0.02 on CMF attribution (Appendix J), but small compared to recent pretraining-scale TDA work that operates on hundreds of billions of tokens (Chang et al., 2025). Scaling to millions of

queries would test whether CMF's mixture decomposition remains stable when individual sources are themselves heterogeneous corpora rather than topic-coherent collections.

Restriction to binary QA. Our evaluation is restricted to binary classification tasks, which allows precise measurement of correctness and attribution effects. Many real-world applications involve open-ended text generation, summarization, or multi-class classification. The estimators themselves are agnostic to the output space — Eq. (1) and Eq. (5) require only a bounded scalar similarity between outputs — so the restriction is evaluative rather than structural: extending CMF and SCM to these settings requires choosing how to map free-form generations into a scalar comparable across queries, and each choice introduces its own biases; candidate signals include token-level perplexity shifts, log-likelihood ratios, or sequence-level embedding distances. We treat this as the most important methodological extension.

Answer-prior ("Yes") bias. Instruction-tuned LLMs are known to exhibit a prior toward affirmative answers, which could in principle contaminate similarity-based attribution: if the model overproduces "Yes," agreement between no-context and with-context outputs may reflect the shared prior rather than source-specific alignment. Two properties of our design bound this risk. First, within each evaluation panel all $(K+1)$ prompts are issued on the *same* query set, so the gold-label distribution is identical across candidate sources by construction; a global answer prior therefore shifts all sources' observed agreement together and is largely absorbed by the base component π_0 rather than attributed to any individual source. Second, the quantity CMF factorizes is the *residual change* between no-context and with-context outputs, so a prior common to both cancels to first order. The residual concern — content-driven label skew that differs across candidate datasets masquerading as source alignment — is testable from collected outputs via per-dataset label distributions, a majority-label baseline, an output-permutation test, and re-estimation on label-balanced subsets. We note that the permutation test must destroy the query-level correspondence between no-context and with-context outputs (e.g., by shuffling with-context outputs across queries within each source condition while preserving marginal answer frequencies): because the similarity in Eq. (1) is equality-based, a global relabeling of answers leaves it — and hence CMF — exactly unchanged, so a naive global Yes/No flip tests nothing. Under a valid output permutation, source separation should collapse toward the uniform weight $1/K$ if the signal were a label-prior artifact.

Retrieval and chunking. Our RAG setup uses a 512-token chunk size, chosen based on the average passage length in BoolQ ($\sim$108 tokens). Modern LLMs support much larger contexts (32K–300K tokens), which could improve retrieval for datasets with longer passages and may interact with attribution stability. We have not systematically studied how chunk size, top-k, reranking, or long-context strategies affect the CMF–SCM gap.

Construct validity: the metric measures parametric-knowledge alignment, not verified exposure. Both SCM and CMF operate on the residual change in the LLM's output when a candidate dataset is supplied as RAG context. This signal is closer to "how much does the model already produce D_k-aligned answers on Q" than to "did D_k appear verbatim in pretraining." The two are correlated but not identical. We see this directly in our results: FakeQ was constructed by us and is unseen by every evaluated checkpoint, yet CMF assigns it $\widehat{\pi}_{\text{FakeQ}} \approx 0.5$, reflecting topical and structural similarity to BoolQ rather than exposure to FakeQ itself. Membership-inference methods that operate on loss or likelihood signals (Shi et al., 2024a; Maini et al., 2024b) address the strict inclusion question more directly, at the cost of requiring loss/logit access; our behavioral framing is complementary, and we caution against reading CMF/SCM scores as exposure probabilities in absolute terms. The temporal/topical-shift confound described by Duan et al. (2024) for membership-inference benchmarks also applies to our BoolQ/FakeQ/Olympic2024 ordering: the three datasets differ in distribution as well as exposure status, and the metric is sensitive to both.

Modeling assumption in CMF. CMF relies on the assumption (Eq. 4) that providing context c_k collapses the mixture to a base component plus the kth dataset component. The synthetic identifiability experiment (Appendix G) shows that CMF recovers the ground-truth mixture weights under this assumption with L_1 error ≤ 0.10 at noise $\sigma \leq 0.10$, but we cannot directly verify the assumption on real LLMs. Real LLMs may exhibit interactions

between sources (e.g. retrieved context cueing the model to recall pre-trained but unrelated facts) that the additive mixture does not capture.

Phi-3 similarity pipeline. Phi-3-mini did not reliably produce single-token boolean answers, so we computed similarity for its outputs via a zero-shot classifier rather than direct token comparison. The ablation in Appendix I shows that absolute attribution values shift by up to one CI width depending on the similarity pipeline used, but the qualitative dataset and method rankings are preserved. Reported Phi-3 numbers should therefore be interpreted as approximate to within $\sim$0.04 of the values that would be obtained under a uniform pipeline across all four models.

Lack of verified ground truth on closed models. For closed-weight models such as GPT-4, we cannot verify which datasets were actually included in pretraining. We use BoolQ as a proxy for "likely seen" data, FakeQ as "synthetic unseen," and Olympic2024 as "unseen to a high level of confidence" (verifiable for the open-weight models whose pretraining cutoff is published, only inferred for GPT-4). Our validation through controlled fine-tuning on Olympic2024 establishes ground truth for the *exposure-progression* direction (attribution should increase with more training on a target dataset) but not for the absolute mapping from π_k to pretraining inclusion. Absolute attribution values should therefore be interpreted relative to each other within a model rather than as exposure probabilities.

Comparison to ContextCite is task-level, not span-level. ContextCite produces per-span attribution within a single retrieved context. We compare against it by summing per-chunk attributions per source, but a finer-grained study would also compare span-level alignment on tasks where individual spans of context contain the answer. We leave this comparison to future work.

Ethical Considerations

Our attribution methods rely on behavioral proxies and prompt-based perturbations. While scalable and applicable to black-box LLMs, these proxies introduce risks of misinterpretation. Attribution scores may reflect prompt sensitivity or stylistic alignment rather than true data influence — especially when prompts are imperfectly calibrated or when retrieval introduces confounds. We therefore frame attribution scores as *diagnostic signals*, not absolute ground truth.

There is also a risk of overfitting attribution to retrieval artifacts: if retrieval systems consistently surface stylistically similar but semantically irrelevant passages, attribution may overestimate their influence. Careful design of the retrieval pipeline and explicit reporting of retrieval configurations are needed to mitigate this risk.

The Olympic 2024 dataset was constructed from publicly available Kaggle data on the Paris 2024 Olympics. We manually generated Yes/No question-answer pairs for evaluation and split them into training, validation, and test partitions. No personally identifying information was collected, and all source material is publicly available.

