Data Attribution in Large Language Models via Bidirectional Gradient Optimization

Frédéric Berdoz, Luca A. Lanzendörfer, Kaan Bayraktar, Roger Wattenhofer

ETH Zurich

{fberdoz, lanzendoerfer, kbayraktar, wattenhofer}@ethz.ch

Abstract

Large Language Models (LLMs) are increasingly deployed across diverse applications, raising critical questions for governance, accountability, and data provenance. Understanding which training data most influenced a model's output remains a fundamental open problem. We address this challenge through training data attribution (TDA) for auto-regressive LLMs by expanding upon the inverse formulation: How would training data be affected if the model had seen the generated output during training? Our method perturbs the base model using bidirectional gradient optimization (gradient ascent and descent) on a generated text sample and measures the resulting change in loss across training samples. Our framework supports attribution at arbitrary data granularity, enabling both factual and stylistic attribution. We evaluate our method against baselines on pretrained models with known datasets, and show that it outperforms previous work on influence metrics, thereby enhancing model interpretability, an essential requirement for accountable AI systems.

Code — https://github.com/ETH-DISCO/DABGO

1 Introduction

Large language models (LLMs) have emerged as transformative tools capable of performing a wide range of tasks, from drafting legal documents (Siino et al. 2025), providing medical diagnoses (Zhou et al. 2025), to assisting scientific research (Luo et al. 2025). Although LLMs are increasingly equipped with external tools to search the Internet and retrieve information, their value remains primarily rooted in their knowledge and creativity acquired during training on vast text corpora. Consequently, current research primarily focuses on developing more effective models given a dataset, whereas the inverse problem of identifying which part of that dataset was most influential given an LLM output remains underexplored, despite its potential to enable traceable and auditable AI systems. Besides improving interpretability, such capabilities would close the feedback loop between training data and model behavior, enabling a wide range of applications, such as model debugging and unlearning (Tanno et al. 2022) and data valuation (Sim, Xu, and Low 2022). We introduce **D**ata **A**ttribution through **B**idirectional **G**radient **O**ptimization (DABGO). This framework estimates data influence by comparing the training loss between two models optimized with respect to the generated text, one model optimized via gradient ascent and the other model optimized via gradient descent. This allows influence to be attributed at any level of data granularity. To validate DABGO, we conduct experiments on open-ended text generation in both factual and stylistic settings, demonstrating that our approach outperforms prior attribution methods. In summary, our contributions are as follows:

- We introduce DABGO, a simple attribution method for open-ended text generation using auto-regressive LLMs.
- We quantitatively demonstrate that our approach significantly outperforms recent attribution baselines.
- We qualitatively demonstrate that DABGO captures both factual content and stylistic characteristics in attributed texts, leveraging the interpretability of our method.

2 Related Work

2.1 Training Data Attribution

Training data attribution (TDA) quantifies how much a given training example influenced a model's prediction on a test input. The gold-standard definition is the counterfactual effect of removing a sample from the training set, but exact computation requires retraining for each sample (Cook 1977) or averaging over all subsets (Ghorbani and Zou 2019) and is therefore infeasible. A wide range of different approaches have been proposed to estimate the influence of samples without retraining, such as datamodels (Ilyas et al. 2022), simulators of alternative training runs (Guu et al. 2023) and checkpoint based influence estimation (Pruthi et al. 2020). Most modern TDA methods rely on the concept of *influence functions* from classical robust statistics (Hampel 1974). Influence functions estimate the effect of an infinitesimal change in the weight of any training sample and can be computed in closed form via a Hessian-adjusted dot product between model gradients of the training and test samples (Koh and Liang 2017). Although promising, Schioppa et al. (2023) mention some practical and fundamental limitations of traditional influence function estimation.

Presented at the AI Governance (AIGOV) Workshop at AAAI 2026.

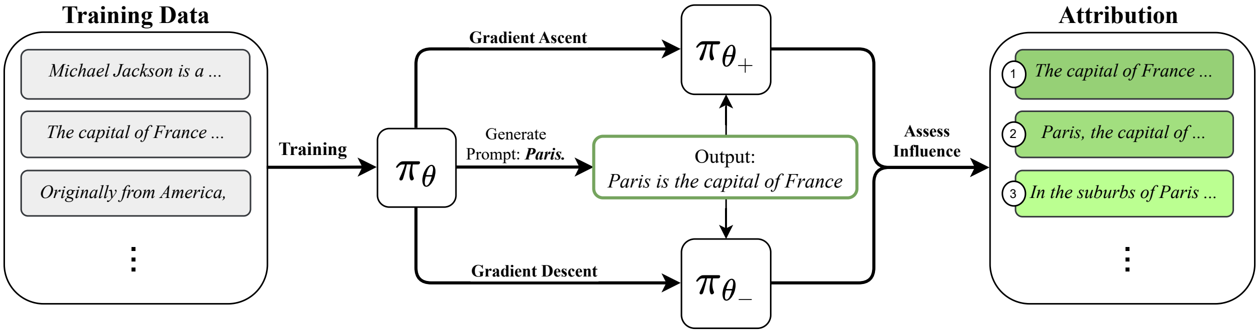

Figure 1: **Overview of DABGO.** We propose a bidirectional attribution technique for training data attribution in LLMs. Starting with a model π_θ trained from scratch, we generate an output sentence from a minimal prompt (e.g., "Paris."). We then apply bidirectional gradient optimization (descent and ascent) on the generated output to obtain two optimized models π_{θ_-} and π_{θ_+}, respectively. To assess the influence of each training sample, we compute its loss under both optimized models and rank samples by the absolute change in loss. This yields a ranking of the most influential training samples responsible for the generated text.

