Sketching Language Models for Context Attribution

Christopher Liaw
Google Research

Aranyak Mehta
Google Research

Grigoris Velegkas
Google Research

Felix Zhou
Yale University*

Abstract

Context attribution asks how removing parts of a prompt changes a language model's output. We study leave-one-out (LOO) attribution for a prompt of n tokens and a fixed measurement function. Computing all LOO values directly requires n model evaluations. Under a stability assumption, we give an algorithm that estimates all LOO values with error proportional to ε and failure probability δ using $\widetilde{O}(\log(n/\delta)/\varepsilon)$ evaluation calls. To complement our upper bound, we prove an $\Omega(\min\{n, 1/\varepsilon\})$ lower bound on the number of model calls in a black-box oracle model, showing that our inverse-error dependence is optimal up to logarithmic factors. Empirically, we verify coefficient decay consistent with stability and evaluate the performance of our sketching algorithm on a small language-model retrieval task.

Technically, our result is based on a sketching algorithm for arithmetic circuits whose Taylor coefficients satisfy the stability condition recently introduced by Gunn [Gun26]. A direct adaptation of their forward differentiation sketches would require $\widetilde{O}(\log(n/\delta)/\varepsilon^2)$ evaluation calls. Our key observation is that we can apply forward differentiation to the backpropagation circuit, with only constant overhead over a forward sketch of the same Taylor degree. This reduces the variance of the sketches and is key to our improved evaluation bound. Our lower bound follows by a reduction to a known matrix–vector query lower bound for estimating the diagonal elements of a random symmetric matrix [ACDH+24].

*Part of this work was completed during an internship at Google Research.

Contents

1 Introduction **1**
- 1.1 Our Contributions . . . 1
- 1.2 Related Works . . . 3

2 Technical Overview **3**
- 2.1 Sketching Arithmetic Circuits. . . . 4
- 2.2 Our Approach . . . 5
- 2.3 Query Complexity Lower Bound . . . 7

3 Preliminaries **8**
- 3.1 Notation . . . 8
- 3.2 Arithmetic Circuits and Computation . . . 9

4 The Toolkit **10**
- 4.1 Gaussian Sketches Of Symmetric Tensors . . . 10
- 4.2 Backward And Forward Differentiation . . . 11
- 4.3 Stability . . . 12
- 4.4 Trimmed Mean Estimation . . . 13

5 Context Attribution **13**
- 5.1 The Context Attribution Problem . . . 14
- 5.2 Our Attribution Algorithm . . . 14
- 5.3 Query Complexity Lower Bound . . . 19

6 Experiments **20**

A Forward Sketch Baseline **28**

B Continuous Masking **28**

C Diagonal Estimation Lower Bound **28**

D AI Disclosure **30**

1 Introduction

How can we reliably determine which pieces of context actually contributed to a model's output? Frontier AI models frequently process documents and external tool outputs in context to improve output quality (*e.g.*, search results). Answering this question helps us assign credit to sources, identify context that can be pruned, and even detect certain poisoning attacks [CSGM24].

Formally, the problem is to counterfactually estimate how removing parts of the context would have affected the model behavior [LKR25]. To be more precise, assume that we are given access to a deterministic measurement function ϕ whose evaluation involves running an expensive model and outputting some scalar metric. For instance, if $(\tau_1, \ldots, \tau_n)$ is a sequence of n tokens forming a prompt to a language model, and $z = (z_1, \ldots, z_m)$ are the generated response tokens, then $\phi(\tau_1, \ldots, \tau_n) = \phi_z(\tau_1, \ldots, \tau_n)$ might be the log-probability that the model would have generated z using the prompt $(\tau_1, \ldots, \tau_n)$.

In this work, we focus on *leave-one-out (LOO)* context attribution, a well-studied and important setting in practice [LMJ16; CSGM24; LKR25] . That is, we quantify the importance of the i-th token by the value of ϕ without the i-th token[1]

$$\phi(\tau_1, \ldots, \tau_{i-1}, \tau_{i+1}, \ldots, \tau_n)\,.$$

Note that there is a trivial algorithm that exactly computes all LOO values using n different evaluations of ϕ, but is likely to be intractable for large n. Our goal is thus to estimate all LOO values with fewer evaluation calls.

> Our main result is an algorithm that provably estimates all n LOO values for context attribution up to error ε and failure probability δ using $\widetilde{O}(\log(n/\delta)/\varepsilon)$ evaluation calls.

Existing approaches for context attribution use surrogate models [CSGM24], gradient-based attribution scores [QSFB24], and other heuristics [LKR25]. While fast in practice, such heuristics offer no formal accuracy guarantees for modern AI models and can fail when context pieces interact nonlinearly or when local linearization underestimates deletion effects. Our context attribution algorithm is the first with provable prediction guarantees. The proof is based on a property of the model called *stability*. Similar notions have recently been studied in the setting of training data attribution [Gun26]. Experimentally, we verify that stability is compatible with modern language models and that our algorithm accurately estimates LOO values on `Llama-3.2-1B`.

1.1 Our Contributions

Before informally stating our results, we introduce the necessary definitions and context.

Let $\boldsymbol{e}_i \in \{0,1\}^n$ denote the canonical basis vector whose i-th entry is 1 and all other entries are 0. We use binary vectors in $\{0,1\}^n$ to indicate which parts of the context are deactivated (1) versus kept active (0): thus $\mathbf{0}$ represents the full prompt, while $\boldsymbol{e}_i$ deactivates the i-th token. Relaxing

[1] Our techniques extend to the setting where there is an arbitrary partition of the tokens and we wish to estimate the value of ϕ after deleting a set of tokens in the partition.

$\{0,1\}^n$ to $[0,1]^n$ naturally encodes "partial" deactivation: each downweight $w_i \in [0,1]$ controls the extent to which the i-th token is suppressed.

We formally define a measurement function $\phi : [0,1]^n \to \mathbb{R}$ as mapping from downweights $\boldsymbol{w} \in [0,1]^n$ to scalar values. Our goal is to estimate the set of all LOO values $\phi(\boldsymbol{e}_1), \ldots, \phi(\boldsymbol{e}_n)$.

Stability. Let $\otimes$ denote the tensor product. Consider the Taylor expansion of ϕ around the origin $\mathbf{0}$,

$$\phi(\boldsymbol{w}) = \sum_{r \geq 0} T_\phi^{(r)}[\boldsymbol{w}^{\otimes r}], \qquad T_\phi^{(r)} := \frac{1}{r!} D^r \phi(\mathbf{0}).$$

Here $T_\phi^{(r)}$ is the degree-r Taylor coefficient tensor. The main structural property upon which our guarantees are based is *stability* (Definition 4.5). We say that ϕ is α-stable if, for every $B > 0$ and every integer $r \geq 1$, its Taylor coefficients satisfy

$$\left\| T_\phi^{(r)} \right\|_F \leq \frac{\alpha(B)}{B^r}. \tag{1}$$

In our informal discussion below, we say simply that ϕ is *stable* when it is α-stable with $\alpha(B) = O(1)$ independent of the context length n and degree r.

Our Contributions. We are now ready to informally state our results.

Theorem 1.1 (Informal; see Theorems 5.1 and 5.3)**.** *For any $\varepsilon, \delta \in (0,1)$, there is a randomized algorithm for estimating LOO values with the following guarantees. For any evaluation function ϕ mapping n downweights to a scalar metric,*

- *the algorithm makes $\widetilde{O}(\min\{n, \log(n/\delta)/\varepsilon\})$ evaluation calls;*
- *if ϕ is stable, then with probability at least $1 - \delta$, all estimates have error at most $O(\varepsilon)$;*
- *any algorithm achieving error $O(\varepsilon)$ requires $\Omega(\min\{n, 1/\varepsilon\})$ evaluation calls.*

The notion of stability was recently introduced by Gunn [Gun26] for the training data attribution problem, where the goal is to predict how removing training examples changes the output of a learning algorithm. Roughly speaking, they are able to estimate k arbitrary subset deletion queries up to error ε and failure probability δ in the data attribution setting using $\widetilde{O}(\log(k/\delta)/\varepsilon^2)$ training runs. Their approach is based on a novel sketching algorithm for stable functions via forward differentiation. Our techniques are based on a refinement of their sketching approach that reduces the variance of the sketch.

Technical Contributions. There is a natural adaptation of Gunn's forward differentiaton scheme to our setting using $\widetilde{O}(\log(n/\delta)/\varepsilon^2)$ evaluation calls. our key observation is to apply forward differentiation to the *gradient circuit* $\nabla\phi : [0,1]^n \to \mathbb{R}^n$ obtained via backpropagation. By classical backpropagation complexity [Lin76; BS83; RHW86], differentiating $\nabla\phi$ incurs only a constant-factor overhead over differentiating ϕ. As it turns out, this significantly reduces the variance of our sketches compared to just using forward differentiation. Moreover, this only incurs constant overhead when computing $\nabla\phi$ using backpropagation and allows us to improve the number of evaluation calls by a quadratic factor in $1/\varepsilon$.

1.2 Related Works

Context Attribution. Context attribution studies how input passages affect a model's response. Existing methods use sparse surrogates fitted to context ablations [CSGM24], gradient saliency [QSFB24], and efficient LOO estimation through KV caching, hierarchical attribution, and proxy models [LKR25]. Related work evaluates generated citations [GYYC23] and improves them using context-ablation rewards [CCSW+25]. We study the number of calls needed to provably approximate all LOO values simultaneously under stability.

Training Data Attribution. Training data attribution studies how training examples affect model predictions. Approaches include influence functions and training-trajectory scores [KL17; PLKS20], learned surrogates and projected gradients [IPEL+22; PGIL+23], and differentiation through training [IE25; CGTM+26]. Accuracy guarantees and corrections for influence-function and Newton-step approximations have also been studied for convex learning problems [RH25; RH26]. Our setting fixes the trained model and attributes a specified measurement to its input context.

Feature Attribution. Feature attribution assigns importance scores to input features to explain a model's prediction. Methods include representation erasure [LMJ16], local surrogate explanations [RSG16], and Shapley-value methods [LL17]. Our LOO target follows the erasure perspective by measuring the effect of a single deletion from the full input. Shapley values instead average marginal contributions across subsets.

2 Technical Overview

We now give a high-level overview of our context attribution scheme.

Notation. We write $[n] \coloneqq \{1, \dots, n\}$. Let

$$T_f^{(r)} \coloneqq \frac{1}{r!} D^r f(\mathbf{0})$$

denote the degree-r Taylor coefficient tensor of f at the origin, and omit the subscript f when the function is clear. We write $\boldsymbol{x}^{\otimes r}$ for the r-fold tensor product and $T^{(r)}[\boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_r]$ for contraction against these vectors, without complex conjugation. The notation $T^{(r)}[\boldsymbol{x}, \cdot]$ contracts only the first index. We use $\|\cdot\|_F$ for the Frobenius norm, the Euclidean norm of the tensor entries. A dagger denotes conjugate transpose, and $\langle \boldsymbol{x}, \boldsymbol{y} \rangle \coloneqq \boldsymbol{x}^\dagger \boldsymbol{y}$.

Setting. We model the evaluation function ϕ as an arithmetic circuit over complex numbers (Definition 3.1). This is a special case of the computation graphs used in deep learning [Bau74; GBC16]. Recall that evaluating ϕ takes $O(\mathrm{size}(\phi))$ arithmetic operations. We note that all of out results translate to modern language models used in practice and this abstraction is mainly to avoid cumbersome technical details. Indeed, our method only requires that the gates in the computation graph have local polynomial approximations around real inputs and we easily implemented our algorithm on `Llama-3.2-1B`.

