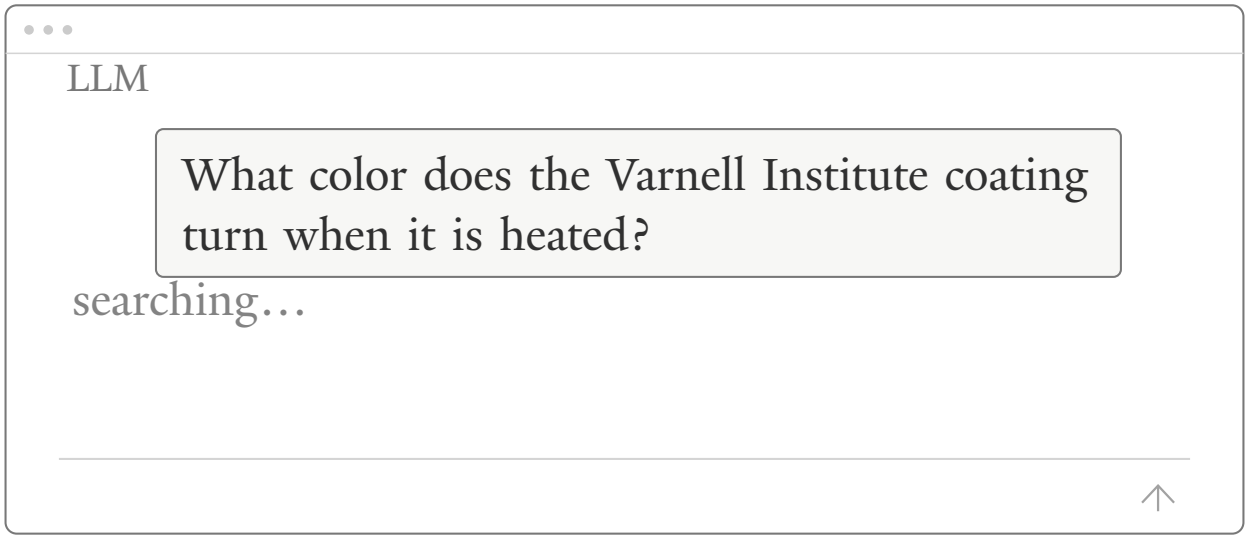


Sketching Language Models for Context Attribution

Christopher Liaw **Aranyak Mehta**
Grigoris Velegkas **Felix Zhou**

Google Research Yale University

Table of Contents

Motivation

Problem Definition

Prior Works

Contributions

Proof Overview

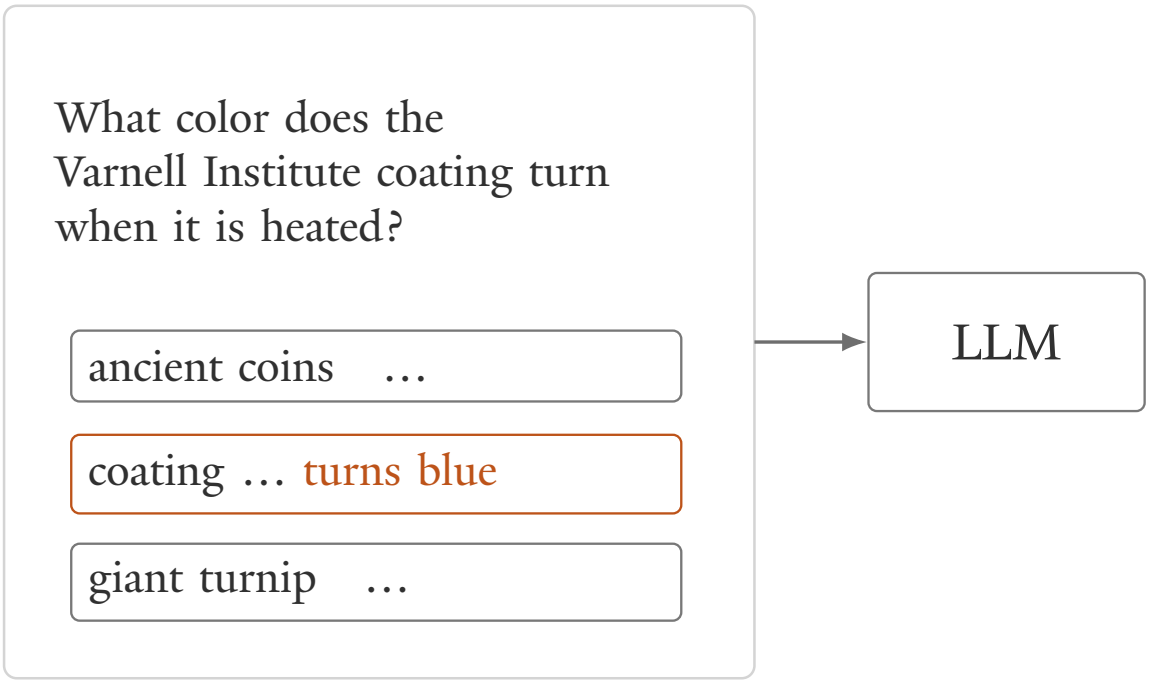
- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