2.2 Forward Influence Function Estimation

Due to the high memory and compute cost associated with inverting the Hessian matrix of a model, most modern TDA methods introduce efficient approximation techniques. (Schioppa et al. 2022; Arnoldi 1951; Agarwal, Bullins, and Hazan 2017; Park et al. 2023; Grosse et al. 2023; Kwon et al. 2024; Chang et al. 2025). Among these, we primarily compare to Chang et al. (2025), who demonstrated promising performance in pretraining data attribution. Their method, TRACKSTAR, utilizes randomly projected gradients to approximate the Hessian.

2.3 Backward Influence Function Estimation

A recent and promising line of work exploits *the mirrored influence hypothesis* formulated by Ko et al. (2024), which states that the influence of a training sample over a model prediction can be accurately estimated by the influence of that same model prediction over the original training example. Reverse influence is typically estimated by unlearning the model on the generated output and measuring the change in training loss between the original and unlearned models. To focus on the parameters most critical to the generated output, the unlearning step is usually performed using a Fisher-regularized gradient ascent step, a method known as machine unlearning (Bourtoule et al. 2021). This approach has the advantage of not using training sample gradients, which greatly improves scalability (Ko et al. 2024), and has been applied in various modalities, including text (Isonuma and Titov 2024), vision (Wang et al. 2024), and audio (Choi et al. 2025). However, prior natural language processing (NLP) studies either restrict themselves to fact tracing using curated facts with single-word predictions (Ko et al. 2024) or operate only at dataset-level granularity (Isonuma and Titov 2024). In contrast, we adapt and extend this line of work to fully open-ended text generation, and we introduce a key refinement: we perform both Fisher-regularized gradient ascent and descent, which we find improves attribution.

3 Methodology

We propose a framework for estimating backward influence functions that enables tracing generated model outputs back to their most influential training samples. Let $\mathcal{D} = \{x^i\}_{i=1}^N$ be the training dataset, where each $x^i = x_0^i x_1^i ... x_{L-1}^i \in \mathcal{X}^L$ represents a contiguous segment of text in the training corpus, L the context window used during training, and $\mathcal{X}$ the token vocabulary. Our method assumes an auto-regressive language model parameterized by θ, denoted π_θ. Given a prompt x_0 of length l_0 (we simplify and slightly abuse notation by representing the entire prompt as a single token x_0) and a partial completion $x_1...x_{t-1}$, the model predicts a probability distribution $\pi_\theta(x_t|x_{0:t-1})$ over $\mathcal{X}$, which can then be used at inference to sample a continuation x_t. With this notation, the likelihood of any sequence $x = x_1...x_l$ given prompt x_0 can be expressed as

$$\pi_\theta(x) = \prod_{t=1}^{l} \pi_\theta(x_t|x_{0:t}), \quad (1)$$

and the negative log-likelihood loss incurred by π_θ on x as

$$\ell(x,\theta) = -\frac{1}{l}\sum_{t=1}^{l} \log \pi_\theta(x_t|x_{0:t-1}). \quad (2)$$

Finally, we denote by π_{θ^*} the pretrained model from which we aim to perform TDA, where

$$\theta^* \approx \arg\min_\theta \frac{1}{N}\sum_{i=1}^{N} \ell(x^i,\theta). \quad (3)$$

3.1 Backward Influence Estimation

To assess which training samples most strongly influenced a particular model completion $\hat{x}_{1:l}$ on a user-defined prompt $\hat{x}_0$, we take inspiration from Ko et al. (2024) and reverse the traditional perspective of influence functions. Rather than analyzing the effect of training samples on model outputs (forward influence), we study how optimizing the model on the test sample $\hat{x}$ affects its behavior on the training data $\mathcal{D}$ (backward influence). We construct two perturbed variants by optimizing on the test sample $\hat{x} = \hat{x}_0\hat{x}_1...\hat{x}_l$ using two symmetric procedures: gradient ascent and gradient descent, yielding two new sets of model weights, $\theta_+^{\hat{x}}$ and $\theta_-^{\hat{x}}$, respectively. Specifically, similar to Wang et al. (2024), we initialize $\theta_{\pm}^{(0)} = \theta^*$ and choose $\theta_{\pm}^{\hat{x}} = \theta_{\pm}^{(M)}$ for some M, where

$$\theta_{\pm}^{(m+1)} \leftarrow \theta_{\pm}^{(m)} \pm \frac{\alpha}{N} F_{\theta^*}^{-1} \nabla \ell(x, \theta), \tag{4}$$

and with α a learning rate and F_{θ^*} the Fisher Information Matrix (FIM), given by

$$F_{\theta^*} := \mathbb{E}_{x \sim \mathcal{D}} \left[\nabla_\theta \log \pi_\theta(x)|_{\theta^*} \, \nabla_\theta \log \pi_\theta(x)|_{\theta^*}^\top \right]. \tag{5}$$