Dual-use considerations. Because CMF and SCM require only input–output access, a third party could in principle probe whether a proprietary or sensitive corpus is behaviorally aligned with a deployed closed-weight model, raising privacy and trade-secret concerns for model providers and data owners. Two factors mitigate this risk. First, the construct-validity limits we document are themselves a safeguard: FakeQ receives nontrivial CMF mass despite provably zero exposure, so CMF scores cannot support definitive membership or inclusion claims, capping their evidentiary value in adversarial settings. Second, provider-side defenses are straightforward: per-account prompt-budget limits raise the cost of assembling the $(K+1) \times m$ prompt matrix, and output randomization (temperature, sampling variation across identical prompts) directly degrades the residual-similarity signal both estimators rely on. We encourage providers weighing such defenses to balance them against the legitimate auditing uses (data valuation, unlearning verification, compliance) that motivate this work.

Finally, as attribution becomes a tool for model auditing, data valuation, and copyright analysis, there is a risk of its use being perceived as definitive. Attribution is fundamentally relative and sensitive to framing; results should be combined with complementary evidence (e.g., training-data inventories, memorization audits) before being used to support claims of inclusion, exclusion, or compensation.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Report. *arXiv:2303.08774*.

Sungmin Cha, Sungjun Cho, Dasol Hwang, Honglak Lee, Taesup Moon, and Moontae Lee. 2024. Learning to unlearn: Instance-wise unlearning for pre-trained classifiers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11186–11194.

Tyler A. Chang, Dheeraj Rajagopal, Tolga Bolukbasi, Lucas Dixon, and Ian Tenney. 2025. Scalable influence and fact tracing for large language model pretraining. In *International Conference on Learning Representations (ICLR)*. ArXiv:2410.17413.

Yuanyuan Chen, Boyang Li, Han Yu, Pengcheng Wu, and Chunyan Miao. 2021. Hydra: Hypergradient data relevance analysis for interpreting deep neural networks. In *AAAI*.

Sang Keun Choe, Hwijeen Ahn, Juhan Bae, Kewen Zhao, Minsoo Kang, Youngseog Chung, Adithya Pratapa, Willie Neiswanger, Emma Strubell, Teruko Mitamura, et al. 2024. What is your data worth to gpt? llm-scale data valuation with influence functions. *arXiv preprint arXiv:2405.13954*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *NAACL-HLT*.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Madry. 2024. ContextCite: Attributing model generation to context. In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2409.00729.

Yibo Deng et al. 2025. DATE-LM: Benchmarking data attribution evaluation for large language models. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*. ArXiv:2507.09424.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. 2024. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*.

Vitaly Feldman and Chiyuan Zhang. 2020. What neural networks memorize and why: Discovering the long tail via influence estimation. *NeurIPS*, 33:2881–2891.

Amirata Ghorbani and James Zou. 2019. Data shapley: Equitable valuation of data for machine learning. In *International conference on machine learning*, pages 2242–2251. PMLR.

Antonio Ginart, Melody Guan, Gregory Valiant, and James Y Zou. 2019. Making ai forget you: Data deletion in machine learning. *NeurIPS*.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9304–9312.

Shahriar Golchin and Mihai Surdeanu. 2024. Time travel in LLMs: Tracing data contamination in large language models. In *International Conference on Learning Representations (ICLR)*. ArXiv:2308.08493.

Jochen Gorski, Frank Pfeuffer, and Kathrin Klamroth. 2007. Biconvex sets and optimization with biconvex functions: a survey and extensions. *Mathematical methods of operations research*, 66(3):373–407.

Zayd Hammoudeh and Daniel Lowd. 2024. Training data influence analysis and estimation: A survey. *Machine Learning*, pages 1–53.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7(3):535–547.

Kaggle. 2024. Paris 2024 olympic summer games.

Pang Wei Koh and Percy Liang. 2017. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR.

Yongchan Kwon, Eric Wu, Kevin Wu, and James Zou. 2024. DataInf: Efficiently estimating data influence in LoRA-tuned LLMs and diffusion models. In *International Conference on Learning Representations (ICLR)*.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *arXiv:2211.05100*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *NeurIPS*.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, et al. 2024. The WMDP benchmark: Measuring and reducing malicious use with unlearning. In *International Conference on Machine Learning (ICML)*. ArXiv:2403.03218.

Jiacheng Liu, Sewon Min, Pang Wei Koh, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2024a. Infini-gram: Scaling unbounded n-gram language models to a trillion tokens. In *First Conference on Language Modeling (COLM)*. ArXiv:2401.17377.

Jiacheng Liu, Sewon Min, Luke Zettlemoyer, Yejin Choi, and Hannaneh Hajishirzi. 2025. OLMoTrace: Tracing language model outputs back to trillions of training tokens. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL)*. ArXiv:2504.07096.

Jiancheng Liu, Parikshit Ram, Yuguang Yao, Gaowen Liu, Yang Liu, PRANAY SHARMA, Sijia Liu, et al. 2024b. Model sparsity can simplify machine unlearning. *Advances in Neural Information Processing Systems*, 36.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C. Lipton, and J. Zico Kolter. 2024a. TOFU: A task of fictitious unlearning for LLMs. In *First Conference on Language Modeling (COLM)*. ArXiv:2401.06121.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. 2024b. LLM dataset inference: Did you train on my dataset? In *Advances in Neural Information Processing Systems (NeurIPS)*. ArXiv:2406.06443.

Elisa Nguyen, Minjoon Seo, and Seong Joon Oh. 2023. A bayesian perspective on training data attribution. *arXiv preprint arXiv:2305.19765*.

Ki Nohyun, Hoyong Choi, and Hye Won Chung. 2022. Data valuation without training of a model. In *The Eleventh International Conference on Learning Representations*.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori B. Hashimoto. 2024. Proving test set contamination in black-box language models. In *International Conference on Learning Representations (ICLR)*. ArXiv:2310.17623.

Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. 2023. Trak: attributing model behavior at scale. In *ICML*.

Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. 2020. Estimating training data influence by tracing gradient descent. *NeurIPS*, 33:19920–19930.

Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. 2024. Model internals-based answer attribution for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ArXiv:2406.13663.

Yulei Qin, Yuncheng Yang, Pengcheng Guo, Gang Li, Hang Shao, Yuchen Shi, Zihan Xu, Yun Gu, Ke Li, and Xing Sun. 2025. Unleashing the power of data tsunami: A comprehensive survey on data assessment and selection for instruction tuning of language models. *TMLR*. Survey Certification.

Ayush Sekhari, Jayadev Acharya, Gautam Kamath, and Ananda Theertha Suresh. 2021. Remember what you want to forget: Algorithms for machine unlearning. *NeurIPS*, 34:18075–18086.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024a. Detecting pretraining data from large language models. In *International Conference on Learning Representations (ICLR)*. ArXiv:2310.16789.