We also note that while the notion of downweights is natural for training algorithms such as stochastic gradient descent, *e.g.*, we downweight the gradient of the i-th training example by w_i, there is no consensus on the specific implementation of downweights in the context attribution setting. This is important as not all implementations will satisfy the stability property that we require. We implement the downweights through continuous masking of attention scores and observe that this seems compatible with transformer-based models. See Appendix B for details.

2.1 Sketching Arithmetic Circuits.

Before overviewing our algorithm, it is instructive to abstractly recall Gunn [Gun26]'s sketch for stable arithmetic circuits f.

Suppose f is stable in the sense that its Taylor coefficients decay as $\left\|T_f^{(r)}\right\|_F \leq \frac{\alpha(B)}{B^r}$. We use this property in two ways:

(1) Bound the error from truncating the Taylor expansion;

(2) Bound the variance of sketching the truncated Taylor expansion.

Indeed, assuming stability holds for $B > 2$, truncating the Taylor expansion at order $N = O(\log(1/\varepsilon))$ introduces at most error $O(\varepsilon)$. However, directly computing the exact Taylor coefficients is expensive even for $r \geq 3$ since the degree-r tensor has n^r entries. A natural idea is to compute projections of the Taylor coefficient tensors using random vectors $g_1, \ldots, g_r \in \mathbb{R}^n$ for $1 \leq r \leq N$

$$T^{(r)}[g_1 \otimes \cdots \otimes g_r]\,.$$

Sketching by Random Complex Projections. As it turns out, it suffices to use the tensor power of a single standard *complex* Gaussian vector $g \sim \mathcal{CN}(0, I_n)$. Indeed, given sketches of the form $T^{(r)}[g^{\otimes r}]$, we can obtain unbiased estimates of $T^{(r)}[w^{\otimes r}]$ as follows.

$$\mathop{\mathbb{E}}_{g}\left[\frac{T^{(r)}[g^{\otimes r}]\cdot\langle g^{\otimes r}, w^{\otimes r}\rangle}{r!}\right] = T^{(r)}\mathop{\mathbb{E}}_{g}\left[\frac{g^{\otimes r}(g^{\otimes r})^\dagger}{r!}\right]w^{\otimes r} = T^{(r)}[w^{\otimes r}]\,.$$

One can bound the variance of this sketch as

$$\mathop{\mathrm{Var}}_{g}\left[\mathrm{Re}\,\frac{T^{(r)}[g^{\otimes r}](g^\dagger w)^r}{r!}\right] \leq \binom{2r}{r}\left\|T^{(r)}\right\|_F^2\|w\|_2^{2r}\,.$$

Let us now focus on the LOO estimates $f(e_1), \ldots, f(e_n)$. We will estimate the degree-r term of the Taylor expansion of $f(e_i)$ for all $r \in [N]$ and then combine the per-degree estimates to an estimate of $f(e_i)$. Fix a degree $r \in [N]$. For independent $g_1, \ldots, g_m \sim \mathcal{CN}(0, I_n)$, define

$$Z_{i,r}^{(j)} := \mathrm{Re}\,\frac{T^{(r)}[g_j^{\otimes r}](g_j^\dagger e_i)^r}{r!}\,, \qquad Z_{i,r} := \frac{1}{m}\sum_{j=1}^m Z_{i,r}^{(j)}\,.$$

Then

$$\mathop{\mathbb{E}}_{g_1,\ldots,g_m}[Z_{i,r}] = \mathop{\mathbb{E}}_{g_1}[Z_{i,r}^{(1)}] = T^{(r)}[e_i^{\otimes r}]\,, \qquad \mathop{\mathrm{Var}}_{g_1,\ldots,g_m}(Z_{i,r}) = \frac{1}{m}\mathop{\mathrm{Var}}_{g_1}(Z_{i,r}^{(1)}) \leq \frac{4^r}{m}\left\|T^{(r)}\right\|_F^2\,.$$

By Chebyshev's inequality, with contant probability, the error at degree r is at most the square root of the variance. Unfortunately, we cannot directly employ $Z_{i,r}$ as we do not have sufficiently strong concentration of the sample average, so we must employ a mean estimation subroutine for heavy-tailed distributions. For instance, using a trimmed mean estimator $\mathrm{Trim}(Z_{i,r}^{(1)}, \dots, Z_{i,r}^{(m)})$ which removes a few of the largest and smallest samples [OOR25], we can show that the estimation error is at most

$$O\left(2^r \left\| T^{(r)} \right\|_F \sqrt{\frac{\log(nN/\delta)}{m}}\right)$$

simultaneously for all LOO queries and degrees $r \in [N]$, with probability at least $1 - \delta$. Assuming stability for $B > 2$, we thus have

$$\left| \sum_{r=1}^{N} T^{(r)}[\boldsymbol{e}_i^{\otimes r}] - \sum_{r=1}^{N} \mathrm{Trim}(Z_{i,r}^{(1)}, \dots, Z_{i,r}^{(m)}) \right| = O\left(\sqrt{\frac{\log(nN/\delta)}{m}}\right).$$

Thus $m = O(\log(nN/\delta)/\varepsilon^2)$ sketches suffice to estimate all LOO values up to error ε.

Sketching by Forward Differentiation. Fix $\boldsymbol{g} \in \mathbb{C}^n$. A key observation is that we can compute these directional derivatives $T^{(r)}[\boldsymbol{g}^{\otimes r}]$ for all $r \in [N]$ simultaneously using forward differentiation in the direction $\boldsymbol{g}$ using $O(\mathrm{size}(f) N \log N)$ operations [Wen64]. This can be interpreted as evaluating $f(z\boldsymbol{g})$ over the polynomial ring $\mathbb{C}[z]/(z^{N+1})$. All in all, the sketch can be efficiently computed in $O(\log(nN/\delta)/\varepsilon^2)$ evaluations and each evaluation takes time $\widetilde{O}(\mathrm{size}(f) \log(1/\varepsilon))$.

2.2 Our Approach

Our algorithms are driven by a refinement of the forward sketches for stable arithmetic circuits.

Backward–Forward Sketches. The counterpart to forward differentiation is backward differentiation [Lin76; RHW86]. Roughly speaking, backpropagation constructs an auxiliary circuit of size $O(\mathrm{size}(f))$ that computes

$$\boldsymbol{w} \mapsto (f(\boldsymbol{w}), \nabla f(\boldsymbol{w})).$$

The advantage of backpropagation is that it computes the exact first-order gradient $\nabla f(\boldsymbol{w})$. However, there is no higher-order analogue in the sense that there is no known $\mathrm{size}(f) \cdot \mathrm{poly}(r)$-time algorithm for computing all degree-r partial derivatives.

Our key observation is that we can perform forward differentiation on the backpropagation circuit itself! In Taylor coefficients, this corresponds to sketches of the form

$$T^{(r)}[\boldsymbol{e}_i \otimes \boldsymbol{g}^{\otimes r-1}], i \in [n].$$

In other words, instead of randomly sketching the entire tensor $T^{(r)}$, we sketch disjoint slices $T^{(r)}[\boldsymbol{e}_i, \cdot]$.

Using our new sketches, we can compute LOO estimates in an almost identical fashion. The constant and linear terms are available exactly, so fix a degree $2 \le r \le N$. As the $\boldsymbol{e}_i$'s are orthonormal,

our estimates simplify to

$$\begin{aligned}
\bar{Z}_{i,r}^{(j)} &:= \mathrm{Re} \sum_{\ell \in [n]} \frac{T^{(r)}[\boldsymbol{e}_\ell \otimes \boldsymbol{g}_j^{\otimes(r-1)}](\boldsymbol{e}_\ell^\top \boldsymbol{e}_i)(\boldsymbol{g}_j^\dagger \boldsymbol{e}_i)^{r-1}}{(r-1)!} \\
&= \mathrm{Re}\, \frac{T^{(r)}[\boldsymbol{e}_i \otimes \boldsymbol{g}_j^{\otimes(r-1)}](\boldsymbol{g}_j^\dagger \boldsymbol{e}_i)^{r-1}}{(r-1)!}, \\
\bar{Z}_{i,r} &:= \frac{1}{m} \sum_{j=1}^{m} \bar{Z}_{i,r}^{(j)}, \\
\mathop{\mathbb{E}}_{\boldsymbol{g}_1,\ldots,\boldsymbol{g}_m}[\bar{Z}_{i,r}] &= \mathop{\mathbb{E}}_{\boldsymbol{g}_1}[\bar{Z}_{i,r}^{(1)}] = T^{(r)}[\boldsymbol{e}_i^{\otimes r}] .
\end{aligned}$$

Example 2.1. *Let us illustrate our estimator on $T^{(2)}$ for the sake of clarity. Sketching the full tensor and estimating $T^{(2)}[\boldsymbol{w}^{\otimes 2}]$ corresponds to*

$$\frac{1}{2}(\boldsymbol{g}^\top T^{(2)} \boldsymbol{g})(\boldsymbol{g}^\dagger \boldsymbol{w})^2 .$$

On the other hand, sketching the slices and then estimating corresponds to

$$\boldsymbol{w}^\top T^{(2)} \boldsymbol{g}(\boldsymbol{g}^\dagger \boldsymbol{w}) .$$

Intuitively, this reduces the amount of randomness in the estimator.

Let us formalize the intuition in Example 2.1. Define

$$v_{i,r} := \binom{2r-2}{r-1} \left\| T^{(r)}[\boldsymbol{e}_i, \cdot] \right\|_F^2 .$$

By the orthogonality of $\boldsymbol{e}_i$'s and a Gaussian moment bound on $\bar{Z}_{i,r}$, we have that

$$\mathop{\mathrm{Var}}_{\boldsymbol{g}_1,\ldots,\boldsymbol{g}_m}(\bar{Z}_{i,r}) = \frac{1}{m} \mathop{\mathrm{Var}}_{\boldsymbol{g}_1}(\bar{Z}_{i,r}^{(1)}) \le \frac{1}{m} v_{i,r} .$$

Here $v_{i,r}$ is an upper bound for the degree-r variance. Similar to Gunn's approach, we will apply a separate trimmed mean $\mathrm{Trim}(\bar{Z}_{i,r}^{(1)}, \ldots, \bar{Z}_{i,r}^{(m)})$ for every i and r. The degree-r error is

$$O\left(\sqrt{\frac{v_{i,r} \log(nN/\delta)}{m}} \right) .$$

Variance of the Estimator. We then add the degree-wise estimates and their error bounds. Define

$$s_i := \sum_{r=2}^{N} \sqrt{v_{i,r}} .$$

Following our per-degree calculation above, the total error for estimating $f(\boldsymbol{e}_i)$ from samples is

$$\left| \sum_{r=2}^{N} T^{(r)}[\boldsymbol{e}_i^{\otimes r}] - \sum_{r=2}^{N} \mathrm{Trim}(\bar{Z}_{i,r}^{(1)}, \ldots, \bar{Z}_{i,r}^{(m)}) \right| = O\left(s_i \sqrt{\frac{\log(nN/\delta)}{m}} \right) .$$

Ignoring the common $\widetilde{O}(\sqrt{1/m})$ term, it follows that the root-mean-square estimation error across LOO queries scales as

$$\begin{aligned}
\sqrt{\sum_i s_i^2} &= \sqrt{\sum_i \left(\sum_{r=2}^{N} \sqrt{v_{i,r}}\right)^2} \leq \sum_{r=2}^{N} \sqrt{\sum_i v_{i,r}} && \text{Minkowski's inequality} \\
&= \sum_{r=2}^{N} \sqrt{\binom{2r-2}{r-1} \sum_i \left\| T^{(r)}[\boldsymbol{e}_i, \cdot] \right\|_F^2} \\
&\leq \sum_{r=2}^{N} \sqrt{4^{r-1} \left\| T^{(r)} \right\|_F^2} \leq \sum_{r=2}^{N} 2^r \left\| T^{(r)} \right\|_F \\
&= O(1)\,.
\end{aligned}$$

The last equality holds under stability for $B > 2$.