Projecting the gradient using the inverse FIM is known as *Elastic Weight Consolidation* (EWC), a method designed to mitigate catastrophic forgetting (Kirkpatrick et al. 2017). Since inverting the exact FIM is computationally intractable, we follow Choi et al. (2025) and use the diagonal approximation

$$\left(\hat{F}_{\theta^*}\right)_{jj} = \frac{1}{NL} \sum_{x^i \in \mathcal{D}} \sum_{t=1}^{L} \left(\left. \frac{\partial \log \pi_\theta(x_t^i | x_{0:t-1}^i)}{\partial \theta} \right|_{\theta = \theta^*} \right)^2 \tag{6}$$

With the perturbed models in hand, we compute the loss of each training sample $x^i \in \mathcal{D}$ under both $\theta_-^{\hat{x}}$ and $\theta_+^{\hat{x}}$. Finally, we define the bidirectional influence score (BIS) of x^i towards the test sample $\hat{x}$ as the absolute change in loss:

$$\mathcal{I}(x^i; \hat{x}) = |\ell(x^i, \theta_-^{\hat{x}}) - \ell(x^i, \theta_+^{\hat{x}})|. \tag{7}$$

Training examples that exhibit large changes in loss under these two models can be interpreted as those potentially most affected by the test sample and, by the *mirrored influence hypothesis* (Ko et al. 2024), most influential in the generation of $\hat{x}$. Our method, *Data Attribution via Bidirectional Gradient Optimization* (DABGO) ranks training samples in descending order of their BIS, enabling targeted inspection and interpretation of the data underlying specific model outputs. A diagram illustrating our method is provided in Fig. 1.

4 Experiments

We evaluate DABGO on two language models based on the GPT-2 architecture (Radford et al. 2019), each trained from scratch on a distinct corpus to support both factual and stylistic attribution. For each generated query $\hat{x}$ we perform $M = 10$ gradient descent and ascent steps following Eq. (4), with learning rate $\alpha = 1 \cdot 10^{-4}$.

4.1 Pretraining

We evaluate DABGO in two complementary settings: factual attribution, using a curated collection of Wikipedia abstracts, and stylistic attribution, using literary texts from the Project Gutenberg archive (Project Gutenberg 2025). The Wikipedia corpus (230M tokens) is derived from the WIT dataset (Srinivasan et al. 2021) and provides a mostly uniform writing style, forcing attribution to rely on factual content rather than stylistic signal. In contrast, the Gutenberg corpus (2M tokens) introduces meaningful stylistic variation across authors and time periods, enabling attribution of stylistic influence. In both cases, we train a GPT-2–style model from scratch with a context window of 256 tokens and create overlapping training blocks using a sliding window of stride 128. Preprocessing includes tokenization, removal of boilerplate, and text normalization.

4.2 Evaluation

We evaluate our attribution method by quantifying the extent to which the top-k attributed training samples influence the likelihood assigned by the model to a given generated test sample.

Tail-Patch Absolute Score. Unlike fact-tracing settings, open-ended generation does not provide a tractable ground truth for attribution: generated text may combine factual, stylistic, or compositional features that do not correspond to any single training example. This makes retrieval- or entailment-based metrics unsuitable, as they assume access to a gold reference passage. Moreover, surface-level overlap is not the objective of training data attribution, which seeks causal influence rather than lexical similarity. Instead, we adopt the *tail-patch absolute* (TPA) evaluation protocol, which quantifies the additive influence of a set of training samples on a given test sample (Chang et al. 2025). TPA offers a computationally efficient proxy for exact influence estimation, which involves retraining from scratch without the attributed samples and is therefore intractable in most practical settings. The evaluation proceeds as follows: for each generated query sentence $\hat{x}$, we identify the top-k most influential training samples using the attribution method under consideration. These samples are then used to perform a single gradient update step on the base model π_{θ^*}, resulting in a perturbed model with parameters θ_k. We compute the likelihood of $\hat{x}$ under both π_{θ^*} and π_{θ_k}, and define the TPA metric τ_k as the absolute change in likelihood, that is

$$\tau_k = \frac{|\pi_{\theta^*}(\hat{x}) - \pi_{\theta_k}(\hat{x})|}{\pi_{\theta^*}(\hat{x})}. \tag{8}$$

This formulation treats attribution as a sensitivity analysis. It is based on the hypothesis that influential training samples should perturb the model's likelihood of $\hat{x}$ more strongly than non-influential samples, either positively or negatively. Using the absolute difference captures the overall coupling between training and test samples. This is particularly important in open-ended generation settings, where influence is not necessarily unidirectional. We observe that computing τ_k on k random training samples yields the lowest values across all tested attribution methods (see Table 1), suggesting that the metric is robust to noise and reflects meaningful influence rather than random variation.