Weijia Shi, Jaechan Wu, Anirudh Ramesh, Stephanie Cardei, Felix Yang, Yujia Zhao, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024b. MUSE: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460*.

Robert J Tibshirani and Bradley Efron. 1993. An introduction to the bootstrap. *Monographs on statistics and applied probability*, 57(1):1–436.

Theodora Worledge, Judy Hanwen Shen, Nicole Meister, Caleb Winston, and Carlos Guestrin. 2024. Unifying corroborative and contributive attributions in large language models. In *2024 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)*, pages 665–683. IEEE.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. 2024. LESS: Selecting influential data for targeted instruction tuning. In *ICML*.

Ruiqi Zhang, Licong Chen, Junxiang Lou, Yujia Hu, Yexin Wang, Lichao Bao, Yang Liu, and Mei Wei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. In *First Conference on Language Modeling (COLM)*. ArXiv:2404.05868.

A Algorithms

We give the full pseudo-code for the two methods introduced in §2.

Input: Instruction-tuned LLM M; queries $Q = \{q_1, \ldots, q_m\}$; datasets $D_1, \ldots, D_K$.

for $q \in Q$ **do**
  Compute output without context: $y = M(q)$;
  for $S \subseteq \{D_1, \ldots, D_K\} \setminus \{D_k\}$ **do**
    Use RAG to create contexts c_S and $c_{S \cup \{D_k\}}$ from the datasets in S and D_k;
    Compute $y \mid c_S = M(q \mid c_S)$ and $y \mid c_{S \cup \{D_k\}} = M(q \mid c_{S \cup \{D_k\}})$;
    Compute similarities s_S and $s_{S \cup \{D_k\}}$;
  end
  Compute $\phi_k(q)$ via Eq. (2);
end

Result: Return the average ϕ_k over the m queries.

Algorithm 1: Shapley Context Method (SCM)

Input: Instruction-tuned LLM M; queries $Q = \{q_1, \ldots, q_m\}$; datasets $D_1, \ldots, D_K$.

for $q \in Q$ **do**
  Compute output without context: $p(y \mid q, c_0) = M(q)$;
  for $k = 1, \ldots, K$ **do**
    Use RAG to create context c_k from dataset D_k;
    Compute the output with context: $p(y \mid q, c_k) = M(q \mid c_k)$;
  end
end
Build matrix P with $P_{k,j} = p(y \mid q_j, c_k)$;
Solve Eq. (6) via alternating projected least squares to obtain $\widehat{\boldsymbol{\pi}}$;

Result: Return the contribution vector $\widehat{\boldsymbol{\pi}}$.

Algorithm 2: Context Mixture Factorization (CMF)

B Prompts

General Question Prompt: *"Read the context provided and answer the following question: {question}"*

Contextual Understanding Prompt: *"Based on the information in the context, what can you conclude about the following question? {question}"*

Summarization Prompt: *"After considering the context below, summarize your answer to this question: {question}"*

Opinion-Based Prompt: *"Given the details in the context, what is your opinion on the following question: {question}"*

Detail Extraction Prompt: *"Extract relevant information from the context to answer this question: {question}"*

Fact-Checking Prompt: *"Using the context provided, verify the accuracy of the following statement: {question}"*

Table 2 shows the average accuracy computed by comparing predictions with the BoolQ ground truth.

C FakeQ construction

FakeQ is a synthetic perturbation of BoolQ designed to preserve topical distribution and template structure while removing surface-text overlap with the original BoolQ corpus.

Generation procedure. Starting from the BoolQ validation split, for each (question, passage, boolean answer) triple we apply the following pipeline: (i) we identify all named entities in the question and passage using spaCy's `en_core_web_lg` model; (ii) we replace each entity with a randomly sampled entity of the same type from a held-out pool drawn from Wikipedia (e.g., person $\rightarrow$ person, location $\rightarrow$ location, date $\rightarrow$ date); (iii) we re-render the question and passage with the substituted entities; (iv) we paraphrase non-entity content with an instruction-tuned model (Mistral-7B-Instruct, temperature 0.0, prompt: "Paraphrase the following passage while preserving meaning and the Yes/No answer to the question:"); (v) we discard any (question, passage) pair where the new passage no longer entails the original boolean answer.

Label verification. For step (v) above, we automatically check label consistency by prompting GPT-4 with the perturbed (question, passage) and comparing its answer to the original BoolQ label. We retain only pairs where GPT-4 agrees with the original label and confidence is high ("Yes" or "No", not "I don't know"). This filter drops $\sim$22% of candidate pairs. We additionally hand-verify a random sample of 200 retained pairs; 188 (94%) were judged correct by the annotator. Because GPT-

4 is used only as a label-consistency filter during dataset construction and not to generate the model outputs that go into attribution, this step should not give GPT-4 an unfair advantage in the FakeQ attribution evaluation. We nonetheless flag the GPT-4 dependence here, and future work should include independent human verification at larger scale or use a labeller drawn from a different model family.

Leakage check. Because BoolQ is publicly available, a model that has memorized BoolQ could in principle recognize the topical structure of a FakeQ question. To bound this, we measure word-level n-gram overlap between each FakeQ passage and the closest BoolQ passage. The mean 5-gram overlap is 0.03 (vs. 0.71 for the unperturbed BoolQ-against-BoolQ baseline), supporting the claim that the surface text differs substantially from any BoolQ training instance.

Generator-style caveat. We acknowledge a potential circularity: the FakeQ paraphraser is Mistral-7B-Instruct-v0.2, which is also one of the four evaluated LLMs. Stylistic alignment between Mistral-generated FakeQ contexts and Mistral's own outputs could in principle inflate Mistral's similarity-based attribution score on FakeQ relative to other LLMs. The 5-gram leakage check above bounds surface-level overlap with BoolQ but does not address this generator-style confound. The Panel B numbers themselves, however, argue against a material effect: generator-style bias would predict Mistral elevated *above* the other models on FakeQ, whereas Mistral's CMF value (0.51) sits strictly inside the cross-model range (0.48–0.53), with GPT-4 — which played no role in FakeQ construction — highest at 0.53. Re-generating FakeQ with a paraphraser outside the evaluation pool and re-running the FakeQ row of Table 1 remains a planned robustness check, and we flag this dependency for transparency.

Scale. The final FakeQ corpus contains 2,000 (question, passage, answer) triples, matching the BoolQ and Olympic2024 evaluation set sizes used in the main paper.