By Markov's inequality, at most $O(1/\varepsilon)$ of the squared error s_i^2's are more than $O(\varepsilon)$. We select the $O(1/\varepsilon)$ largest s_i's and evaluate the corresponding values $f(\boldsymbol{e}_i)$ exactly in $O(\text{size}(f)/\varepsilon)$ time. Every remaining query has $s_i = O(\sqrt{\varepsilon})$, so its total sampling error is

$$O\left(\sqrt{\frac{\varepsilon \log(nN/\delta)}{m}}\right).$$

Thus $m = O(\log(nN/\delta)/\varepsilon)$ sketches suffice for ε-error.

In order to estimate s_i, we can estimate the $v_{i,r}$ using a trimmed mean estimator $\widehat{v}_{i,r}$ from samples. Using a fourth-moment bound, we can achieve constant relative error estimates using $O(4^{N-1}\log(nN/\delta))$ sketches. For $N = \log_4(1/\varepsilon)$, this is $O(\log(nN/\delta)/\varepsilon)$. We can thus estimate s_i up to constant relative error

$$\widehat{s}_i := \sum_{r=2}^{N} \sqrt{\widehat{v}_{i,r}}\,.$$

All in our, our algorithm requires $O(\log(nN/\delta)/\varepsilon)$ model calls with each call taking $\widetilde{O}(\text{size}(f)\log(1/\varepsilon))$ operations.

2.3 Query Complexity Lower Bound

In Section 5.3, we complement our algorithm with a nearly matching lower bound in the number of model calls in an oracle model.

Oracle Model. We assume that we are given access to a stable function $f : [0,1]^n \to \mathbb{R}$ in the form of an oracle. For any input $\boldsymbol{g} \in \mathbb{C}^n$, and integer $r \geq 1$, the oracle provides

$$(\boldsymbol{w}, \boldsymbol{g}, r) \mapsto \left(f(\boldsymbol{w}), \boldsymbol{\nabla} f(\boldsymbol{w}), D_{\boldsymbol{g}}^1 \boldsymbol{\nabla} f(\boldsymbol{w}), \ldots, D_{\boldsymbol{g}}^r \boldsymbol{\nabla} f(\boldsymbol{w})\right).$$

Thus it provides exact zero-th and first-order information at $\boldsymbol{w}$, in addition to directional derivatives of arbitrary order,

$$D_{\boldsymbol{g}}^k \boldsymbol{\nabla} f(\boldsymbol{w}) := \frac{\partial^k}{\partial z^k} \boldsymbol{\nabla} f(\boldsymbol{w} + z\boldsymbol{g})\bigg|_{z=0},$$

exactly as our algorithm requires. As we will see, the terms $r > 1$ will not be useful for the lower bound instance.

Diagonal Estimation Lower Bound. We show that our $1/\varepsilon$ model calls is optimal up to logarithmic factors in an oracle model. It suffices to consider quadratic functions $f(\boldsymbol{w}) = \boldsymbol{w}^\top A\boldsymbol{w}$ for a random symmetric matrix $A \in \mathbb{R}^{n\times n}$. Note that f is stable since its higher-order Taylor coefficients are all zero. Observe that

$$\begin{aligned} f(\boldsymbol{w}) &= \boldsymbol{w}^\top A\boldsymbol{w}\,, \\ \boldsymbol{\nabla} f(\boldsymbol{w}) &= 2A\boldsymbol{w}\,, \\ D^1_{\boldsymbol{g}}\boldsymbol{\nabla} f(\boldsymbol{w}) &= 2A\boldsymbol{g}\,, \\ D^r_{\boldsymbol{g}}\boldsymbol{\nabla} f(\boldsymbol{w}) &= 0\,, \quad \forall r > 1\,. \end{aligned}$$

Hence the oracle essentially returns two matrix-vector products $A\boldsymbol{w}, A\boldsymbol{g}$ in every query.

Since $\boldsymbol{e}_i^\top A\boldsymbol{e}_i = A_{ii}$, estimating the leave-one-out values in this oracle model essentially amounts to estimating the diagonal of A from matrix–vector products. By a known lower bound for this problem due to Amsel, Chen, Duman Keles, Halikias, Musco, and Musco [ACDH+24], we need $\Omega(\min\{n, 1/\varepsilon\})$ queries and thus oracle calls to achieve estimates of error ε.

3 Preliminaries

In this section, we establish the notation use throughout the rest of this work and review the necessary background.

3.1 Notation

For a positive integer n, write $[n] \coloneqq \{1, \ldots, n\}$. Let $\boldsymbol{e}_i$ denote the i-th standard basis vector, and let $\mathbf{1}_D \in \mathbb{R}^n$ denote the indicator vector of $D \subseteq [n]$. For a complex vector $\boldsymbol{x}$, write $\overline{\boldsymbol{x}}$ for its complex conjugate and $\boldsymbol{x}^\dagger \coloneqq \overline{\boldsymbol{x}}^\top$ for its conjugate transpose. All logarithms are natural unless a base is specified.

Tensors And Inner Products. For tensors $S, T \in (\mathbb{C}^n)^{\otimes r}$ with $r \geq 1$, we use the Frobenius inner product

$$\langle S, T\rangle \coloneqq \sum_{j_1,\ldots,j_r=1}^{n} \overline{S_{j_1,\ldots,j_r}} T_{j_1,\ldots,j_r}, \qquad \|T\|_F^2 \coloneqq \langle T, T\rangle,$$

which is conjugate-linear in its first argument. For vectors, we use the Euclidean norm $\|\boldsymbol{x}\|_2$. For a (complex) tensor $T \in (\mathbb{C}^n)^{\otimes r}$ and vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_r \in \mathbb{C}^n$, write

$$T[\boldsymbol{x}_1 \otimes \cdots \otimes \boldsymbol{x}_r] \coloneqq \sum_{j_1,\ldots,j_r=1}^{n} T_{j_1,\ldots,j_r} (\boldsymbol{x}_1)_{j_1} \cdots (\boldsymbol{x}_r)_{j_r} \in \mathbb{C},$$

and abbreviate this as $T[\boldsymbol{x}^{\otimes r}]$ when all arguments equal $\boldsymbol{x}$. Note that contraction does not conjugate the arguments. We write $T[\boldsymbol{x}, \cdot] \in (\mathbb{C}^n)^{\otimes(r-1)}$ for the partial contraction that retains $r-1$ indices,

and $T[\cdot, \boldsymbol{x}^{\otimes(r-1)}] \in \mathbb{C}^n$ for the partial contraction that retains a single index. We identify order-zero tensors with scalars.

Let $\mathrm{Sym}^r(\mathbb{C}^n)$ denote the subspace of $(\mathbb{C}^n)^{\otimes r} \cong \mathbb{C}^{n^r}$ consisting of tensors invariant under permutations of their indices, and let $\Pi_{\mathrm{sym}}^{(n,r)}$ denote its orthogonal projector.

Taylor Coefficients. For a function $f : \mathbb{C}^n \to \mathbb{C}$ that is holomorphic in a neighborhood of the origin, define its normalized Taylor tensors by

$$T_f^{(r)} := \frac{1}{r!} D^r f(\mathbf{0}).$$

When the function is clear, we omit the subscript f. For $r \geq 1$, the tensor $T^{(r)}$ belongs to $\mathrm{Sym}^r(\mathbb{C}^n)$. The degree-zero term $T^{(0)} = f(\mathbf{0})$ is a complex scalar, and $T^{(1)} = \boldsymbol{\nabla} f(\mathbf{0})$ is a complex vector.

Complex Gaussians. Let $\boldsymbol{a}, \boldsymbol{b}$ be independent $\mathcal{N}(0, I_n)$ vectors. Let

$$\boldsymbol{g} := \frac{\boldsymbol{a} + \mathrm{i}\boldsymbol{b}}{\sqrt{2}}$$

denote a standard complex Gaussian vector and write $\boldsymbol{g} \sim \mathcal{CN}(0, I_n)$. Note that $\mathbb{E}_{\boldsymbol{g}}[g_i \overline{g_j}]$ is 1 when $i = j$ and 0 otherwise, and $\mathbb{E}_{\boldsymbol{g}}[g_i g_j] = 0$ for all i, j. For a complex random variable X, write $\mathrm{Var}_X(X) := \mathbb{E}_X |X - \mathbb{E}_X X|^2$.

3.2 Arithmetic Circuits and Computation

Computation. We count complex arithmetic operations as constant time and assume that complex numbers can be stored exactly in constant space. We further assume random access to memory with a word size large enough for each lookup to take constant time.

Arithmetic Circuits. We formally represent a neural network as an *arithmetic circuit* over complex numbers. This is a special case of the computation graphs used in deep learning [Bau74; GBC16].

Definition 3.1 (Arithmetic Circuit)**.** *An arithmetic circuit $f : \mathbb{C}^n \to \mathbb{C}^p$ is a finite directed acyclic graph with labeled vertices. Each vertex with in-degree zero is labeled by an input index $i \in [n]$ or a complex constant. Each vertex with positive in-degree is labeled by $+$ or $\times$ and computes the sum or product of its incoming values. Its value is passed to all of its children. We do not restrict the in-degree or out-degree. Each vertex with out-degree zero is also labeled by an output index in $[p]$. Input indices may be reused, and each output index labels exactly one vertex. For an input $\boldsymbol{w} \in \mathbb{C}^n$, a vertex labeled by input index i receives w_i. Evaluating the gates in topological order gives $f(\boldsymbol{w})$ in output-index order. We define its size, denoted* $\mathrm{size}(f)$*, as the maximum of n, p, and the number of edges.*

A scalar-valued arithmetic circuit with constants computes a polynomial. Its Taylor expansion is therefore finite and exact, including at complex inputs:

$$f(\boldsymbol{w}) = \sum_{r \geq 0} T^{(r)}[\boldsymbol{w}^{\otimes r}], \qquad \boldsymbol{w} \in \mathbb{C}^n,$$

where $T^{(r)} = 0$ above the degree of the polynomial.

We work under Definition 3.1 to avoid cumbersome technical details. Gates such as ReLU, GELU, and Softmax may be replaced by chosen polynomial approximations before applying the results. Indeed, we easily implemented the procedure on the open-weight `Llama-3.2-1B` model.

4 The Toolkit

We now present our algorithm design tool kit to obtain our context attribution algorithm. In Section 4.1, we establish some basic property of complex Gaussian sketches. Then in Section 4.2, we describe our main sketching subroutine which involves both backpropagation and forward differentiation. Next, we formally define stability and explore basic intuition and consequences of this definition in Section 4.3. Finally, we state the mean estimation subroutine we employ for potentially heavy-tailed distributions in Section 4.4.

4.1 Gaussian Sketches Of Symmetric Tensors

Let g be a standard complex Gaussian vector. The following proposition explains why contractions against $g^{\otimes r}$ suffice to sketch symmetric tensors. We note that the same fact does not hold for real Gaussian vectors and it is necessary to use *complex* Gaussians.

Proposition 4.1 (Complex Gaussian Moments; [Har13, Eq. (9)])**.** *Let $g \sim \mathcal{CN}(0, I_n)$. The following hold for every integer $r \geq 1$.*

(i) The Gaussian moment operator satisfies

$$\mathop{\mathbb{E}}_{g}\left[g^{\otimes r}(g^{\otimes r})^{\dagger}\right] = r!\,\Pi_{\mathrm{sym}}^{(n,r)}.$$

(ii) For every tensor $A \in (\mathbb{C}^n)^{\otimes r}$,

$$\mathop{\mathbb{E}}_{g}|A[g^{\otimes r}]|^2 = r!\left\|\Pi_{\mathrm{sym}}^{(n,r)}A\right\|_F^2 \leq r!\|A\|_F^2.$$

The next lemma records the moments needed by our estimator. Its first conclusion will let us estimate the size of a tensor contraction, and its second conclusion gives an unbiased estimate of the contraction in a prescribed direction and bounds its second moment.