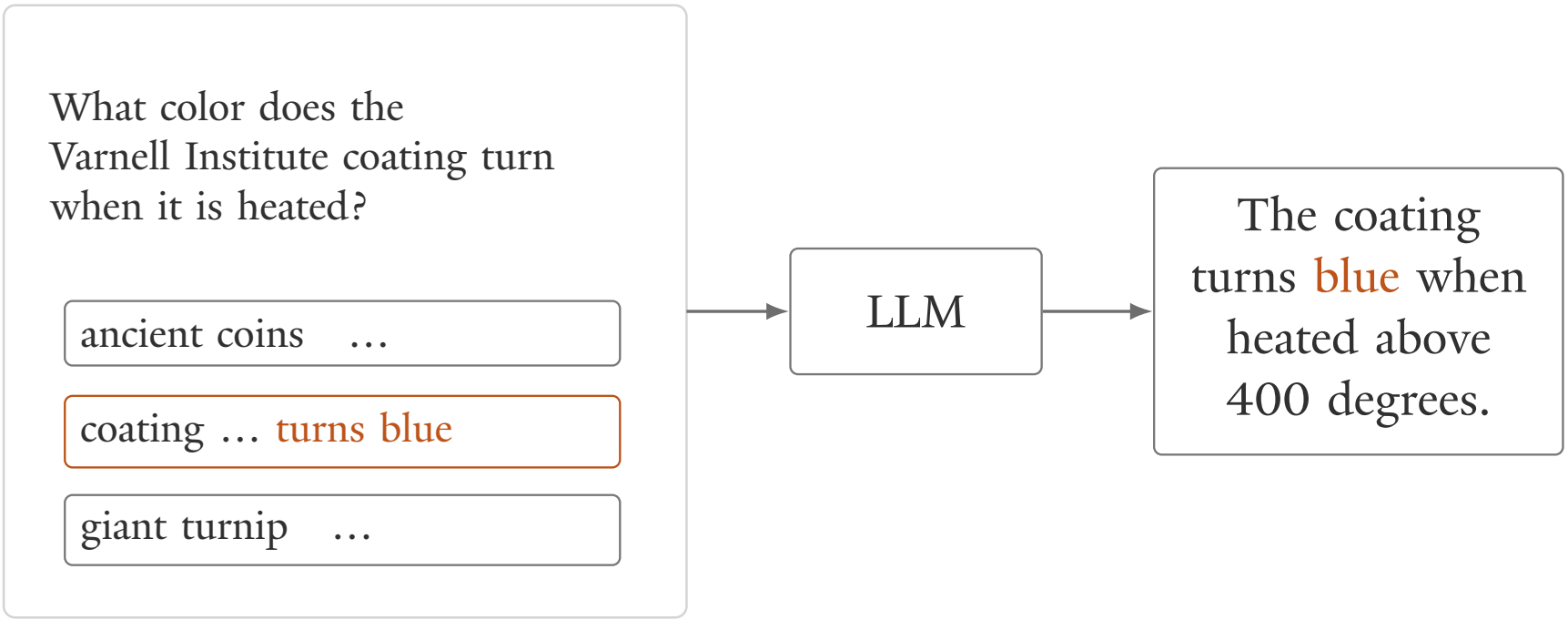
Motivation

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

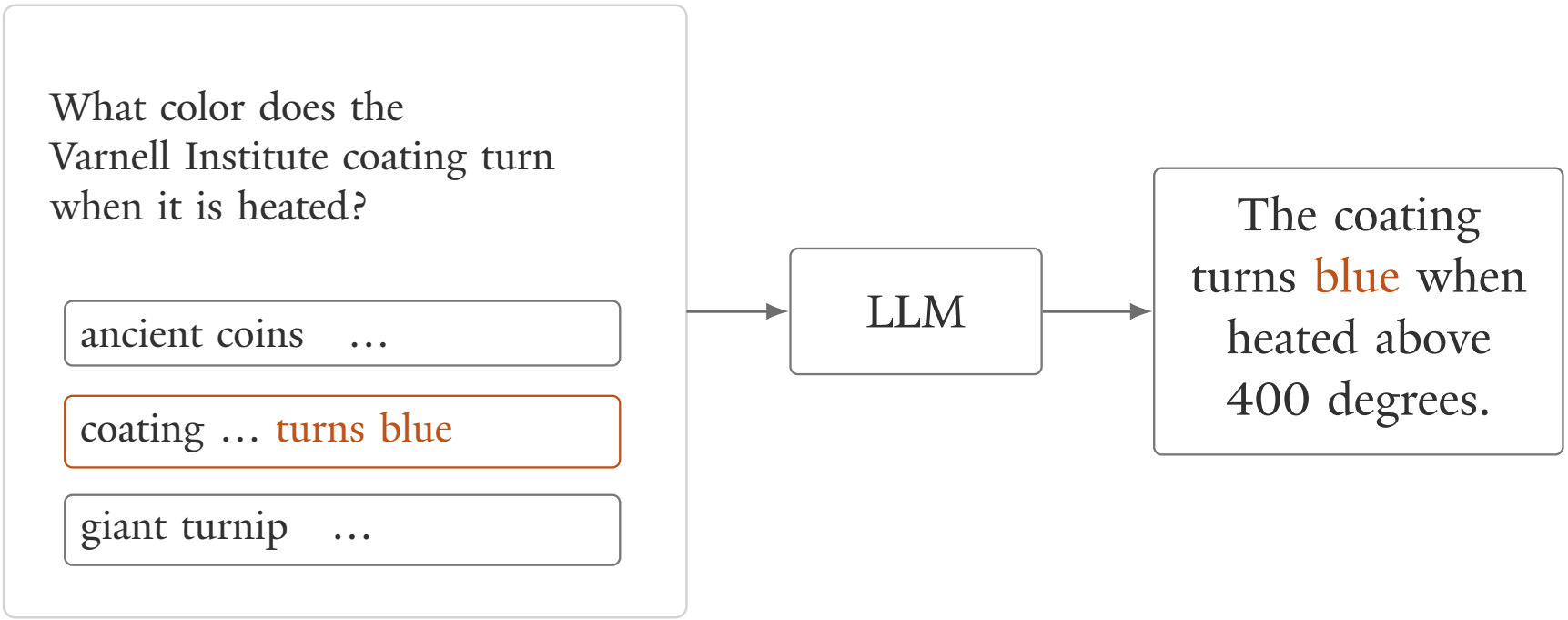
Motivation


What color does the Varnell Institute coating turn when it is heated?
ancient coins …
coating … turns blue
giant turnip …
LLM

Motivation


What color does the Varnell Institute coating turn when it is heated?
ancient coins …
coating … turns blue
giant turnip …
LLM
The coating turns blue when heated above 400 degrees.

Motivation



How much did each context chunk contribute to this answer?

Table of Contents

Motivation

Problem Definition

Prior Works

Contributions

Proof Overview
- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

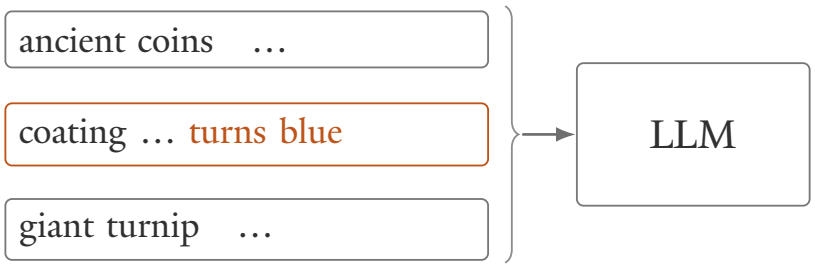
Context Attribution

- Fix an LLM.