Example. *Original BoolQ:* "Is the Eiffel Tower located in Paris?" / "The Eiffel Tower (French: tour Eiffel) is a wrought-iron lattice tower on the Champ de Mars in Paris, France." / Yes. *Perturbed FakeQ:* "Is the Burj Khalifa located in Hanoi?" / "The Burj Khalifa (Vietnamese: tháp Burj Khalifa) is a stainless-steel skyscraper on the Thánh Long peninsula in Hanoi, Vietnam." / Yes. (The label is preserved because the new passage entails the new question.)

D Olympic 2024 Dataset

The Olympic 2024 dataset was constructed by sourcing textual snippets from publicly available Kaggle datasets (Kaggle, 2024) related to the Paris 2024 Olympics. Since no structured QA dataset existed for this event, question-answer pairs were manually generated for each context snippet to create a well-defined evaluation framework. For each retrieved context, three binary (Yes/No) question-answer pairs were created: one for training, one for validation, and one for testing. The training set was used to fine-tune the models, allowing them to incorporate Olympic-related information. The validation set was used to evaluate attribution metrics, ensuring that the methods assessed data influence without direct exposure during training. The test set was reserved for evaluating generalization after fine-tuning and unlearning experiments. This design follows the structure of BoolQ, making it easier to analyze how fine-tuning impacts attribution and how well unlearning methods reduce the model's reliance on the dataset.

E Effectiveness of RAG

To evaluate the effectiveness of context provision via RAG, we measured the accuracy of various LLMs when answering questions from the BoolQ dataset. The experiment compared performance across different scenarios: without any context, with only the BoolQ context, with contexts from five other datasets, and with all datasets combined. Results are summarized in Table 2.

Table 2: Accuracy of LLMs with different contexts.

Context	Bloomz	GPT-4	Mistral	Phi-3
No Context	0.43	0.73	0.68	0.70
BoolQ as RAG	0.74	0.87	0.85	0.82
Five Datasets Only	0.35	0.64	0.60	0.45
All Data + BoolQ	0.73	0.84	0.82	0.83

The baseline (No Context) setting reveals inherent differences in the LLMs' capabilities. GPT-4 has the highest baseline accuracy at 0.73, followed by Phi-3 at 0.70 and Mistral 7B at 0.68, indicating robust pretraining. Bloomz shows lower accuracy at 0.43, highlighting its dependency on contextual data. When BoolQ context is provided

via RAG, all models show significant accuracy improvements. Providing context from five datasets (excluding BoolQ) leads to accuracy drops, confirming that less relevant data are less helpful for BoolQ queries. Combining all datasets with the BoolQ context yields slight decreases for GPT-4 and Mistral 7B, suggesting that additional datasets introduce noise.

F Per-source attribution breakdowns

Table 3 reports CMF's full raw mixture weights on the BoolQ evaluation (discussed in §4); Tables 4–5 report the companion SCM and TRAK per-source attribution on BoolQ. Tables 6 and 7 report the same for FakeQ; Tables 8 and 9 for Olympic2024.

Table 3: Raw CMF mixture weights $\boldsymbol{\pi}$ on the BoolQ evaluation (columns sum to 1.00 by the simplex constraint in Eq. 6). The last row reports the source-normalized BoolQ weight $\widehat{\pi}_{\text{BoolQ}}$ used in Table 1.

Source	Bloomz	GPT-4	Mistral	Phi-3
π_{Base}	0.05	0.08	0.06	0.05
π_{BoolQ}	0.60	0.62	0.61	0.59
$\pi_{\text{Chemistry}}$	0.09	0.06	0.07	0.08
$\pi_{\text{Natural Sci}}$	0.07	0.06	0.07	0.07
π_{History}	0.08	0.07	0.07	0.07
π_{Biology}	0.06	0.05	0.06	0.07
π_{Law}	0.05	0.06	0.06	0.07
Column sum	**1.00**	**1.00**	**1.00**	**1.00**
$\widehat{\pi}_{\text{BoolQ}} = \frac{\pi_{\text{BoolQ}}}{1-\pi_{\text{Base}}}$	0.63	0.67	0.65	0.62

Table 4: SCM Shapley values ϕ_k on BoolQ.

Source	Bloomz	GPT-4	Mistral 7B	Phi-3
ϕ_{BoolQ}	0.48	0.59	0.57	0.50
$\phi_{\text{Chemistry}}$	0.10	0.08	0.09	0.10
$\phi_{\text{Natural Sci}}$	0.12	0.09	0.10	0.11
ϕ_{History}	0.11	0.10	0.10	0.11
ϕ_{Biology}	0.10	0.07	0.08	0.10
ϕ_{Law}	0.09	0.07	0.08	0.08

Table 5: TRAK scores per source for Phi-3 and Bloomz on BoolQ.

Source	Phi-3	Bloomz
BoolQ	0.58	0.61
Chemistry	−0.08	−0.10
Natural Science	0.20	0.22
History	0.18	0.19
Biology	−0.05	−0.06
Law	0.07	0.09

FakeQ per-source breakdown. Attribution values for FakeQ are noticeably lower than for BoolQ. For example, ϕ_{FakeQ} ranges from 0.28 to 0.36 across the LLMs, compared to 0.48 to 0.59 for ϕ_{BoolQ}. CMF assigns higher scores than SCM, with $\frac{\pi_{\text{FakeQ}}}{1-\pi_{\text{Base}}}$ ranging from 0.48 to 0.53, consistent with CMF detecting residual structural similarity between FakeQ and BoolQ.

Table 6: SCM Shapley values ϕ_k on FakeQ.

Source	Bloomz	GPT-4	Mistral 7B	Phi-3
ϕ_{FakeQ}	0.32	0.36	0.33	0.28
$\phi_{\text{Chemistry}}$	0.08	0.06	0.07	0.08
$\phi_{\text{Natural Sci}}$	0.09	0.07	0.08	0.09
ϕ_{History}	0.08	0.07	0.08	0.08
ϕ_{Biology}	0.07	0.05	0.06	0.07
ϕ_{Law}	0.06	0.05	0.06	0.06

Table 7: Raw CMF mixture weights $\boldsymbol{\pi}$ on FakeQ (columns sum to 1.00). The last row reports the source-normalized FakeQ weight $\widehat{\pi}_{\text{FakeQ}}$ used in Table 1.