Retraining from Scratch. To validate our main evaluation metric, the tail-patch score, we perform a counterfactual evaluation via full retraining. For each attribution method (BM25, TRACKSTAR, GECKO and DABGO), we identify the top-k influential training samples $\mathcal{D}_k^{\hat{x}} \subset \mathcal{D}$ for a given query $\hat{x}$. We then remove these samples to construct a reduced training set $\mathcal{D}_{-k}^{\hat{x}} = \mathcal{D} \setminus \mathcal{D}_k^{\hat{x}}$, and retrain a new model

	Wikipedia							Gutenberg						
Method	1	3	5	7	10	15	20	1	3	5	7	10	15	20
Random	1.6	4.3	4.7	6.3	8.8	8.7	8.9	42.9	46.5	45.6	45.3	50.0	53.5	42.8
Bm25	**159.1**	90.8	67.3	59.4	57.0	68.7	68.3	75.7	90.3	95.0	110.2	103.0	103.5	104.4
TrackStar	3.9	6.3	7.4	8.5	8.9	9.7	10.3	61.7	57.2	50.8	51.4	56.2	54.8	60.6
Gecko	43.5	55.6	55.2	42.1	29.9	18.2	14.5	**127.5**	**143.9**	177.7	**291.4**	215.8	312.4	258.5
Dabgo	133.8	**101.2**	**75.9**	**73.9**	**77.7**	**89.3**	**76.7**	98.3	112.2	**258.6**	235.6	**355.1**	**377.0**	**470.7**

Table 1: **Quantitative analysis of Dabgo against baselines.** We report the tail-patch absolute scores for varying numbers of top-k proponents, as defined in Eq. (8). Bold indicates the best-performing method, and underlining marks the second best. Results are shown for both factual (Wikipedia) and stylistic (Gutenberg) attribution. In the factual setting, Dabgo consistently outperforms all baselines across values of k, except at $k = 1$. Here Bm25 also identifies a relevant training sample. This is likely due to strong lexical overlap between generated and training sequences in the factual setting, which Bm25 is well-suited to capture. In the stylistic setting, our method performs competitively with Gecko at small k and outperforms as k increases, reflecting that stylistic influence is typically distributed across multiple training samples rather than a single passage.

$\pi_{\theta_{-k}}$ on this subset. We keep all training hyperparameters identical to the base model π_θ. To quantify the effect of removal, we compare the loss of the test sample $\hat{x}$ under the retrained model, $\ell(\hat{x}, \theta_{-k})$, to its original loss under the base model, $\ell(\hat{x}, \theta)$. A higher value of $\ell(\hat{x}, \theta_{-k})$ indicates greater influence of removed training samples, validating the effectiveness of the attribution method. We perform this experiment on four test samples from the Gutenberg models with $k = 20, 50, 100$, corresponding to 0.1%, 0.3%, and 0.6% of the full training set. As shown in Fig. 2, models retrained without samples attributed by Dabgo consistently yield higher losses compared to those trained without samples from Bm25, TrackStar or Gecko. While computationally expensive, this procedure corresponds to the ground-truth definition of influence and provides validation that the tail-patch score is a meaningful and efficient evaluation method.

Comparison Baselines. To evaluate the quality of the identified influential samples, we compare our method against several baselines. These include TrackStar (Chang et al. 2025), which estimates influence via per-sample loss gradients, Best Matching 25 (Bm25) (Robertson and Walker 1994), a classical term-based retrieval method that ranks documents (sequences in our case) by estimating their relevance to a query, and Gecko (Lee et al. 2024), a text-embedding based retrieval method. Similar to the authors in the open-ended text generation experiment (Chang et al. 2025), we do not use the query-specific Hessian matrix approximation for TrackStar. We also include random training samples as a noise baseline to contextualize the inherent variability in τ_k. Evaluation is based on the average absolute change in probability across multiple query samples.

4.3 Factual Attribution

Generation. We sample subject entities from the Wikipedia dataset (Srinivasan et al. 2021) and use them to construct generation prompts. For each subject s, we prompt the model with $\hat{x}_0 = s$. This simple prompting strategy minimizes prompt-induced bias, while enabling targeted generation that reduces the number of relevant samples to attribute, thereby facilitating both qualitative and quantitative evaluation of Dabgo. We generate continuations using Top-50 sampling with temperature 1 and a repetition penalty of 1.5. This setup produces standalone sentences or factual statements about the subject. Examples of our prompts and continuations can be found in Appendix A.