Lemma 4.2 (Gaussian Tensor Moments)**.** *Let $S \in \mathrm{Sym}^r(\mathbb{R}^n)$, $x \in \mathbb{R}^n$, and $g \sim \mathcal{CN}(0, I_n)$, where $r \geq 1$. The following hold.*

(i) The second and fourth moments satisfy

$$\mathop{\mathbb{E}}_{g}|S[g^{\otimes r}]|^2 = r!\|S\|_F^2, \qquad \mathop{\mathbb{E}}_{g}|S[g^{\otimes r}]|^4 \leq (2r)!\|S\|_F^4.$$

(ii) The directional sketch satisfies

$$\mathop{\mathbb{E}}_{g}\left[\frac{S[g^{\otimes r}](g^{\dagger}x)^r}{r!}\right] = S[x^{\otimes r}]$$

and

$$\mathop{\mathbb{E}}_{g}\left|\frac{S[g^{\otimes r}](g^{\dagger}x)^{r}}{r!}\right|^{2} \leq \binom{2r}{r}\|S\|_{F}^{2}\|x\|_{2}^{2r}.$$

Proof. For Lemma 4.2(i), symmetry of S and Proposition 4.1(ii) give

$$\mathop{\mathbb{E}}_{g}|S[g^{\otimes r}]|^{2} = r!\|S\|_{F}^{2}.$$

Applying Proposition 4.1(ii) in order $2r$ gives

$$\begin{aligned}\mathop{\mathbb{E}}_{g}|S[g^{\otimes r}]|^{4} &= \mathop{\mathbb{E}}_{g}|(S\otimes S)[g^{\otimes 2r}]|^{2}\\ &= (2r)!\left\|\Pi_{\mathrm{sym}}^{(n,2r)}(S\otimes S)\right\|_{F}^{2}\\ &\leq (2r)!\|S\|_{F}^{4}.\end{aligned}$$

For Lemma 4.2(ii), write

$$Y = \frac{S[g^{\otimes r}](g^{\dagger}x)^{r}}{r!}.$$

Since S is real and $x^{\otimes r}$ is symmetric, Proposition 4.1(i) gives

$$\begin{aligned}\mathop{\mathbb{E}}_{g}Y &= \frac{1}{r!}\left\langle S, \mathop{\mathbb{E}}_{g}\left[g^{\otimes r}(g^{\otimes r})^{\dagger}\right]x^{\otimes r}\right\rangle\\ &= \langle S, x^{\otimes r}\rangle = S[x^{\otimes r}].\end{aligned}$$

Since x is real, Proposition 4.1(ii) in order $2r$ gives

$$\begin{aligned}\mathop{\mathbb{E}}_{g}|Y|^{2} &= \frac{1}{(r!)^{2}}\mathop{\mathbb{E}}_{g}\left|(S\otimes x^{\otimes r})[g^{\otimes 2r}]\right|^{2}\\ &= \binom{2r}{r}\left\|\Pi_{\mathrm{sym}}^{(n,2r)}(S\otimes x^{\otimes r})\right\|_{F}^{2}\\ &\leq \binom{2r}{r}\|S\|_{F}^{2}\|x\|_{2}^{2r}.\end{aligned}$$

□

4.2 Backward And Forward Differentiation

The standard method for computing the gradient of a scalar-valued arithmetic circuit f is backward differentiation (backpropagation) [RHW86]. While it computes the full first-order gradient vector exactly in $O(\mathrm{size}(f))$ time, explicitly computing or storing the full order-r derivative tensor requires $\binom{n+r-1}{r}$ entries in general, even after accounting for symmetry.

Proposition 4.3 (Backward Differentiation; [RHW86], [CKW11, Theorem 9.10])**.** *Let $f : \mathbb{C}^{n} \to \mathbb{C}$ be an arithmetic circuit. Backpropagation constructs an arithmetic circuit of size $O(\mathrm{size}(f))$ computing*

$$w \longmapsto \big(f(w), \nabla f(w)\big).$$

On the other hand, forward differentiation allows us to compute all *directional* derivatives up to order N in time $O(\mathrm{size}(f) N \log N)$. Roughly speaking, forward differentiation at point $\boldsymbol{x}$ in direction $\boldsymbol{g}$ is accomplished by evaluating $f(\boldsymbol{x} + z\boldsymbol{g})$ over the polynomial ring $\mathbb{C}[z]/(z^{N+1})$. Here z is a formal polynomial variable.

Proposition 4.4 (Forward Differentiation; [Wen64], [Gun26, Section 3.1])**.** *Let $F : \mathbb{C}^n \to \mathbb{C}^p$ be an arithmetic circuit, $\boldsymbol{g} \in \mathbb{C}^n$ a direction, and $N \geq 1$ an integer. Evaluating F over $\mathbb{C}[z]/(z^{N+1})$ at input $z\boldsymbol{g}$ computes*

$$[z^r]F(z\boldsymbol{g}) = \frac{1}{r!} D^r F(\boldsymbol{0})[\boldsymbol{g}^{\otimes r}] \ \in \mathbb{C}^p, \qquad 0 \leq r \leq N,$$

in $O(\mathrm{size}(F)\, N \log N)$ arithmetic operations and $O(\mathrm{size}(F)\, N)$ working storage.

Our key observation is that the circuit obtained by backward differentiation is itself an arithmetic circuit, with only constant-factor overhead. We can therefore apply forward differentiation to the backpropagation circuit and obtain sliced sketches essentially for free! We refer to this procedure as a *backward–forward sketch*.

Specifically, for a direction $\boldsymbol{g} \in \mathbb{C}^n$ and an integer $N \geq 1$, a backward–forward sketch computes the coefficients of $\boldsymbol{\nabla} f(z\boldsymbol{g})$ through degree $N-1$, giving

$$\boldsymbol{c}_r(\boldsymbol{g}) := \frac{1}{r}[z^{r-1}]\boldsymbol{\nabla} f(z\boldsymbol{g}) = T^{(r)}[\cdot, \boldsymbol{g}^{\otimes(r-1)}] \ \in \mathbb{C}^n, \qquad 1 \leq r \leq N.$$

By Propositions 4.3 and 4.4, a single sketch computes these vectors in $O(\mathrm{size}(f)\, N \log N)$ arithmetic operations and $O(\mathrm{size}(f)\, N)$ working storage. We retain f as an output of the gradient circuit, so the constant coefficients also give $f(\boldsymbol{0})$ and $\boldsymbol{\nabla} f(\boldsymbol{0})$.

4.3 Stability

The key structural property introduced by Gunn [Gun26] is *stability*.

Definition 4.5 (Stability)**.** *Let f be an arithmetic circuit and let $\alpha : (0, \infty) \to [0, \infty]$. We say that f is α-*stable *if, for every $B > 0$ and every integer $r \geq 1$, the coefficients of the Taylor expansion about the origin satisfy*

$$\left\| T^{(r)} \right\|_F \leq \frac{\alpha(B)}{B^r}.$$

To see why this might be a natural property, recall that a complex-valued function $f : \mathbb{C}^n \to \mathbb{C}$ has a *modulus of continuity at* $\boldsymbol{0}$ if, for every $\boldsymbol{x} \neq \boldsymbol{0}$,

$$|f(\boldsymbol{x}) - f(\boldsymbol{0})| \leq \alpha(\|\boldsymbol{x} - \boldsymbol{0}\|_2).$$

If this is the case, and f is in addition holomorphic, then Cauchy's estimate [Rud87, Theorem 10.26] gives

$$\left\| T^{(r)} \right\|_{\mathrm{op}} \leq \frac{\alpha(B)}{B^r}.$$

Here

$$\left\| T^{(r)} \right\|_{\mathrm{op}} := \sup_{\|\boldsymbol{u}\|_2 = 1} |T^{(r)}[\boldsymbol{u}^{\otimes r}]|,$$

and the bound holds for every $B > 0$ and $r \geq 1$. Stability strengthens this coefficient bound by controlling the Frobenius norm.

An immediate consequence of Definition 4.5 is the following bound on the Taylor truncation error.

Lemma 4.6 (Taylor Truncation)**.** *Suppose f is α-stable. For every nonzero $\boldsymbol{w} \in \mathbb{C}^n$ with $\alpha(4\|\boldsymbol{w}\|_2) < \infty$ and every integer $N \geq 1$,*

$$\left| f(\boldsymbol{w}) - \sum_{r=0}^{N} T^{(r)}[\boldsymbol{w}^{\otimes r}] \right| \leq \frac{4^{-N}}{3} \alpha(4\|\boldsymbol{w}\|_2).$$

Proof. For every $r > N$, stability gives

$$|T^{(r)}[\boldsymbol{w}^{\otimes r}]| \leq \left\| T^{(r)} \right\|_F \|\boldsymbol{w}\|_2^r \leq 4^{-r} \alpha(4\|\boldsymbol{w}\|_2).$$

Sum over $r > N$ and use $\sum_{r>N} 4^{-r} = 4^{-N}/3$. □

4.4 Trimmed Mean Estimation

A key subroutine of our context attribution algorithm is estimating the mean of a random variable given only a second moment bound. In general, this is insufficient for the exponential concentration needed for logarithmic dependence on the inverse failure probability $1/\delta$. We therefore use the trimmed mean, a classical mean estimator for possibly heavy-tailed observations. In 1-D, we remove the k smallest and largest observations and take the average of the remaining observations.

Definition 4.7 (Trimmed Mean; [LM19, Section 2.3])**.** *For real numbers $y_1, \ldots, y_m$, write $y_{(1)} \leq \cdots \leq y_{(m)}$ for their nondecreasing rearrangement. For an integer $0 \leq k < m/2$, define*

$$\mathrm{Trim}_k(y_1, \ldots, y_m) := \frac{1}{m - 2k} \sum_{j=k+1}^{m-k} y_{(j)}.$$

The next proposition bounds the error of the trimmed mean under a finite-variance assumption.

Proposition 4.8 (Trimmed-Mean Concentration; [OOR25, Theorem 5.1.1])**.** *Let $X_1, \ldots, X_m$ be independent, identically distributed real random variables with mean μ and finite variance σ^2. Fix $\delta \in (0, 1)$ and suppose $m \geq 64 \log(16/\delta)$. Then, with probability at least $1 - \delta$ over $X_1, \ldots, X_m$,*

$$\left| \mathrm{Trim}_{\lfloor \log(16/\delta) \rfloor}(X_1, \ldots, X_m) - \mu \right| \leq 7\sigma \sqrt{\frac{\log(16/\delta)}{m}}.$$

5 Context Attribution

In this section, we study the context attribution problem (Section 5.1) and present our algorithm for estimating LOO values (Section 5.2). We complement our upper bounds on the number of evaluation calls with a matching lower bound in the inverse error parameter $1/\varepsilon$ (Section 5.3).

5.1 The Context Attribution Problem

Before presenting our main result, we remind the reader of the setting of the context attribution problem.

The Setting. Let $\phi : \mathbb{C}^n \to \mathbb{C}$ be an arithmetic circuit that is real-valued on $[0,1]^n$. Its input consists of downweights $\boldsymbol{w} = (w_1, \ldots, w_n)$, where $w_i = 0$ includes the i-th token and $w_i = 1$ deletes it. Our goal is to estimate all LOO values $\phi(\boldsymbol{e}_1), \ldots, \phi(\boldsymbol{e}_n)$. For transformer-based language models, Appendix B describes an implementation of these downweights using continuous attention masks. The algorithm computes backward–forward sketches of ϕ and evaluates a few high variance LOO values exactly.

5.2 Our Attribution Algorithm

We now present our main result for estimating LOO values.