Source	Bloomz	GPT-4	Mistral	Phi-3
π_{Base}	0.06	0.09	0.07	0.06
π_{FakeQ}	0.46	0.48	0.47	0.45
$\pi_{\text{Chemistry}}$	0.10	0.09	0.10	0.10
$\pi_{\text{Natural Sci}}$	0.10	0.08	0.09	0.10
π_{History}	0.10	0.09	0.09	0.10
π_{Biology}	0.08	0.09	0.09	0.09
π_{Law}	0.10	0.08	0.09	0.10
Column sum	**1.00**	**1.00**	**1.00**	**1.00**
$\widehat{\pi}_{\text{FakeQ}} = \frac{\pi_{\text{FakeQ}}}{1-\pi_{\text{Base}}}$	0.49	0.53	0.51	0.48

Olympic2024 per-source breakdown. Olympic2024 attribution values are significantly lower than both BoolQ and FakeQ; $\phi_{\text{Olympic2024}}$ ranges from 0.07 to 0.11. CMF assigns lower attribution to Olympic2024 than to BoolQ/FakeQ, with $\frac{\pi_{\text{Olympic2024}}}{1-\pi_{\text{Base}}}$ ranging from 0.15 to 0.20 — slightly above SCM, indicating CMF retains some signal even in the post-cutoff regime.

Table 8: SCM Shapley values ϕ_k on Olympic2024.

Source	Bloomz	GPT-4	Mistral 7B	Phi-3
$\phi_{\text{Olympic2024}}$	0.08	0.11	0.09	0.07
$\phi_{\text{Chemistry}}$	0.04	0.03	0.04	0.05
$\phi_{\text{Natural Sci}}$	0.05	0.04	0.04	0.05
ϕ_{History}	0.05	0.04	0.05	0.05
ϕ_{Biology}	0.04	0.03	0.03	0.04
ϕ_{Law}	0.03	0.02	0.03	0.03

G Synthetic identifiability experiment

To test whether CMF actually recovers the mixture weights $\boldsymbol{\pi}$ under the modeling assumption

Table 9: Raw CMF mixture weights $\boldsymbol{\pi}$ on Olympic2024 (columns sum to 1.00). The last row reports the source-normalized Olympic2024 weight $\widehat{\pi}_{\text{Olympic2024}}$ used in Table 1. On this post-cutoff dataset no candidate source matches the model's parametric knowledge, so the mixture distributes diffusely; for bloomz and Phi-3 the target weight is comparable to individual control weights (0.14–0.16 vs. 0.15–0.16), and only for GPT-4 and Mistral 7B is the target weight clearly above individual controls. The source-normalized ratio in the bottom row is correspondingly close to the $1/(K_{\text{src}}{+}1) \approx 0.143$ uniform baseline induced by the entropy regularizer at $\lambda{=}10^{-1}$. We interpret the Olympic2024 column primarily as a control that anchors the lower end of the exposure continuum, rather than as an attribution signal that meaningfully distinguishes target from controls.

Source	**Bloomz**	**GPT-4**	**Mistral 7B**	**Phi-3**
π_{Base}	0.08	0.11	0.09	0.08
$\pi_{\text{Olympic2024}}$	0.15	0.18	0.16	0.14
$\pi_{\text{Chemistry}}$	0.16	0.15	0.15	0.16
$\pi_{\text{Natural Sci}}$	0.15	0.14	0.15	0.16
π_{History}	0.16	0.15	0.15	0.16
π_{Biology}	0.15	0.14	0.15	0.15
π_{Law}	0.15	0.13	0.15	0.15
Column sum	**1.00**	**1.00**	**1.00**	**1.00**
$\widehat{\pi}_{\text{Olympic2024}} = \frac{\pi_{\text{Olympic2024}}}{1-\pi_{\text{Base}}}$	0.16	0.20	0.18	0.15

in Eq. (4), we construct a synthetic experiment with known ground truth. We fix $K{=}3$ candidate sources plus a base component and set ground-truth weights, written in source-first order to match Table 10, $(\pi_1^*, \pi_2^*, \pi_3^*, \pi_0^*) = (0.40, 0.30, 0.20, 0.10)$. We sample $m{=}200$ latent "query truth" values $\widetilde{p}_k(y \mid q) \in (0, 1)$ uniformly per source and form $P = \Pi\widetilde{P} + \mathcal{N}(0, \sigma^2)$ following Eq. (4). We run CMF and SCM with 100 random initializations per setting and report mean recovered $\hat{\boldsymbol{\pi}}$ and L_1 error.

CMF recovers the rank-ordering of sources at all noise levels we tested. The L_1 recovery error grows roughly linearly in σ but remains substantially smaller than that of SCM under the same simulation (a $2\times$ to $3\times$ gap), supporting the use of CMF as the primary estimator and SCM as a simpler comparison method. The relative ordering of the recovered $\hat{\pi}_k$ is preserved across all 100 initializations for $\sigma \leq 0.10$.

G.1 Scaling the number of candidate sources

The identifiability experiment above fixes $K{=}3$. Because CMF's prompt cost is linear in K ($K{+}1$ prompts per query), the binding constraint at larger source counts is identifiability rather than budget: condition (a) in §2 predicts degradation when candidate sources become mutually indistinguishable on the query set. Two failure regimes follow from the structure of Eq. (5). First, as K grows with mutually distinguishable sources, each additional source contributes its own with-context row to P, so pairwise rank agreement between recovered and true weights should degrade gracefully; exact recovery of the *full* source ordering, by contrast, requires all $\binom{K}{2}$ pairwise comparisons to resolve correctly simultaneously, and should therefore collapse much earlier than pairwise agreement — fine-grained total orders over many sources are the first casualty of scale. Second, when sources share latent content, the corresponding rows of $\widetilde{P}$ become nearly collinear and the factorization approaches the unidentifiable regime, inflating both recovery error and rank inversions regardless of the query budget m; this failure mode is governed by source similarity, not by K itself. Both regimes can be probed with the synthetic protocol above — (i) scaling $K \in \{3, 5, 10, 20, 50\}$ with independent sources (ground-truth weights proportional to $K{+}2{-}k$ with $\pi_0^*{=}0.10$, reducing to the setting above at $K{=}3$), measuring L_1 error, Kendall's τ, and exact-order recovery; and (ii) fixing $K{=}10$ while interpolating each source's latent query-truth vector toward a shared common component, $\widetilde{p}_k \leftarrow (1{-}\rho)\,\widetilde{p}_k + \rho\,\widetilde{p}_{\text{common}}$ — and ρ sweeps directly stress condition (a). The same protocol extends to structured misspecification: generating P with cross-source interaction terms (violating the additivity of Eq. 3, per Remark 1) and fitting the additive CMF quantifies the estimator's behavior when its modeling assumption fails, complementing the additive-noise robustness in Table 10. Practically, we recommend (i) checking the empirical distinguishability of candidate sources (e.g., the conditioning of the with-context block of P) before trusting fine-grained rankings at large K, and (ii) reading CMF output at large K as a coarse partition into high/medium/low-attribution tiers rather

Table 10: Synthetic recovery of source-first weights $(\pi_1^*, \pi_2^*, \pi_3^*, \pi_0^*) = (0.40, 0.30, 0.20, 0.10)$ from $P = \Pi\widetilde{P} + \mathcal{N}(0, \sigma^2)$. Values are means over 100 random initializations (standard deviations in parentheses). The final two columns report L_1 recovery error for CMF and SCM under the same simulated data.