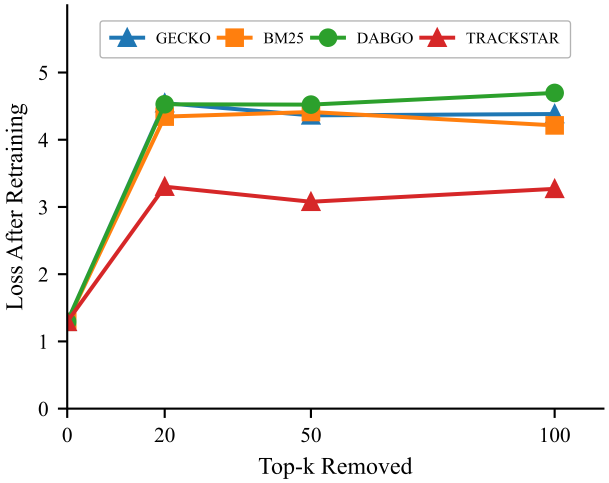

Figure 2: **Loss after retraining without top-k.** We plot the average final loss across four generated sentences $\hat{x}$ from five distinct Gutenberg queries, after removing their top-k proponents from the training set and retraining the model. The resulting increase in loss provides strong evidence that our method accurately identifies influential training samples. Moreover, the alignment between this empirical effect and our tail patch absolute scores supports the validity of our main evaluation metric.

Attribution. Unlike previous work, which often attributes human-written answers to handcrafted prompts (e.g., attributing why the model assigns probability to *"France"* when prompted with *"Paris is in"*), we perform attribution on full model-generated sentences, making it both more realistic and more challenging. We compute τ_k over 25 query samples across varying values of k and present our results in Table 1. We observe that our proposed method DABGO yields consistently stronger attribution performance than evaluated baselines. We further conduct ablations isolating the effect of gradient ascent (unlearning-style updates used in prior work (Choi et al. 2025; Wang et al. 2024; Isonuma and Titov 2024)) and descent (finetuning-style updates). While ascent alone already yields strong performance, combining both ascent and descent, leads to the highest attribution accuracy (cf. Table 2). We attribute this to the generative nature of LLMs: model outputs differ in how "likely" they are by the underlying training distribution, and ascent and descent capture complementary signals. A qualitative examples is provided Appendix B.

Method	**Tail-patch absolute [%] for different k**						
	1	3	5	7	10	15	20
Descent	16.6	22.5	26.4	28.8	32.4	37.4	40.3
Ascent	131.6	83.4	58.2	67.7	59.3	58.9	58.5
DABGO	**133.8**	**101.2**	**75.9**	**73.9**	**77.7**	**89.3**	**76.7**

Table 2: **Ablation study on uni-directional attribution methods** using the Wikipedia dataset. We compare the top-k proponents identified by three different influence estimation methods: Ascent-only: $\ell(\cdot, \theta) - \ell(\cdot, \theta_+)$, Descent-only: $\ell(\cdot, \theta_-) - \ell(\cdot, \theta)$, and DABGO. For each method, we report the tail-patch absolute scores τ_k, averaged over 25 sequences generated from distinct prompts.

4.4 Stylistic Attribution

Generation. In this experiment, the model is prompted with a short sentence sampled from a book, and a continuation is generated (in this case, we do not use a repetition penalty to avoid stylistic bias). More specifically, for each author, we select a validation chunk not seen during training and extract the first complete sentence from that chunk to use as a prompt. We filter for prompts of moderate length to ensure the resulting generation remains within the model context window and has reasonable length. Although no ground-truth attribution targets are available, we make use of metadata associated with the prompt (e.g., author and book title) to assume that influential training samples should exhibit stylistic similarity to the origin of the source. To construct the evaluation queries, we sample one segment of text from each author represented in the subset.

Attribution. Quantitative results for the model trained on the subset are reported in the right half of Table 1. DABGO consistently outperforms BM25 and TRACKSTAR in this setting. The stylistic attribution setting shows the limitations of word-based retrieval methods such as BM25 and the improvement in attribution with GECKO. This is also observable in the qualitative results (cf. Table 4), where we observe that DABGO attributes to a sample originating from the same author as the segment of text used in the prompt, while also more accurately covering the topic and narrative which was generated.

4.5 Interpretability

Beyond ranking training samples, our method supports attribution at arbitrary levels of granularity. In particular, we refine our influence computation to the token level by measuring the absolute difference in per-token log-likelihoods under the updated models. Specifically, for the token at position t in sample x^i, we compute

$$\sigma(x_t^i; \hat{x}) = |\log \pi_{\theta_+^{\hat{x}}}(x_t^i | x_{0:t-1}^i) - \log \pi_{\theta_-^{\hat{x}}}(x_t^i | x_{0:t-1}^i)|. \quad (9)$$

This enables finer-grained interpretability by localizing which parts of a training sample contribute most to its attribution score. In Table 4, we show examples from both factual and stylistic settings, highlighting the most influential tokens within each attributed training sample. This approach is especially beneficial for stylistic attribution, where subtle differences in sentence structure, syntax, and phrasing play a key role. For instance, the stylistic example shown in Table 4 emphasizes first-person narratives, while the factual example highlights complete descriptive segments like "centered on the city of Rome."