Theorem 5.1 (LOO Attribution). *Let $\phi : \mathbb{C}^n \to \mathbb{C}$ be an arithmetic circuit that is real-valued on $[0,1]^n$, and fix $\varepsilon, \delta \in (0,1)$. There is a randomized algorithm for estimating all LOO values with the following guarantees.*

(i) The algorithm takes

$$\widetilde{O}\left(\frac{\mathrm{size}(\phi)\log(n/\delta)}{\varepsilon}\right)$$

arithmetic operations and uses

$$\widetilde{O}\left(\mathrm{size}(\phi) + \frac{n\log(n/\delta)}{\varepsilon}\right)$$

auxiliary space.

(ii) If ϕ is α-stable and $\alpha(4) < \infty$, then, with probability at least $1-\delta$ over the Gaussian directions, simultaneously for every $i \in [n]$,

$$|\widehat{L}_i - \phi(\boldsymbol{e}_i)| \le \varepsilon\alpha(4).$$

(iii) For $N = \log_4(1/\varepsilon)$, the total number of evaluation calls is

$$\widetilde{O}\left(\min\left\{n, \frac{\log(n/\delta)}{\varepsilon}\right\}\right).$$

As previously remarked, our algorithm improves the dependence on $1/\varepsilon$ from quadratic to linear in the number of arithmetic operations compared with direct forward differentiation sketches (Theorem A.1).

The rest of this subsection is dedicated to proving Theorem 5.1.

Degree-Wise Estimates. Our strategy is to sketch the degree-r Taylor coefficients separately for each $r \in [N]$ and then add the resulting estimates and error bounds.

For simplicity, we write $T^{(r)} = T_\phi^{(r)}$. Recall from Section 4.2 that for a direction $g \in \mathbb{C}^n$ and an integer $N \geq 1$, a backward–forward sketch computes the coefficients of $\nabla\phi(zg)$ through degree $N-1$, giving

$$c_r(g) := \frac{1}{r}[z^{r-1}]\nabla\phi(zg) = T^{(r)}[\cdot, g^{\otimes(r-1)}] \in \mathbb{C}^n, \qquad 1 \leq r \leq N.$$

Moreover, the constant and linear terms are available exactly.

For independent $g_1, \ldots, g_m \sim \mathcal{CN}(0, I_n)$, define, for every $i \in [n]$ and $2 \leq r \leq N$, the following estimates of the degree-r Taylor expansion along e_i

$$\bar{Z}_{i,r}^{(j)} := \mathrm{Re}\,\frac{e_i^\top c_r(g_j)(g_j^\dagger e_i)^{r-1}}{(r-1)!}.$$

As in Section 2.2, define a variance proxy for each degree and an error proxy for their sum:

$$v_{i,r} := \binom{2r-2}{r-1}\left\|T^{(r)}[e_i, \cdot]\right\|_F^2, \qquad s_i := \sum_{r=2}^{N}\sqrt{v_{i,r}}.$$

Algorithm Description. We first describe the algorithm for a truncation degree N, m backward–forward sketches, and K exact evaluations. We choose these parameters in the proof of the theorem. If $n \leq m + K$, we can trivially evaluate all LOO values exactly. Thus we proceed assuming otherwise.

Our first step is to compute backward–forward sketches of the degree-r Taylor tensors simultaneously for each $r \in [N]$. For each random complex Gaussian $g \in \mathbb{C}^n$, this gives us sketches of the form $T^{(r)}[e_i \otimes g^{\otimes r-1}]$ for $i \in [n], r \in [N]$. For each LOO value $\phi(e_i)$ and Taylor degree $2 \leq r \leq N$, we estimate the degree-r Taylor term and its variance proxy using trimmed means with $k = \log(32nN/\delta)$:

$$\widehat{Z}_{i,r} \leftarrow \mathrm{Trim}_k(\bar{Z}_{i,r}^{(1)}, \ldots, \bar{Z}_{i,r}^{(m)}),$$
$$\widehat{v}_{i,r} \leftarrow \frac{\binom{2r-2}{r-1}}{(r-1)!}\mathrm{Trim}_k\left(|e_i^\top c_r(g_1)|^2, \ldots, |e_i^\top c_r(g_m)|^2\right).$$

This gives an estimate $\widehat{s}_i \leftarrow \sum_{r=2}^{N}\sqrt{\widehat{v}_{i,r}}$ of the error proxy s_i. We evaluate the K LOO values with the largest estimated error proxies exactly and output the sketched estimates for the rest.

Pseudocode. The pseudocode can be found in Algorithm 1.

We are now ready to prove our main result.

Proof of Theorem 5.1. We first control the degree-wise estimates and their variance proxies, keeping N, m, K general. We then use stability to bound the error after selecting K queries for exact evaluation. Finally, we choose the parameters and bound the complexity.

Degree-Wise Estimates. Suppose ϕ is α-stable and $\alpha(4) < \infty$. Apply Lemma 4.2(ii) to the symmetric order-$(r-1)$ tensor $T^{(r)}[e_i, \cdot]$ and the direction e_i. The degree-r estimates are unbiased and satisfy

$$\mathop{\mathbb{E}}_{g_j}\bar{Z}_{i,r}^{(j)} = T^{(r)}[e_i^{\otimes r}], \qquad \mathop{\mathrm{Var}}_{g_j}(\bar{Z}_{i,r}^{(j)}) \leq v_{i,r}.$$

Algorithm 1: LOO Attribution

Input: The circuit ϕ; parameters N, m, K; a failure probability δ.

Output: Leave-one-out estimates.

Set $k \leftarrow \log(32nN/\delta)$.

Draw independent $\boldsymbol{g}_1, \dots, \boldsymbol{g}_m \sim \mathcal{CN}(0, I_n)$.

for $j \in [m]$ **do**

  Compute the forward differentiation of $\boldsymbol{\nabla}\phi$ at $z\boldsymbol{g}_j$ through degree $N-1$.

  Set

$$\boldsymbol{c}_r(\boldsymbol{g}_j) \leftarrow \frac{1}{r}[z^{r-1}]\boldsymbol{\nabla}\phi(z\boldsymbol{g}_j), \qquad 2 \le r \le N\,.$$

for $i \in [n]$ **do**

  Compute

$$\bar{Z}_{i,r}^{(j)} \leftarrow \mathrm{Re}\,\frac{\boldsymbol{e}_i^\top \boldsymbol{c}_r(\boldsymbol{g}_j)(\boldsymbol{g}_j^\dagger \boldsymbol{e}_i)^{r-1}}{(r-1)!}\,, \qquad 2 \le r \le N\,, j \in [m]\,.$$

  For every $2 \le r \le N$, compute

$$\widehat{Z}_{i,r} \leftarrow \mathrm{Trim}_k\big(\bar{Z}_{i,r}^{(1)}, \dots, \bar{Z}_{i,r}^{(m)}\big)\,,$$
$$\widehat{v}_{i,r} \leftarrow \frac{\binom{2r-2}{r-1}}{(r-1)!}\,\mathrm{Trim}_k\big(|\boldsymbol{e}_i^\top \boldsymbol{c}_r(\boldsymbol{g}_1)|^2, \dots, |\boldsymbol{e}_i^\top \boldsymbol{c}_r(\boldsymbol{g}_m)|^2\big)\,.$$

  Set $\widehat{s}_i \leftarrow \sum_{r=2}^N \sqrt{\widehat{v}_{i,r}}$.

Let $F \subseteq [n]$ index the K largest values of $\widehat{s}_i$, breaking ties arbitrarily.

Evaluate $\phi(\boldsymbol{e}_i)$ exactly for every $i \in F$.

Return estimates

$$\widehat{L}_i = \begin{cases} \phi(\boldsymbol{e}_i)\,, & i \in F\,, \\ \phi(\boldsymbol{0}) + \boldsymbol{e}_i^\top \boldsymbol{\nabla}\phi(\boldsymbol{0}) + \sum_{r=2}^N \widehat{Z}_{i,r}\,, & i \notin F\,. \end{cases}$$

Moreover, the second- and fourth-moment bounds in Lemma 4.2(i) give the following mean and variance for the squared contractions:

$$\mathop{\mathbb{E}}_{g_j} |e_i^\top c_r(g_j)|^2 = (r-1)! \left\| T^{(r)}[e_i, \cdot] \right\|_F^2,$$

and

$$\mathop{\mathrm{Var}}_{g_j} \left(|e_i^\top c_r(g_j)|^2 \right) \le \binom{2r-2}{r-1} \left(\mathop{\mathbb{E}}_{g_j} |e_i^\top c_r(g_j)|^2 \right)^2 \le 4^{r-1} \left(\mathop{\mathbb{E}}_{g_j} |e_i^\top c_r(g_j)|^2 \right)^2.$$

Multiplying these observations by $\binom{2r-2}{r-1}/(r-1)!$ gives mean $v_{i,r}$ and variance at most $4^{r-1} v_{i,r}^2$. Thus the same sketches estimate both the degree-r value and its variance proxy.

Apply the guarantees of the trimmed mean estimator (Proposition 4.8) to each variance-proxy estimate and value estimate, each with failure probability $\delta/(2nN)$. Suppose

$$m \ge 256 \cdot 4^{N-1} \log(32nN/\delta).$$

By a union bound, with probability at least $1-\delta$ over the random complex Gaussians $g_1, \dots, g_m$, simultaneously for every $i \in [n]$ and $2 \le r \le N$,

$$|\widehat{Z}_{i,r} - T^{(r)}[e_i^{\otimes r}]| \le 7\sqrt{\frac{v_{i,r} \log(32nN/\delta)}{m}},$$
$$\frac{1}{2} v_{i,r} \le \widehat{v}_{i,r} \le \frac{3}{2} v_{i,r}.$$

Consequently,

$$\frac{s_i}{\sqrt{2}} \le \widehat{s}_i \le \sqrt{\frac{3}{2}}\, s_i.$$

Selecting Exact Evaluations. By stability at $B = 4$, $\left\| T^{(r)} \right\|_F \le \alpha(4)/4^r$.

Since $\sum_i \left\| T^{(r)}[e_i, \cdot] \right\|_F^2 = \left\| T^{(r)} \right\|_F^2$, Minkowski's inequality gives

$$\begin{aligned} \left(\sum_{i=1}^n s_i^2 \right)^{1/2} &\le \sum_{r=2}^N \sqrt{\binom{2r-2}{r-1}} \left\| T^{(r)} \right\|_F \\ &\le \alpha(4) \sum_{r=2}^\infty \frac{2^{r-1}}{4^r} = \frac{\alpha(4)}{4}. \end{aligned}$$

We now apply Markov's inequality to bound the error proxies of queries that are not selected. Let F be the set of selected queries and fix $i \notin F$. Every selected query $j \in F$ has $\widehat{s}_j \ge \widehat{s}_i$, and hence $s_j^2 \ge s_i^2/3$. Averaging over the K selected queries gives

$$s_i^2 \le \frac{3}{K} \sum_{j \in F} s_j^2 \le \frac{3\alpha(4)^2}{16K}.$$

Adding the degree-wise error bounds gives

$$\left|\sum_{r=2}^{N}\widehat{Z}_{i,r}-\sum_{r=2}^{N}T^{(r)}[e_i^{\otimes r}]\right|\le 7s_i\sqrt{\frac{\log(32nN/\delta)}{m}}$$
$$\le 4\alpha(4)\sqrt{\frac{\log(32nN/\delta)}{mK}}.$$

The constant and linear terms are exact. By Lemma 4.6, the Taylor remainder at e_i satisfies

$$\left|\phi(e_i)-\sum_{r=0}^{N}T^{(r)}[e_i^{\otimes r}]\right|\le\frac{4^{-N}}{3}\alpha(4).$$

Combining these two bounds ensures that each query i that was not selected satisfies

$$|\widehat{L}_i-\phi(e_i)|\le\alpha(4)\left(\frac{4^{-N}}{3}+4\sqrt{\frac{\log(32nN/\delta)}{mK}}\right).$$

The error is zero for selected queries as we evaluate those exactly.