σ	$\hat{\pi}_1$ (true 0.40)	$\hat{\pi}_2$ (true 0.30)	$\hat{\pi}_3$ (true 0.20)	$\hat{\pi}_0$ (true 0.10)	L_1 err. (CMF)	L_1 err. (SCM)
0.00	0.41 (0.01)	0.29 (0.01)	0.21 (0.01)	0.09 (0.01)	0.04	0.08
0.05	0.40 (0.02)	0.30 (0.02)	0.22 (0.02)	0.08 (0.02)	0.06	0.12
0.10	0.38 (0.04)	0.31 (0.04)	0.23 (0.04)	0.08 (0.04)	0.10	0.22
0.20	0.35 (0.07)	0.32 (0.07)	0.22 (0.07)	0.11 (0.06)	0.17	0.38

than as a total order. Our main experiments use K_{src}=6, a moderate source count; the synthetic results above establish recovery directly only at K=3, and the considerations here caution against extrapolating fine-grained total orders to substantially larger K.

H Regularization sensitivity

To justify the choice of regularization coefficients $\lambda_\pi = \lambda_{\widetilde{P}}$ used throughout the paper, we sweep over six values and measure the spread of recovered $\hat{\boldsymbol{\pi}}$ across 100 random initializations on the same synthetic instance as Appendix G with σ=0.05. We report the trace of the covariance matrix of $\hat{\boldsymbol{\pi}}$ as a single-scalar stability measure; lower is better.

Table 11: Stability of recovered $\hat{\boldsymbol{\pi}}$ across random initializations as a function of regularization strength $\lambda = \lambda_\pi = \lambda_{\widetilde{P}}$.

λ	$\mathrm{tr}\big(\mathrm{Cov}(\hat{\boldsymbol{\pi}})\big)$	Note
0	0.184	Many minima; solutions collapse
10^{-4}	0.121	Still unstable
10^{-3}	0.058	Improving
10^{-2}	0.019	Stable
10^{-1}	0.006	**Used throughout paper**
1	0.003	Over-regularized: $\hat{\boldsymbol{\pi}} \to \mathbf{1}/(K{+}1)$

The solution becomes effectively stable across initializations at $\lambda \geq 10^{-2}$. We use $\lambda = 10^{-1}$ as a conservative default that retains stability without collapsing the recovered $\boldsymbol{\pi}$ toward the uniform prior. At $\lambda = 1$ the recovered weights converge to within 0.02 of $1/(K{+}1)$, confirming that the regularizer dominates the data term in that regime.

I Phi-3 similarity pipeline ablation

Phi-3-mini did not reliably produce single-token boolean responses (§3.1), so the main results use a zero-shot classifier to map its outputs to {Yes, No, IDK} before computing similarity. To check that this pipeline choice does not drive the reported attribution rankings, we re-run SCM and CMF on Phi-3 with three alternative similarity pipelines: (i) the zero-shot classifier (default), (ii) direct token forced-choice (constraining decoding to the Yes/No/IDK tokens), and (iii) embedding cosine similarity between the model's free-form output and the reference answers.

Absolute attribution shifts by up to ± 0.04 depending on the similarity pipeline, but the BoolQ>Olympic2024 ordering and the CMF>SCM ordering are preserved under all three pipelines. We conclude that the Phi-3 pipeline choice affects absolute scores by approximately one CI width but does not flip qualitative conclusions.

J Scaling and convergence diagnostic

We measure how the CMF estimate of π_{BoolQ} for Bloomz stabilizes as the number of queries m increases.

The point estimate stabilizes by m=500 and the CI half-width drops below 0.02 by m=2,000. We use m=2,000 throughout the main paper; this provides a favorable balance between statistical precision and compute cost, and matches the precision needed to distinguish the reported CMF–SCM gaps.

K Full unlearning results

We report CMF and SCM attribution to Olympic2024 for both bloomz-1b1 and Phi-3-mini, across three fine-tuning learning rates and three unlearning methods (GA = Gradient Ascent, RL = Random Labels, AS = Adversarial Samples). Lower post-unlearning attribution indicates more effective unlearning.

The qualitative ranking RL $\preceq$ GA $\prec$ AS is preserved across both LLMs, both estimators (SCM and CMF), and all three fine-tuning learning rates. The CMF-SCM ratio is largest in the highest-LR regime, consistent with the main-paper observation that CMF–SCM gaps grow with the strength of the underlying signal.

Table 12: Phi-3 similarity pipeline ablation: SCM and CMF attribution to the target source on BoolQ and Olympic2024 under three pipelines. Values are mean±95% bootstrap CI half-width over m=2,000 queries.

	BoolQ		Olympic2024	
Pipeline	SCM	CMF	SCM	CMF
Zero-shot classifier (default)	0.50±0.04	0.62±0.02	0.07±0.03	0.10±0.02
Direct token forced-choice	0.53±0.04	0.64±0.02	0.09±0.03	0.11±0.02
Embedding cosine	0.46±0.04	0.58±0.03	0.06±0.03	0.09±0.03

Table 13: CMF attribution to BoolQ for Bloomz as a function of query count m, with 95% bootstrap CI half-width over B=1000 resamples.

m	π_{BoolQ}	CI half-width
100	0.61	±0.08
500	0.62	±0.04
2,000	0.63	±0.02
10,000	0.63	±0.01

Table 14: CMF attribution to Olympic2024 for bloomz-1b1 across fine-tuning learning rates and unlearning methods.

FT LR	**Pre-FT**	**Post-FT**	**+GA**	**+RL**	**+AS**
1×10^{-5}	0.16	0.41	0.18	0.16	0.22
5×10^{-5}	0.16	0.62	0.31	0.28	0.38
1×10^{-4}	0.16	0.85	0.34	0.31	0.41

Table 15: SCM attribution to Olympic2024 for bloomz-1b1 across fine-tuning learning rates and unlearning methods.

FT LR	**Pre-FT**	**Post-FT**	**+GA**	**+RL**	**+AS**
1×10^{-5}	0.08	0.32	0.12	0.10	0.15
5×10^{-5}	0.08	0.54	0.21	0.19	0.27
1×10^{-4}	0.08	0.75	0.25	0.22	0.30

Table 16: CMF attribution to Olympic2024 for Phi-3-mini across fine-tuning learning rates and unlearning methods.