4.6 Limitations and Future Work

DABGO remains computationally expensive: it requires two full passes over the training corpus, which is impractical for industrial-scale LLMs trained on hundreds of billions of tokens (Meta AI 2024). Accordingly, our experiments are limited to controlled settings. Although DABGO supports arbitrary granularity, it attributes influence at the individual-sample level and does not account for interactions between samples. Subgroups of examples may exert synergistic or adversarial influence that is not recoverable from per-sample estimates. While Isonuma and Titov (2024) consider attribution at the dataset level, subgroup-level attribution remains an open problem and is even more computationally demanding. Finally, while the highlighted tokens in attributed passages offer qualitative interpretability, we do not establish their causal role. Token-level attribution remains unverified beyond observed loss changes, and isolating true causal components is an important direction for future work.

5 Conclusion

We present DABGO, a novel framework for training data attribution in autoregressive language models, based on backward influence estimation. By comparing training loss under models optimized via gradient ascent and descent, our method identifies influential training samples without expensive per-sample gradient computation. DABGO supports attribution at arbitrary granularity, applies to both factual and stylistic outputs, and enables fine-grained interpretability analysis, providing transparency critical for accountable AI systems. Empirical results demonstrate that DABGO qualitatively and quantitatively outperforms baselines on multiple attribution tasks. Additional experiments with models retrained from scratch further validate its effectiveness, showing strong agreement with ground-truth influence metrics. Unlike prior approaches that rely on handcrafted prompts and static completions, DABGO can attribute in fully open-ended generation settings, offering a practical tool for interpretable and responsible deployment of LLMs.

References

Agarwal, N.; Bullins, B.; and Hazan, E. 2017. Second-Order Stochastic Optimization for Machine Learning in Linear Time. *Journal of Machine Learning Research*, 18(116): 1–40.

Arnoldi, W. E. 1951. The Principle of Minimized Iterations in the Solution of the Matrix Eigenvalue Problem. *Quarterly of Applied Mathematics*, 9(1): 17–29.

Bourtoule, L.; Chandrasekaran, V.; Choquette-Choo, C. A.; Jia, H.; Travers, A.; Zhang, B.; Lie, D.; and Papernot, N. 2021. Machine Unlearning. In *Proceedings of the IEEE Symposium on Security and Privacy*, 141–159.

Chang, T. A.; Rajagopal, D.; Bolukbasi, T.; Dixon, L.; and Tenney, I. 2025. Scalable Influence and Fact Tracing for Large Language Model Pretraining. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Choi, W.; Koo, J.; Cheuk, K. W.; Serrà, J.; Martínez-Ramírez, M. A.; Ikemiya, Y.; Murata, N.; Takida, Y.; Liao, W.-H.; and Mitsufuji, Y. 2025. Large-Scale Training Data Attribution for Music Generative Models via Unlearning. arXiv:2506.18312.

Cook, R. D. 1977. Detection of Influential Observation in Linear Regression. *Technometrics*, 19(1): 15–18.

Ghorbani, A.; and Zou, J. 2019. Data Shapley: Equitable Valuation of Data for Machine Learning. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Grosse, R.; Bae, J.; Anil, C.; Elhage, N.; Tamkin, A.; Tajdini, A.; Steiner, B.; Li, D.; Durmus, E.; Perez, E.; Hubinger, E.; Lukošiūtė, K.; Nguyen, K.; Joseph, N.; McCandlish, S.; Kaplan, J.; and Bowman, S. R. 2023. Studying Large Language Model Generalization with Influence Functions. arXiv:2308.03296.

Guu, K.; Webson, A.; Pavlick, E.; Dixon, L.; Tenney, I.; and Bolukbasi, T. 2023. Simfluence: Modeling the Influence of Individual Training Examples by Simulating Training Runs. arXiv:2303.08114.

Hampel, F. R. 1974. The Influence Curve and Its Role in Robust Estimation. *Journal of the American Statistical Association*, 69(346): 383–393.

Ilyas, A.; Park, S. M.; Engstrom, L.; Leclerc, G.; and Madry, A. 2022. Datamodels: Predicting Predictions from Training Data. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Isonuma, M.; and Titov, I. 2024. Unlearning Traces the Influential Training Data of Language Models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Kirkpatrick, J.; Pascanu, R.; Rabinowitz, N.; Veness, J.; Desjardins, G.; Rusu, A. A.; Milan, K.; Quan, J.; Ramalho, T.; Grabska-Barwinska, A.; et al. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13): 3521–3526.

Ko, M.; Kang, F.; Shi, W.; Jin, M.; Yu, Z.; and Jia, R. 2024. The Mirrored Influence Hypothesis: Efficient Data Influence Estimation by Harnessing Forward Passes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Koh, P. W.; and Liang, P. 2017. Understanding Black-Box Predictions via Influence Functions. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Kwon, Y.; Wu, E.; Wu, K.; and Zou, J. 2024. DataInf: Efficiently Estimating Data Influence in LoRA-Tuned LLMs and Diffusion Models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Lee, J.; Dai, Z.; Ren, X.; Chen, B.; Cer, D.; Cole, J. R.; Hui, K.; Boratko, M.; Kapadia, R.; Ding, W.; et al. 2024. Gecko: Versatile Text Embeddings Distilled from Large Language Models. arXiv:2403.20327.