Parameter Choice And Complexity. Choose

$$N=\log_4(1/\varepsilon),\qquad K=\frac{1}{\varepsilon},\qquad m=\frac{256}{\varepsilon}\log(32nN/\delta).$$

Since $4^{N-1}\le\varepsilon^{-1}$, this satisfies the sample-size condition for estimating the variance proxies. For every query that is not selected, the error is at most

$$\alpha(4)\left(\frac{\varepsilon}{3}+\frac{\varepsilon}{4}\right)\le\varepsilon\alpha(4).$$

This proves Theorem 5.1(ii).

By Propositions 4.3 and 4.4, each backward–forward sketch computes all retained coefficients in $O(\mathrm{size}(\phi)N\log N)$ arithmetic operations and $O(\mathrm{size}(\phi)N)$ auxiliary space. Thus the m sketches take $O(\mathrm{size}(\phi)mN\log N)$ arithmetic operations. Using successive powers, forming all observations takes $O(nmN)$ operations. Each trimmed mean can be computed in $O(m)$ time by selection, and the exact evaluations take $O(\mathrm{size}(\phi)K)$ operations. Since $\mathrm{size}(\phi)\ge n$, the total is

$$O\big(\mathrm{size}(\phi)(mN\log N+K)\big).$$

Each sketch stores one length-n direction and $N-1$ length-n coefficient vectors, using $O(nmN)$ space in total. Together with the Taylor evaluations, this uses auxiliary space

$$O\big(\mathrm{size}(\phi)N+nmN\big).$$

Substituting the parameters proves Theorem 5.1(i). A single sketch returns every retained degree simultaneously, so estimating the degrees separately does not increase the number of model calls. The total number of model calls is $\min\{n,m+K\}$, which proves Theorem 5.1(iii). □

5.3 Query Complexity Lower Bound

In this subsection, we present a lower bound for the number of evaluations that is required to estimate all LOO values of an α-stable arithmetic circuit ϕ in an oracle access model that matches the way our algorithm accesses ϕ.

Oracle Access. We assume we can obtain $\phi(\mathbf{0})$ and $\nabla\phi(\mathbf{0})$ for free. For $\boldsymbol{w}, \boldsymbol{g} \in \mathbb{C}^n$ and an integer $N \geq 1$, the oracle provides

$$(\boldsymbol{w}, \boldsymbol{g}, N) \longmapsto \big(\phi(\boldsymbol{w}), \nabla\phi(\boldsymbol{w}), D_{\boldsymbol{g}}^1 \nabla\phi(\boldsymbol{w}), \ldots, D_{\boldsymbol{g}}^N \nabla\phi(\boldsymbol{w})\big).$$

Each query counts as one unit, regardless of N, and its inputs may depend on all previous responses. These queries include the backward–forward sketches and exact evaluations used by our algorithm. We remark that this is a black-box oracle lower bound and does not rule out algorithms with unrestricted access to the circuits.

We use the following lower bound for diagonal estimation.

Lemma 5.2 (Diagonal Estimation Lower Bound; Adapted From [ACDH+24, Lemma 3 and Fact 2])**.** *There is a universal constant $c > 0$ such that the following holds. Any randomized adaptive procedure that, for every real positive semidefinite $n \times n$ matrix A with $\|A\|_F \leq 1/16$, outputs estimates satisfying*

$$\Pr_{\text{seed}}\left(\sum_{i=1}^{n} |\widehat{L}_i - A_{ii}|^2 \leq \frac{c^2}{n}\right) \geq \frac{2}{3}$$

requires $\Omega(n)$ real matrix–vector queries.

We include a proof in Appendix C for the sake of completeness.

Theorem 5.3 (Oracle Query Lower Bound)**.** *For every $n \geq 1$ and sufficiently small $\varepsilon > 0$, any randomized adaptive procedure that, for every real symmetric $n \times n$ matrix A with $\|A\|_F \leq 1/16$, estimates the LOO values of $\phi_A(\boldsymbol{w}) = \boldsymbol{w}^\top A \boldsymbol{w}$ with*

$$\Pr_{\text{seed}}\left(\max_{i\in[n]} |\widehat{L}_i - A_{ii}| \leq \varepsilon\right) \geq \frac{2}{3}$$

requires $\Omega(\min\{n, \varepsilon^{-1}\})$ queries in the preceding oracle model. The probability is over the procedure's random seed. These functions are α-stable for $\alpha(B) = B^2/16$, so $\alpha(4) = 1$.

For constant failure probability, Theorem 5.3 matches the dependence on $1/\varepsilon$ in Theorems 5.1(ii) and 5.1(iii) up to logarithmic factors.

Proof. For a quadratic $\phi_A(\boldsymbol{w}) = \boldsymbol{w}^\top A \boldsymbol{w}$,

$$\nabla\phi_A(\boldsymbol{w}) = 2A\boldsymbol{w}, \qquad D_{\boldsymbol{g}}^1 \nabla\phi_A(\boldsymbol{w}) = 2A\boldsymbol{g}, \qquad D_{\boldsymbol{g}}^r \nabla\phi_A(\boldsymbol{w}) = 0 \quad (r > 1).$$

The oracle response is therefore determined by $A\boldsymbol{w}$ and $A\boldsymbol{g}$. Note that each complex matrix–vector product can be simulated using two real products, by applying A separately to the real and imaginary parts. Thus one oracle query can be simulated using a constant number of real matrix–vector queries.

Let c be the constant in Lemma 5.2 and set

$$n' = \min\{n, c/\varepsilon\}.$$

Apply the hypothetical LOO estimation procedure to positive semidefinite $n' \times n'$ matrices A, extended by zero-padding to $n \times n$ matrices. On every successful run,

$$\sum_{i=1}^{n'} |\widehat{L}_i - A_{ii}|^2 \le n'\varepsilon^2 \le \frac{c^2}{n'}.$$

The procedure therefore satisfies the diagonal-estimation guarantee in dimension n'. By Lemma 5.2, this means that the hypothetical procedure must use $\Omega(n')$ real matrix–vector queries, which corresponds to $\Omega(\min\{n, \varepsilon^{-1}\})$ oracle calls.

Finally, $T^{(2)}_{\phi_A} = A$ and every other Taylor tensor is zero. Since $\|A\|_F \le 1/16$, these functions are α-stable for $\alpha(B) = B^2/16$, with $\alpha(4) = 1$. □

6 Experiments

In this section, we examine Taylor coefficient decay and compare backward–forward sketches with forward-only sketches on a small retrieval task. We also evaluate how well the standard deviation proxy tracks the empirical standard deviation across trials.

Setup. We perform LOO attribution for entire context chunks rather than individual tokens. The only adjustment from Section 5 is that we assign one downweight w_i to all tokens in chunk i. We prompt `Llama-3.2-1B-Instruct`[2] at 64-bit precision with a retrieval task whose answer is in one of 12 chunks (the 7th chunk) and all other chunks contain irrelevant information. We show the full prompt below.

```
System: Answer the question using only the provided text.
User: Documents:
[1] The Calder Museum has opened an exhibition of ancient coins recovered
    from a shipwreck off the Pellan coast. Divers spent two summers bringing
    the pieces to the surface, and conservators cleaned each one by hand.
    The exhibition will stay open until the end of the year, and entry is
    free for children under twelve.
[2] The most borrowed book at the Pinecrest public library this year was a
    novel about a young painter who leaves home to find her lost brother.
    The librarians say they had to order six extra copies to keep up with
    demand. The author, Celia Moraine, will visit the library next season to
    read from the book.
[3] Four climbers from the Ardel Alpine Club reached the summit of Mount
    Keldar after a long ascent of its northern ridge. Their food and tents
    were carried up the lower slopes by a string of camels hired from a
    nearby village. The expedition leader, Sigrun Halvo, said the view from
    the top was worth every freezing night on the ice.
```

[2] https://huggingface.co/unsloth/Llama-3.2-1B-Instruct

[4] From next month, the overnight train between Ostbury and Lanmere will run only on Fridays. The railway company says the change will free up carriages for its busy coastal routes. Passengers who already hold tickets for other nights can exchange them at any station without paying a fee. A new timetable will be posted on the station noticeboards soon.

[5] On the first morning of their trip down the Ormsby River, Marta Jensen caught a trout that weighed almost four kilograms. Her brother had brought a new rod but spent most of the day untangling his line. They cooked their catch over a small fire and returned to the cabin before dark.

[6] At the Eastfield garden show, first prize went to a giant turnip grown by Walter Pemberton, who has kept an allotment for forty years. He said the secret was rainwater collected in barrels and a lot of patience. The judges also praised his neat rows of herbs and the hand-painted signs along his paths.

[7] Chemists at the Varnell Institute have created a ceramic coating that turns blue when it is heated above four hundred degrees. The coating could warn engineers when pipes in power plants begin to overheat. Professor Tomasz Ellery, who led the project, said the change is fully reversible and survives thousands of heating cycles.

[8] Last spring the Hollowbrook bakery introduced a new pastry filled with lemon jam and toasted almonds. The head baker, Imelda Corvan, spent three months testing the recipe in a small kitchen behind the shop. On opening day the queue stretched past the post office, and the first tray sold out in under an hour. Regular customers now reserve a box every weekend.

[9] The lantern festival in Sorrel Bay is held every year in April, when the evenings are mild and the bay is calm. Thousands of paper lanterns are released over the water at sunset, and the streets are decorated with ribbons. Visitors are advised to book rooms early because the town has only three small hotels.

[10] A new statue of the poet Edwin Larke now stands in Harlow Square. The sculptor cast it in bronze at a small foundry outside the city, and it took nine months to finish. On the unveiling day the mayor read one of Larke's poems aloud, and pigeons settled on the statue's shoulders within minutes.

[11] The spring concert at Marlowe Hall featured a young soloist, Yuki Hartmann, playing a rarely heard piece for oboe and orchestra. The audience gave her a standing ovation that lasted nearly five minutes. Critics praised her calm stage presence and the clarity of the slow movement.

[12] The winter soup at the Dovecote Inn is made with barley, smoked ham and plenty of black pepper. The recipe has been handed down by the innkeeper's family for four generations. Travellers who arrive cold from the moors often order a second bowl, and the inn sells jars of it at the counter.

Question: What color does the Varnell Institute coating turn when it is heated?

The evaluation function ϕ is the sum of the token log-probabilities of the following completion, generated once with greedy decoding. We hold this completion fixed across deletion weights and trials.

```
Assistant: The Varnell Institute coating turns blue when it is heated above
400 degrees
```

We implement backward–forward sketches in JAX, using its `jet` module to apply forward differentiation to the value-and-gradient computation of ϕ. We use `complex128` for complex arithmetic. Each trial runs on one NVIDIA H100 GPU.

Empirically, context attributions are often concentrated on a few sources [CSGM24; LKR25]. Prior work uses gradient norms and embedding similarities to rank influential sources [LKR25]. In our setting, chunk 7 contains the answer, so deleting it is expected to strongly affect the completion's log-probability. We perform a standard pre-processing step by ranking queries by $|\partial_i \phi(\mathbf{0})|$ and evaluate the query with the largest score, chunk 7, exactly. We then set $g_7 = 0$ and sketch the remaining indices. The other coordinates are independent standard complex Gaussians. This applies the sketch to the restricted function $\phi\big|_{w_7=0}$, with chunk 7 kept in the prompt. Our analysis in Section 5 holds under the stability of this restricted function.