FT LR	**Pre-FT**	**Post-FT**	**+GA**	**+RL**	**+AS**
1×10^{-5}	0.10	0.37	0.16	0.15	0.20
5×10^{-5}	0.10	0.58	0.34	0.31	0.40
1×10^{-4}	0.10	0.82	0.37	0.33	0.42

Table 17: SCM attribution to Olympic2024 for Phi-3-mini across fine-tuning learning rates and unlearning methods.

FT LR	**Pre-FT**	**Post-FT**	**+GA**	**+RL**	**+AS**
1×10^{-5}	0.07	0.30	0.13	0.11	0.17
5×10^{-5}	0.07	0.52	0.23	0.20	0.30
1×10^{-4}	0.07	0.72	0.28	0.24	0.33

L Reproducibility details

Model checkpoints. Bloomz: `bigscience/bloomz-1b1` (1.1B parameters). Mistral 7B: `mistralai/Mistral-7B-Instruct-v0.2` (7.3B). Phi-3: `microsoft/Phi-3-mini-4k-instruct` (3.8B). GPT-4: OpenAI API `gpt-4-turbo-2024-04-09` snapshot.

Decoding. Temperature 0.0 (greedy) with `max_new_tokens=8` for the boolean answer plus IDK. For Phi-3 we use `max_new_tokens=64` to capture the longer outputs the zero-shot classifier then maps to {Yes, No, IDK}. No top-p or top-k sampling.

Similarity function. For models that emit single-token answers, $\text{sim}(y_1, y_2) = \mathbb{1}[y_1 = y_2]$ at the token level (Yes/No/IDK). For Phi-3 we apply the zero-shot classifier before this comparison.

Bootstrap. 95% bootstrap CIs use B=1,000 resamples of the query set, fixed random seed (42). Aggregates are means over evaluation examples after the per-query aggregation procedure described in §2.1 (SCM) and §2.2 (CMF).

Hardware. A single AWS EC2 g6.16xlarge (4×L4 24 GB) instance. Total compute: ∼80 GPU-hours for CMF across all (model, dataset) pairs, ∼360 GPU-hours for SCM, ∼60 GPU-hours for TRAK-style scoring.

GPT-4 API budget. GPT-4 calls used approximately 60,000 API requests across all experiments. This figure reflects (i) CMF on the full m=2,000 query set for each of the three datasets (∼42,000 calls; 7 prompts per query), (ii) the no-context prompt and retrieval-cosine baseline (negligible additional calls beyond CMF's no-context prompt), and (iii) ContextCite on a 500-query subsample per dataset (∼96,000 calls of the 64-ablation budget). We did *not* run SCM on GPT-4 (which would have required up to 2^6=64 subset prompts per query,

$\sim$384,000 additional calls) due to API cost. The total is therefore approximately 1.4×10^5 calls; the $\sim$60,000 figure in the headline budget reflects only the CMF and retrieval-cosine portions, with ContextCite on GPT-4 reported separately above.

Software. Python 3.11, PyTorch 2.3, HuggingFace Transformers 4.42, FAISS-CPU 1.8, scikit-learn 1.5 (for the LASSO step inside our ContextCite implementation and the Phi-3 zero-shot classifier).

M Extended deep dive into RAG noise effect

For the BoolQ dataset (Table 18), the per-pair residual $r_{k,S} = s_{S\cup\{D_k\}} - s_S$ varies substantially across LLMs — from -0.32 for Bloomz (strong context reliance and weak parametric overlap) to -0.03 for GPT-4 (consistent with prior exposure); Mistral 7B and Phi-3 show smaller residual magnitudes. We interpret these only within each model's own no-context behavior, not as direct exposure probabilities across models. This cross-model heterogeneity is the dispersion that SCM aggregates into a single Shapley statistic; CMF absorbs it into the explicit π_0 term, which is why CMF's CIs are tighter.

Table 18: Mean similarities s_S and residuals $r_{k,S}$ on BoolQ across different LLMs (standard deviations in parentheses).

LLM	s_S	$s_{S\cup\{D_k\}}$	$r_{k,S}$
Bloomz	0.86 (0.03)	0.54 (0.02)	-0.32 (0.02)
GPT-4	0.76 (0.02)	0.73 (0.02)	-0.03 (0.02)
Mistral 7B	0.73 (0.03)	0.79 (0.03)	0.06 (0.02)
Phi-3	0.60 (0.03)	0.63 (0.02)	0.03 (0.02)

For the FakeQ dataset, Bloomz exhibits a moderately negative residual (-0.16) and a slightly reduced similarity score ($s_S = 0.78$), indicating that although FakeQ is not identical to BoolQ, its design allows the model to relate to it effectively. GPT-4 maintains a low residual (-0.03), reinforcing its robustness on datasets that resemble those seen during training. Mistral 7B and Phi-3 show small positive residuals (0.03 and 0.04, respectively), suggesting that these models benefit from the added FakeQ context while exhibiting less direct alignment compared to BoolQ.

For the Olympic2024 dataset, Bloomz has a less negative residual (-0.13) and a reduced similarity score ($s_S = 0.72$), highlighting limited alignment of this post-cutoff dataset with the model's pre-existing knowledge. GPT-4 continues to display a low residual (-0.03), showing relatively stable behavior even on post-cutoff data. Mistral 7B and Phi-3 exhibit small positive residuals (0.03 and 0.05, respectively); we do not interpret these as evidence of exposure — Olympic2024 post-dates the open-weight checkpoints' training cutoffs — but as a mix of generalization to topically adjacent content, retrieval artifacts, and broad world knowledge already in the model's parametric distribution.

Table 19: Mean similarities s_S and residuals $r_{k,S}$ for FakeQ across different LLMs (standard deviations in parentheses).

LLM	s_S	$s_{S\cup\{D_k\}}$	$r_{k,S}$
Bloomz	0.78 (0.04)	0.62 (0.03)	-0.16 (0.03)
GPT-4	0.69 (0.03)	0.66 (0.02)	-0.03 (0.02)
Mistral 7B	0.65 (0.04)	0.68 (0.03)	0.03 (0.03)
Phi-3	0.58 (0.04)	0.62 (0.03)	0.04 (0.03)

Table 20: Mean similarities s_S and residuals $r_{k,S}$ for Olympic2024 across different LLMs (standard deviations in parentheses).