Luo, Z.; Yang, Z.; Xu, Z.; Yang, W.; and Du, X. 2025. LLM4SR: A Survey on Large Language Models for Scientific Research. arXiv:2501.04306.

Meta AI. 2024. Llama 3.1 Model Card and Prompt Formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/. Accessed: 2025-11-13.

Park, S. M.; Georgiev, K.; Ilyas, A.; Leclerc, G.; and Madry, A. 2023. TRAK: Attributing Model Behavior at Scale. In *Proceedings of the International Conference on Machine Learning (ICML)*, 27074–27113.

Project Gutenberg. 2025. Project Gutenberg Archive. https://www.gutenberg.org. Accessed: 2025-11-13.

Pruthi, G.; Liu, F.; Sundararajan, M.; and Kale, S. 2020. Estimating Training Data Influence by Tracing Gradient Descent. arXiv:2002.08484.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language Models are Unsupervised Multitask Learners.

Robertson, S. E.; and Walker, S. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *Proceedings of the Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Schioppa, A.; Filippova, K.; Titov, I.; and Zablotskaia, P. 2023. Theoretical and Practical Perspectives on What Influence Functions Do. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Schioppa, A.; Zablotskaia, P.; Vilar, D.; and Sokolov, A. 2022. Scaling Up Influence Functions. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Siino, M.; Falco, M.; Croce, D.; and Rosso, P. 2025. Exploring LLMs Applications in Law: A Literature Review on Current Legal NLP Approaches. *IEEE Access*, 13: 18253–18276.

Sim, R. H. L.; Xu, X.; and Low, B. K. H. 2022. Data Valuation in Machine Learning: Ingredients, Strategies, and Open Challenges. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Srinivasan, K.; Raman, K.; Chen, J.; Bendersky, M.; and Najork, M. 2021. WIT: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*.

Tanno, R.; Pradier, M. F.; Nori, A.; and Li, Y. 2022. Repairing Neural Networks by Leaving the Right Past Behind. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Wang, S.-Y.; Hertzmann, A.; Efros, A.; Zhu, J.-Y.; and Zhang, R. 2024. Data Attribution for Text-to-Image Models by Unlearning Synthesized Images. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*.

Zhou, S.; Xu, Z.; Zhang, M.; Xu, C.; Guo, Y.; Zhan, Z.; Fang, Y.; Ding, S.; Wang, J.; Xu, K.; Xia, L.; Yeung, J.; Zha, D.; Cai, D.; Melton, G. B.; Lin, M.; and Zhang, R. 2025. Large Language Models for Disease Diagnosis: A Scoping Review. *npj Artificial Intelligence*, 1(1): 9.

A Prompting Strategy

In Table 3 we show some examples of our prompting strategy for our models and the types of outputs we received.

Prompt	Generated Text
Mount Everest.	Mount Everest is the highest point in the Himalayas at 8,948 m above sea level.
World War I.	It was a global military conflict that embroiled most of the world's great powers, assembled in two opposing alliances: the Entente and the Central Powers. The immediate cause of the war was the June 28, 1914 assassination of Archduke Franz Ferdinand, heir to the Austro-Hungarian throne, by Gavrilo Princip, a Bosnian Serb citizen of Austria–Hungary and member of the Black Hand.
Art Deco.	Art Deco is a style of visual arts, architecture and design that first appeared in France just before World War I.

Table 3: **Prompting Strategy.** We show three examples of model outputs for subject-only prompts of the form *"[subject]."*. The model tends to generate standalone factual statements guided by the topic, without relying on a relational or narrative context. While minor factual errors may occur (e.g., Everest's height is off by 100m), this does not undermine the attribution setting.

B Additional Qualitative Examples

Table 4 shows factual attribution example from Wikipedia and a stylistic attribution example from Project Gutenberg.