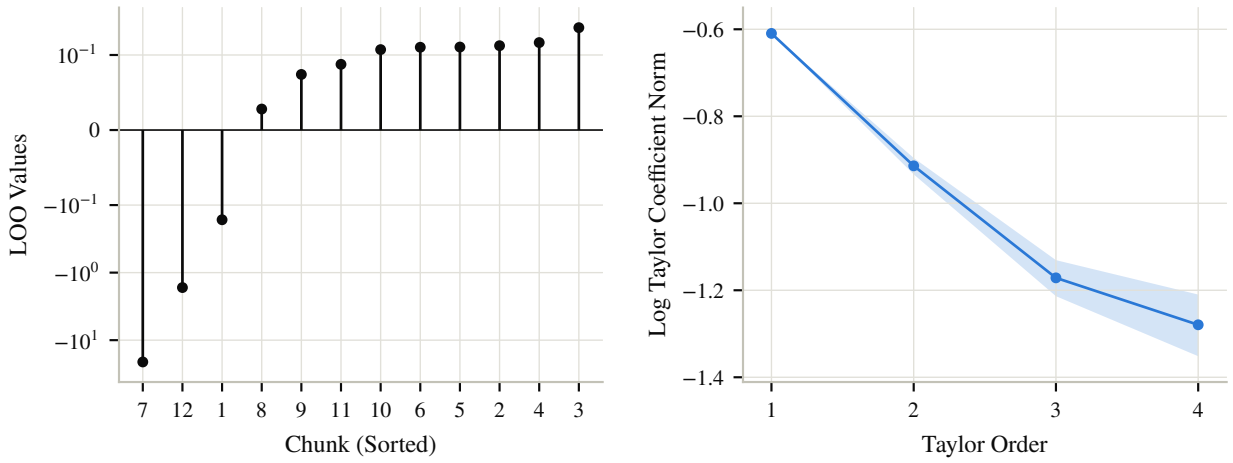
We run 128 independent trials with $m = 8$ sketches per trial. For each chunk and degree, each trimmed mean removes the smallest and largest observations ($k = 1$) and averages the remaining six. We report Taylor degrees $N = 1, \ldots, 4$. Figures 1, 2 and 3 show the leave-one-out values and restricted Taylor coefficient norms, the Taylor approximations against both sketches, and the standard deviation proxies, respectively. The experiments in this section are extremely minimal, and intended as a proof of concept. For this small setting of $n = 12$, it may be faster to just evaluate all LOO values exactly with brute force. The key benefit of our algorithm is that the number of required sketches should scale logarithmically with n rather than linearly, as in the brute-force algorithm.

Exact Values & Stability. Figure 1 shows that removing chunk 7 causes the largest decrease in the completion's log-probability. Panel (a) plots the exact differences $\phi(\boldsymbol{e}_i) - \phi(\mathbf{0})$ for all twelve chunks, sorted by value on a symmetric logarithmic scale. Panel (b) shows the estimated log Frobenius norms of the Taylor tensors of the restricted function $\phi\big|_{w_7=0}$. We estimate these norms from all 1,024 Gaussian directions using the second-moment identity in Lemma 4.2(i). The shaded region is a 90% bootstrap interval computed from 1,000 resamples. The estimated norms decrease over the four displayed orders.

Estimates via Sketching. Figure 2 compares reference Taylor polynomials with our backward–forward estimates and the forward-sketch baseline (Theorem A.1). For a deletion weight $z \in [0, 1]$, the target is $\phi(z\boldsymbol{e}_i) - \phi(\mathbf{0})$, and its degree-$N$ Taylor approximation is

$$\sum_{r=1}^{N} z^r T^{(r)}[\boldsymbol{e}_i^{\otimes r}].$$

We compute the reference coefficients by numerical contour integration, checking agreement across integration radii. The sketch curves show the mean across trials before exact correction,



(a) Leave-one-out values

(b) Taylor coefficient norms

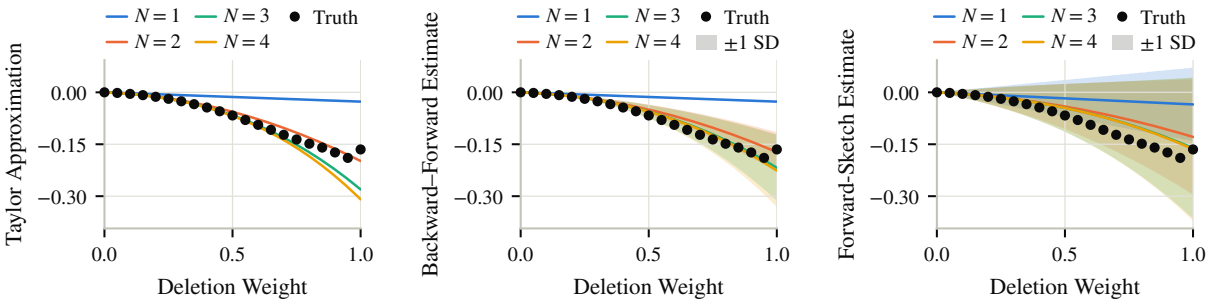
Figure 1: **The restricted function exhibits coefficient decay consistent with stability.** (a) Exact LOO differences $\phi(\boldsymbol{e}_i) - \phi(\mathbf{0})$, sorted by value. (b) Estimated log Frobenius norms of the Taylor tensors of $\phi\big|_{w_7=0}$, for $r = 1, \ldots, 4$. Shading shows a 90% bootstrap interval.

with one empirical standard deviation shaded. At the same sketch budget, backward–forward sketches have smaller standard deviations compared to forward-only sketches.

Error Proxy. Figure 3 compares the empirical standard deviation across trials with the estimated standard deviation proxy

$$\sum_{r=2}^{N} z^r \sqrt{\frac{\widehat{v}_{i,r}}{m}}.$$

The factor z^r accounts for the deletion weight in the degree-r term. We average this proxy across trials and compare it with the standard deviation of the corresponding sketch estimates. For the displayed chunks and degrees, the proxy stays above the empirical standard deviation and tracks its growth as z increases. The proxy describes sampling variability. Taylor truncation and bias from trimming are separate sources of prediction error.



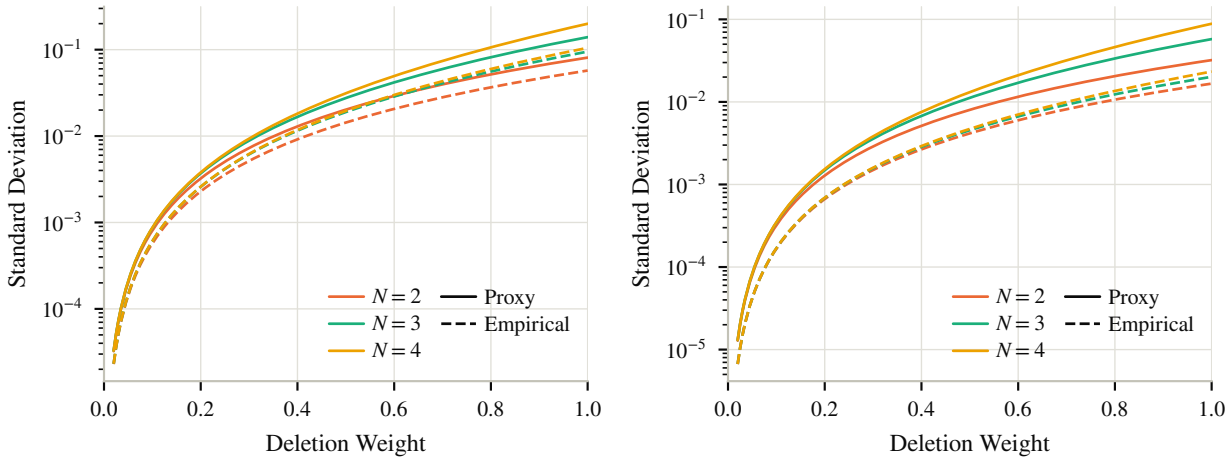
(a) Chunk 1, Taylor (b) Chunk 1, backward sketch (c) Chunk 1, forward sketch

(d) Chunk 4, Taylor (e) Chunk 4, backward sketch (f) Chunk 4, forward sketch

(g) Chunk 6, Taylor (h) Chunk 6, backward sketch (i) Chunk 6, forward sketch

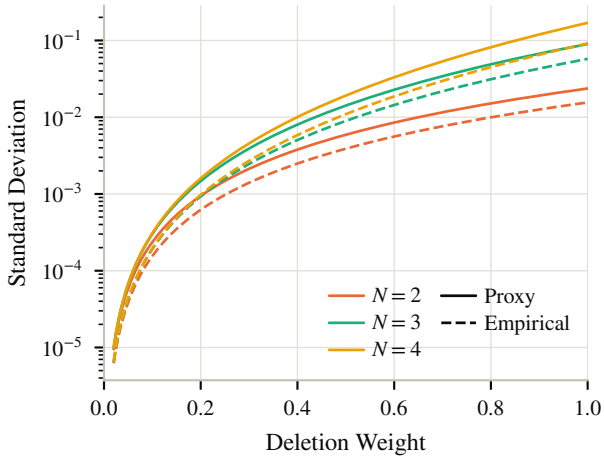
(j) Chunk 12, Taylor (k) Chunk 12, backward sketch (l) Chunk 12, forward sketch

Figure 2: **Backward–forward sketches have lower variance compared to forward-only sketches at the same sketch budget.** Rows correspond to chunks $1, 4, 6, 12$. Columns show reference Taylor polynomials, backward–forward estimates, and forward-sketch estimates for $N = 1, \ldots, 4$. Dots denote direct evaluations of $\phi(z\boldsymbol{e}_i) - \phi(\mathbf{0})$. Sketch curves show trial means with one standard deviation shaded.

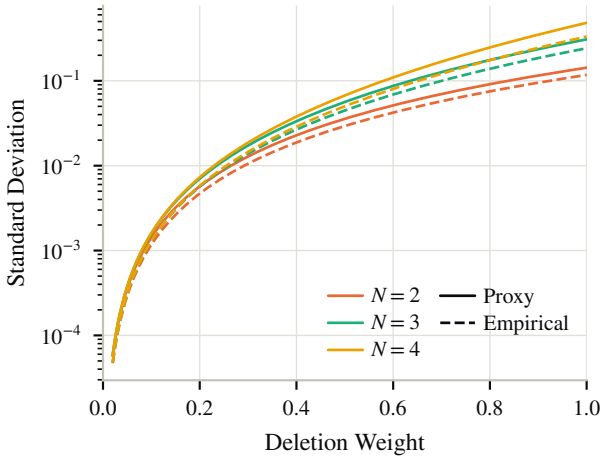


(a) Chunk 1

(b) Chunk 4



(c) Chunk 6



(d) Chunk 12

Figure 3: **The standard deviation proxy closely tracks the empirical standard deviation.** Solid curves show the mean proxy across trials; dashed curves show the empirical standard deviation across trials. Panels correspond to chunks $1, 4, 6, 12$, with Taylor degrees $N = 2, 3, 4$.

References

[ACDH+24] Noah Amsel, Tyler Chen, Feyza Duman Keles, Diana Halikias, Cameron Musco, and Christopher Musco. *Fixed-Sparsity Matrix Approximation from Matrix-Vector Products*. 2024. arXiv: 2402.09379v3 [cs.DS] (cit. on pp. i, 8, 19, 28, 29).

[Bau74] F. L. Bauer. "Computational Graphs and Rounding Error". In: *SIAM Journal on Numerical Analysis* 11.1 (1974), pp. 87–96 (cit. on pp. 3, 9).

[BS83] Walter Baur and Volker Strassen. "The Complexity of Partial Derivatives". In: *Theoretical Computer Science* 22.3 (1983), pp. 317–330 (cit. on p. 2).

[CCSW+25] Yung-Sung Chuang, Benjamin Cohen-Wang, Zejiang Shen, Zhaofeng Wu, Hu Xu, Xi Victoria Lin, James R. Glass, Shang-Wen Li, and Wen-tau Yih. "SelfCite: Self-Supervised Alignment for Context Attribution in Large Language Models". In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2025 (cit. on p. 3).