LLM	s_S	$s_{S\cup\{D_k\}}$	$r_{k,S}$
Bloomz	0.72 (0.05)	0.59 (0.04)	-0.13 (0.04)
GPT-4	0.64 (0.03)	0.61 (0.03)	-0.03 (0.03)
Mistral 7B	0.60 (0.04)	0.63 (0.03)	0.03 (0.03)
Phi-3	0.55 (0.05)	0.60 (0.04)	0.05 (0.04)

N Additional robustness baselines

This appendix reports additional robustness checks: a no-context/with-context agreement baseline, open-weight membership-inference baselines (Min-K% Prob, Min-K%++, LLM Dataset Inference), a paired target-vs-control test on CMF's per-source weights, a cross-method AUC comparison for separating likely-seen from post-cutoff data, and a sanity check against the official `context-cite` implementation. The main-text pointer is at the end of §4.5.

N.1 Agreement baseline

We define the per-query *agreement rate* as the fraction of queries on which the model's no-context answer equals its with-context answer (Table 21). Agreement recovers the same coarse BoolQ>FakeQ>Olympic2024 ordering as CMF, which is consistent with the construct-validity caveat in §5: the ordering reflects parametric answerability, not verified exposure. Agreement is,

however, a single dataset-level stability statistic. It does not recover per-source weights, does not estimate π_{Base}, does not compare a target source against background controls, and does not produce the monotonic source-level response under controlled fine-tuning shown in §4.3 (Fig. 1). CMF's added value is the explicit mixture decomposition and its behavior under controlled exposure.

Table 21: No-context/with-context *agreement baseline*: per-query fraction where the model's no-context and with-context answers agree. Mean ± 95% bootstrap CI half-width over m=2,000 queries.

Dataset	Bloomz	GPT-4	Mistral	Phi-3
BoolQ	0.83±0.02	0.91±0.01	0.87±0.01	0.85±0.02
FakeQ	0.54±0.02	0.62±0.02	0.58±0.02	0.56±0.02
Olympic2024	0.43±0.03	0.51±0.03	0.47±0.03	0.45±0.03

N.2 Open-weight membership-inference baselines

We evaluate Min-K% Prob (Shi et al., 2024a), its Min-K%++ refinement, and the LLM Dataset Inference test of Maini et al. (2024b) on the three open-weight models in our study. None of these methods is applicable to GPT-4 under our closed-weight API setting because they require token log-probabilities or full per-token likelihoods that the chat API does not expose for arbitrary continuations. Table 22 reports AUC for distinguishing BoolQ (likely-seen) from Olympic2024 (post-cutoff) per model; CMF's AUC under the same task is reported for reference. Min-K% and Min-K%++ recover a weaker membership signal than CMF across all three open-weight models. Table 23 reports the LLM Dataset Inference p-value per (model, dataset); the test flags BoolQ in every case and fails to reject the null for FakeQ and Olympic2024, qualitatively agreeing with CMF.

Table 22: AUC for distinguishing BoolQ from Olympic2024 on open-weight models, K=20% for Min-K% variants. Mean ± 95% bootstrap CI half-width over m=2,000 queries.

Method	Bloomz	Mistral	Phi-3
Min-K% Prob	0.68±0.02	0.61±0.03	0.66±0.02
Min-K%++	0.71±0.02	0.64±0.03	0.69±0.02
CMF (ref.)	**0.86±0.02**	**0.84±0.03**	**0.83±0.03**

N.3 Paired target-vs-control test on CMF

Per query, we compare the CMF target-source weight against the strongest control-source weight using a Wilcoxon signed-rank test (Table 24). On BoolQ and FakeQ the target dominates the strongest control with $p < 0.01$ for every model. On Olympic2024 the separation is significant for GPT-4 (p=0.02) and Mistral 7B (p=0.04) but not for Bloomz (p=0.18) or Phi-3 (p=0.23). We interpret this honestly: the per-source signal is weaker on smaller models and on post-cutoff data, consistent with the softened wording on Table 9. The metric should be read cautiously in this regime, and absolute weights should be compared to within-model controls rather than as exposure probabilities.

Table 23: LLM Dataset Inference (Maini et al., 2024b) p-values on the open-weight models. Lower p indicates stronger evidence against the "not in training" null hypothesis. The test flags BoolQ across all three models and fails to reject for FakeQ and Olympic2024, qualitatively agreeing with CMF.

Dataset	Bloomz	Mistral	Phi-3
BoolQ	<0.001	<0.001	<0.001
FakeQ	0.41 *n.s.*	0.38 *n.s.*	0.45 *n.s.*
Olympic2024	0.52 *n.s.*	0.49 *n.s.*	0.47 *n.s.*

Table 24: Wilcoxon signed-rank test: does the CMF target-source weight exceed the strongest control across queries? *n.s.* marks $p \geq 0.05$.

Dataset	Bloomz	GPT-4	Mistral	Phi-3
BoolQ	<0.001	<0.001	<0.001	<0.001
FakeQ	<0.001	<0.001	<0.001	<0.01
Olympic2024	0.18 *n.s.*	0.02	0.04	0.23 *n.s.*

N.4 Cross-method AUC for seen-vs-post-cutoff separation

Table 25 reports the AUC for separating likely-seen (BoolQ) from post-cutoff (Olympic2024) examples, pooled across queries and averaged over models on which a given method is applicable (CMF/SCM/Agreement/Retrieval-cosine over all four models; Min-K% Prob over the three open-weight models). CMF yields the strongest seen-vs-post-cutoff separation among the tested methods; retrieval-cosine is near chance, which is consistent with the headline observation that retrieval relevance alone does not explain the attribution ordering.

N.5 Sanity check against the official `context-cite` implementation

We re-run ContextCite on Phi-3-mini using the official `context-cite` library released by Cohen-

Table 25: Cross-method AUC for distinguishing likely-seen BoolQ from post-cutoff Olympic2024. Mean ± 95% bootstrap CI half-width.

Method	AUC
CMF (ours)	**0.85±0.02**
SCM (ours)	0.79±0.03
Agreement baseline	0.78±0.02
Min-K% Prob (K=20%)	0.66±0.02
Retrieval-cosine	0.51±0.02

Wang et al. (2024) and aggregate per-chunk attributions to dataset level using the same procedure as our re-implementation. Table 26 shows that the official implementation reproduces the same qualitative ordering on Phi-3-mini (BoolQ<FakeQ<Olympic2024 for ContextCite, i.e., the predicted inversion relative to CMF/SCM), with all differences within bootstrap uncertainty. The Olympic2024 inversion reported in Table 1 is therefore not an artifact of our adaptation.

Table 26: ContextCite source-level attribution on Phi-3-mini: our re-implementation vs. the official `context-cite` library. Mean ± 95% bootstrap CI half-width.

Dataset	Our reimpl.	Official	Δ
BoolQ	0.46±0.03	0.48±0.03	+0.02
FakeQ	0.54±0.03	0.55±0.03	+0.01
Olympic2024	0.66±0.03	0.69±0.03	+0.03