Method	Top Proponent
	Model: GPT-2 trained on a subset of Wikipedia abstracts **Prompt**: *Ancient Rome.* **Generated**: *It was a civilization of ancient Rome centered on the city of Rome and its surrounding region.*
BM25	**From Wikipedia article: *Culture of ancient Rome*** The culture of ancient Rome existed throughout the almost 1200-year history of the civilization of Ancient Rome. The term refers to the culture of the Roman Republic, later the Roman Empire, which at its peak covered an area from Lowland Scotland and Morocco to the Euphrates. Life in ancient Rome revolved around the city of Rome, its famed seven hills, and its monumental architecture such as the Colosseum, Trajan's Forum, and the Pantheon. The city also had several theaters, gymnasia, and many taverns, baths, and brothels. Throughout the territory under ancient Rome's control, residential architecture ranged from very modest houses to country villas, and in the capital city of Rome, there were imperial residences on the elegant Palatine Hill, from which the word palace is derived. The vast majority of the population lived in the city center, packed into insulae. The city of Rome was the largest megalopolis of that time, with a population that may well have exceeded one million people, with a high-end estimate of 3.6 million and a low-end estimate of 450,000.
TrackStar	**From Wikipedia article: *List of rulers in the British Isles*** This is a list of rulers in the British Isles. In 1603, King James VI of Scotland also became James I of England, joining the crowns of England and Scotland in personal union. By royal proclamation, James styled himself "King of Great Britain", but no such kingdom was actually created until 1707, when England and Scotland united to form the new Kingdom of Great Britain, with a single British parliament sitting at Westminster, during the reign of Queen Anne.
Gecko	**From Wikipedia article: *Campaign history of the Roman military*** From its origin as a city-state on the peninsula of Italy in the 8th century BC, to its rise as an empire covering much of Southern Europe, Western Europe, Near East and North Africa to its fall in the 5th century AD, the political history of Ancient Rome was closely entwined with its military history. The core of the campaign history of the Roman military is an aggregate of different accounts of the Roman military's land battles, from its initial defense against and subsequent conquest of the city's hilltop neighbors on the Italian peninsula, to the ultimate struggle of the Western Roman Empire for its existence against invading Huns, Vandals and Germanic tribes. These accounts were written by various authors throughout and after the history of the Empire. Following the First Punic War, naval battles were less significant than land battles to the military history of Rome due to its encompassment of lands of the periphery and its unchallenged dominance of the Mediterranean Sea. The Roman army battled first against its tribal neighbours and Etruscan towns within Italy, and later came to dominate the Mediterranean and at its height the provinces of Britannia and Asia Minor.
Dabgo **(ours)**	**From Wikipedia article: *Outline of Ancient Rome*** The following outline is provided as an overview of and topical guide to ancient Rome: Ancient Rome – former civilization that thrived on the Italian Peninsula as early as the 8th century BC. Located along the Mediterranean Sea and centered on the city of Rome, it expanded to become one of the largest empires in the ancient world.
	Model: GPT-2 trained on Project Gutenberg archive **Prompt (from Edward P. Roe's *His Sombre Rivals*)**: *On reaching Centerville, we found everything in the direst confusion.* **Generated**: *As we succeeded in taking it a letter from us to which we saw the enemy. I lowered the flank of our way, for the time that the enemy had given me a roar of fare, though not only was there. Your men had left me, but I could not help feeling that, at length, I saw a soldier among them, and that he would be sure to be shut out the march. As it was, I was quite sure, that he could not bring himself to any possible excuse for a surgeon. I could not have missed my life since*
BM25	**From Alice I. Jones's *Unveiling a Parallel: A Romance*** the performance was quite as distasteful to him as to me, the only difference being that he was familiar with the idea and I was not. As I passed out, I observed that many of the women were vaporizing and many of the men smoking. I suppose it was, in part, the intolerable abomination of these
TrackStar	**From J. Berg Esenwein's *The Art of Public Speaking*** both; it may be physical or spiritual, or both the mental image is at once the most lawless and the most law-abiding child that has ever been born of the mind. First of all, as its name suggests, the process of imagination—for we are thinking of it now as a process rather than as a faculty—is memory at work. Therefore we must consider it primarily as 1. Reproductive Imagination. We see or hear or feel or taste or smell something and the sensation passes away. Yet we are conscious of a greater or lesser ability to reproduce such feelings at will. Two considerations, in general, will govern the vividness
Gecko	**From Edward P. Roe's *His Sombre Rivals*** could judge, was the position of affairs between twelve and one, although I can give you only my impressions. It appeared to me that our men were fighting well, gradually and steadily advancing, and closing in upon the enemy. Still, I cannot help feeling that if we had followed up our success by the determined charge of one brigade that would hold together, the hill might have been swept, and victory made certain. "I had taken my position near Rickett's and Griffin's batteries on the right of our line, and decided to follow them up, not only because they were doing splendid work, but also for the reason that they would
Dabgo **(ours)**	**From Edward P. Roe's *His Sombre Rivals*** Beauregard, but also Johnson from the Shenandoah. "My hope was exceedingly intensified by the appearance of a long line of troops emerging from the woods on our flank and rear, for I never dreamed that they could be other than our own re-enforcements. Suddenly I caught sight of a flag which I had learned to know too well. The line halted a moment, muskets were levelled, and I found myself in a perfect storm of bullets. I assure you I made a rapid change of base, for when our line turned I should be between two fires. As it was, I was cut twice

Table 4: **Qualitative comparison for a factual (top) and a stylistic (bottom) attribution example.** We show the top proponent retrieved by Dabgo and other baseline methods. Highlighted tokens correspond to the largest loss differences in Dabgo, as defined in Eq. (9). In the factual case, the model is prompted with *"Ancient Rome."* and generates a sentence containing *"centered on the city of Rome,"* which Dabgo successfully attributes to a training sample containing this exact phrase (also highlighted as the most influential segment), from the Wikipedia article *Outline of Ancient Rome*. While BM25 also retrieves a relevant training sample, TrackStar does not manage to surface semantically meaningful content. In the stylistic example, Dabgo and Gecko are the only methods that retrieves a thematically consistent, first-person battlefield narrative. In Dabgo (several *"I"* are highlighted along with words such as *"flag"*, *"fires"*, *"bullets"*). In addition, Dabgo returns as the main proponent a passage from the same author and book as the prompt segment used to generate the completion.