[CGTM+26] Yuxi Chen, Hamza Golubovic, Han Tong, Arian Maleki, and Andrew Ilyas. *Data Attribution via Sketched Metadifferentiation*. 2026. arXiv: 2609.15044 [stat.ML] (cit. on p. 3).

[CKW11] Xi Chen, Neeraj Kayal, and Avi Wigderson. "Partial Derivatives in Arithmetic Complexity and Beyond". In: *Foundations and Trends in Theoretical Computer Science* 6.1–2 (2011), pp. 1–138 (cit. on p. 11).

[CSGM24] Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. "ContextCite: Attributing Model Generation to Context". In: *Advances in Neural Information Processing Systems*. Vol. 37. 2024 (cit. on pp. 1, 3, 22).

[GBC16] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016 (cit. on pp. 3, 9).

[Gun26] Sam Gunn. "How to sketch a learning algorithm". In: *arXiv preprint arXiv:2604.07328* (2026) (cit. on pp. i, 1, 2, 4, 12, 28).

[GYYC23] Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. "Enabling Large Language Models to Generate Text with Citations". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2023, pp. 6465–6488 (cit. on p. 3).

[Har13] Aram W. Harrow. *The Church of the Symmetric Subspace*. 2013. arXiv: 1308.6595 [quant-ph] (cit. on p. 10).

[IE25] Andrew Ilyas and Logan Engstrom. *MAGIC: Near-Optimal Data Attribution for Deep Learning*. 2025. arXiv: 2504.16430 [cs.LG] (cit. on p. 3).

[IPEL+22] Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. "Datamodels: Understanding Predictions with Data and Data with Predictions". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 9525–9587 (cit. on p. 3).

[KL17] Pang Wei Koh and Percy Liang. "Understanding Black-box Predictions via Influence Functions". In: *Proceedings of the 34th International Conference on Machine Learning*. Vol. 70. Proceedings of Machine Learning Research. PMLR, 2017, pp. 1885–1894 (cit. on p. 3).

[Lin76] Seppo Linnainmaa. "Taylor Expansion of the Accumulated Rounding Error". In: *BIT Numerical Mathematics* 16.2 (1976), pp. 146–160 (cit. on pp. 2, 5).

[LKR25] Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. "AttriBoT: A Bag of Tricks for Efficiently Approximating Leave-One-Out Context Attribution". In: *International Conference on Learning Representations*. 2025 (cit. on pp. 1, 3, 22).

[LL17] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 30. 2017 (cit. on p. 3).

[LM19] Gábor Lugosi and Shahar Mendelson. "Mean Estimation and Regression under Heavy-Tailed Distributions: A Survey". In: *Foundations of Computational Mathematics* 19.5 (2019), pp. 1145–1190 (cit. on p. 13).

[LMJ16] Jiwei Li, Will Monroe, and Dan Jurafsky. "Understanding Neural Networks through Representation Erasure". In: *arXiv preprint arXiv:1612.08220* (2016) (cit. on pp. 1, 3).

[OOR25] Roberto I. Oliveira, Paulo Orenstein, and Zoraida F. Rico. *Finite-Sample Properties of the Trimmed Mean*. 2025. arXiv: 2501.03694 [math.ST] (cit. on pp. 5, 13).

[PGIL+23] Sung Min Park, Kristian Georgiev, Andrew Ilyas, Guillaume Leclerc, and Aleksander Madry. "TRAK: Attributing Model Behavior at Scale". In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, 2023, pp. 27074–27113 (cit. on p. 3).

[PLKS20] Garima Pruthi, Frederick Liu, Satyen Kale, and Mukund Sundararajan. "Estimating Training Data Influence by Tracing Gradient Descent". In: *Advances in Neural Information Processing Systems*. Vol. 33. 2020 (cit. on p. 3).

[QSFB24] Jirui Qi, Gabriele Sarti, Raquel Fernández, and Arianna Bisazza. "Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation". In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2024, pp. 6037–6053 (cit. on pp. 1, 3).

[RH25] Ittai Rubinstein and Samuel Hopkins. "Rescaled Influence Functions: Accurate Data Attribution in High Dimension". In: *Advances in Neural Information Processing Systems*. Vol. 38. 2025 (cit. on p. 3).

[RH26] Ittai Rubinstein and Samuel B. Hopkins. "On the Accuracy of Newton Step and Influence Function Data Attributions". In: *Proceedings of the 43rd International Conference on Machine Learning*. Vol. 306. Proceedings of Machine Learning Research. PMLR, 2026, pp. 106088–106134 (cit. on p. 3).

[RHW86] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536 (cit. on pp. 2, 5, 11).

[RSG16] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2016, pp. 1135–1144 (cit. on p. 3).

[Rud87] Walter Rudin. *Real and Complex Analysis*. 3rd ed. McGraw-Hill, 1987 (cit. on p. 12).

[Wen64] R. E. Wengert. "A Simple Automatic Derivative Evaluation Program". In: *Communications of the ACM* 7.8 (1964), pp. 463–464 (cit. on pp. 5, 12).

A Forward Sketch Baseline

As a baseline comparison to our main Theorem 5.1, we state an adaptation of Gunn's data attribution algorithm via forward differentiation sketching to our context attribution setting.

Theorem A.1 (Forward-Sketch Baseline; Adapted From [Gun26, Theorem 2]). *Let $\phi : \mathbb{C}^n \to \mathbb{C}$ be an α-stable arithmetic circuit that is real-valued on $[0,1]^n$, with $\alpha(4) < \infty$. For any $\varepsilon, \delta \in (0,1)$, forward sketches give estimates satisfying*

$$\max_{i \in [n]} |\widehat{L}_i - \phi(\boldsymbol{e}_i)| \leq \varepsilon \alpha(4)$$

with probability at least $1 - \delta$ over the sketch randomness, using $\widetilde{O}(\mathrm{size}(\phi) \log(n/\delta)/\varepsilon^2)$ operations and $\widetilde{O}(\mathrm{size}(\phi) + n \log(n/\delta)/\varepsilon^2)$ auxiliary space.

B Continuous Masking

We implement the downweights $\boldsymbol{w} \in [0,1]^n$ of Section 5.1 as a continuous attention mask. Fix the model parameters and the tokenized input of length L, and let $p_1 < \cdots < p_n$ be the positions of the context tokens $\tau_1, \ldots, \tau_n$. For each position $j \in [L]$, define its survival weight

$$\chi_j(\boldsymbol{w}) := \begin{cases} 1 - w_i, & \text{if } j = p_i, \\ 1, & \text{otherwise,} \end{cases}$$

so $w_i = 0$ keeps token τ_i and $w_i = 1$ masks it.

Every attention head in every layer uses these weights. At query position t, a head outputs

$$\boldsymbol{o}_t(\boldsymbol{w}) := \frac{\sum_{j=1}^{L} M_{tj}\, \chi_j(\boldsymbol{w}) \exp\big(a_{tj}(\boldsymbol{w})\big)\, \boldsymbol{u}_j(\boldsymbol{w})}{\sum_{j=1}^{L} M_{tj}\, \chi_j(\boldsymbol{w}) \exp\big(a_{tj}(\boldsymbol{w})\big)},$$

where $a_{tj}(\boldsymbol{w})$ and $\boldsymbol{u}_j(\boldsymbol{w})$ are the query–key logits and value vectors, and $M_{tj} \in \{0,1\}$ is the fixed causal mask. We assume every attention row attends to at least one position outside the context, such as a beginning-of-sequence token, so the denominator is positive on $[0,1]^n$. If the logits and values are holomorphic in $\boldsymbol{w}$, then so is $\boldsymbol{o}_t$ wherever the denominator is nonzero.

Setting $\boldsymbol{w} = \boldsymbol{e}_i$ removes the terms for token τ_i from both sums, giving ordinary softmax attention over the remaining tokens. Since all other operations act on each position separately, the remaining hidden states agree layer by layer with a run on the prompt with τ_i deleted, keeping the positions of the other tokens fixed. Hence $\phi(\boldsymbol{e}_i)$ is exactly the LOO value of token τ_i.

C Diagonal Estimation Lower Bound

Lemma 5.2 (Diagonal Estimation Lower Bound; Adapted From [ACDH+24, Lemma 3 and Fact 2]). *There is a universal constant $c > 0$ such that the following holds. Any randomized adaptive procedure that, for every real positive semidefinite $n \times n$ matrix A with $\|A\|_F \leq 1/16$, outputs estimates satisfying*

$$\Pr_{\text{seed}} \left(\sum_{i=1}^{n} |\widehat{L}_i - A_{ii}|^2 \leq \frac{c^2}{n} \right) \geq \frac{2}{3}$$

requires $\Omega(n)$ real matrix–vector queries.

We use the following two results in the proof.

Proposition C.1 (Gaussian Gram Matrix Moment; Consequence Of [ACDH+24, Fact 2])**.** *Let G be an $r \times n$ matrix of independent standard real Gaussians. Then*

$$\mathop{\mathbb{E}}_{G} \left\| G^\top G \right\|_F^2 = nr(r+n+1).$$

Lemma C.2 (Gaussian Gram Diagonal Estimation; cf. [ACDH+24, Lemma 3])**.** *There are universal constants $b, \beta > 0$ and a universal integer $L \geq 1$ such that the following holds for every $n \geq 1$. Let G be an $Ln \times n$ matrix of independent standard real Gaussians. Any randomized adaptive procedure that uses at most bn real matrix–vector queries to $G^\top G$ and outputs $\widehat{L}_1, \ldots, \widehat{L}_n$ satisfies*

$$\Pr_{G,\mathrm{seed}} \left(\sum_{i=1}^{n} |\widehat{L}_i - (G^\top G)_{ii}|^2 < \beta n^2 \right) \leq \frac{1}{4}.$$

This is the diagonal case of the cited lemma, with constant failure probability and a sufficiently large fixed ratio of rows to columns.

Proof of Lemma 5.2. Let b, β, L be the constants in Lemma C.2. Draw an $Ln \times n$ matrix G of independent standard real Gaussians and set

$$A = \frac{G^\top G}{64Ln^{3/2}}.$$

By Proposition C.1,

$$\mathop{\mathbb{E}}_{G} \|A\|_F^2 = \frac{1 + 1/L + 1/(Ln)}{64^2} \leq \frac{3}{64^2}.$$

Since $\|A\|_F^2 \geq 1/16^2$ whenever $\|A\|_F > 1/16$,

$$\Pr_{G}(\|A\|_F > 1/16) \leq 16^2 \mathop{\mathbb{E}}_{G} \|A\|_F^2 \leq \frac{3}{16} < \frac{1}{4}.$$

Choose $c > 0$ so that $(64L)^2 c^2 < \beta$. Suppose a procedure satisfies the claimed accuracy guarantee using at most bn queries. Run it on A, simulating each query by one query to $G^\top G$ and stopping after its query budget. Multiply each returned estimate by $64Ln^{3/2}$. On every successful run,

$$\begin{aligned} \sum_{i=1}^{n} \left| 64Ln^{3/2}\widehat{L}_i - (G^\top G)_{ii} \right|^2 &= (64L)^2 n^3 \sum_{i=1}^{n} |\widehat{L}_i - A_{ii}|^2 \\ &\leq (64L)^2 c^2 n^2 < \beta n^2. \end{aligned}$$

The matrix A is positive semidefinite and lies in the required Frobenius ball with probability at least $3/4$. The procedure's guarantee therefore makes this event have probability at least $(2/3)(3/4) = 1/2$. This contradicts Lemma C.2, which bounds its probability by $1/4$. □

D AI Disclosure

We used Gemini 4 and ChatGPT 5.6 Pro / Astra extensively during the development of the work. We used these tools to brainstorm ideas about technical approaches, carry over tedious calculations, and also help with writing code for the experimental portion of the work. The authors have verified the correctness of the results and take responsibility for them.