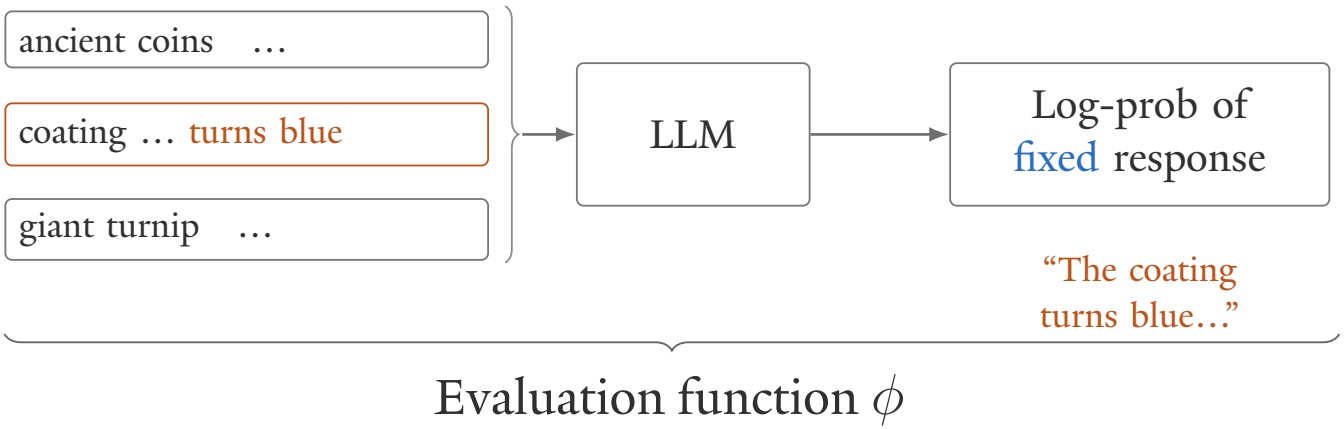
Context Attribution

- Fix an LLM.
- Supply a context of n chunks.



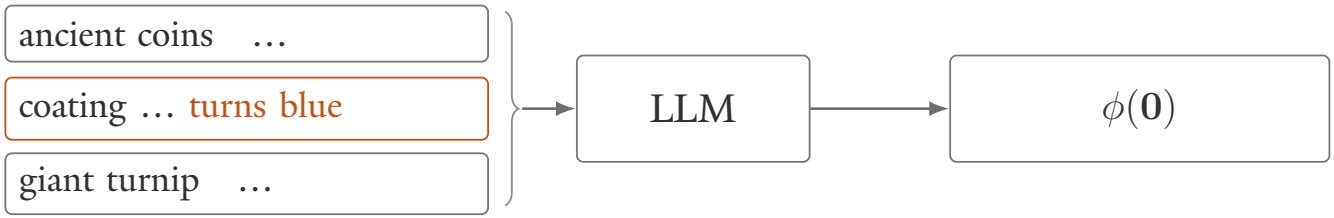
Context Attribution

- Fix an LLM.
- Supply a context of n chunks.
- Fix a scalar metric of the output.
  - Examples: accuracy or log-probability of a fixed response.



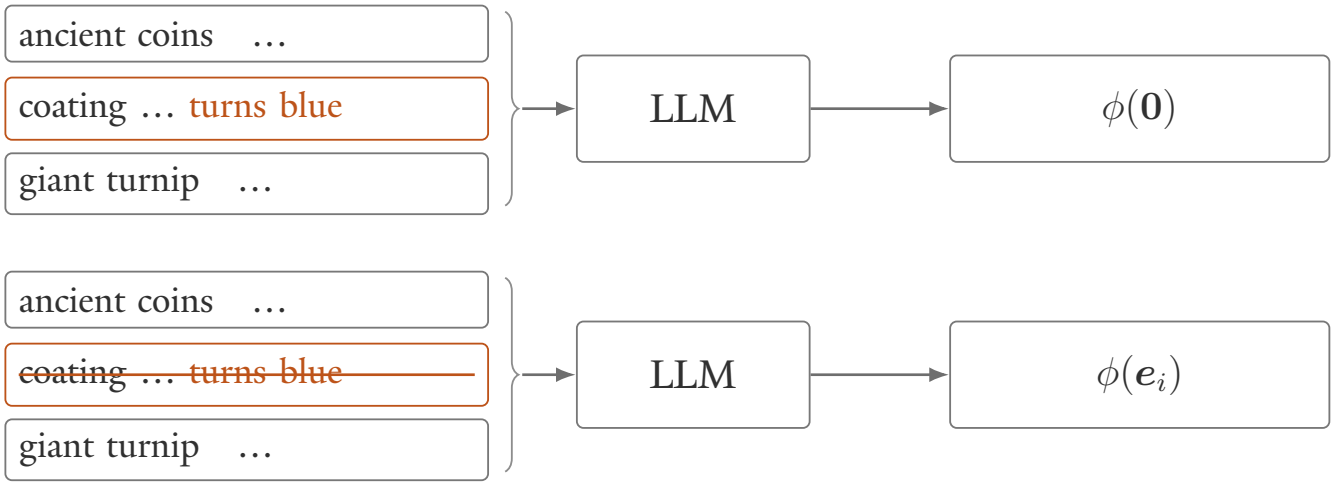
Leave-One-Out Context Attribution

For $\boldsymbol{w} \in \{0,1\}^n$, $w_i = 1$ deletes chunk i.



Leave-One-Out Context Attribution

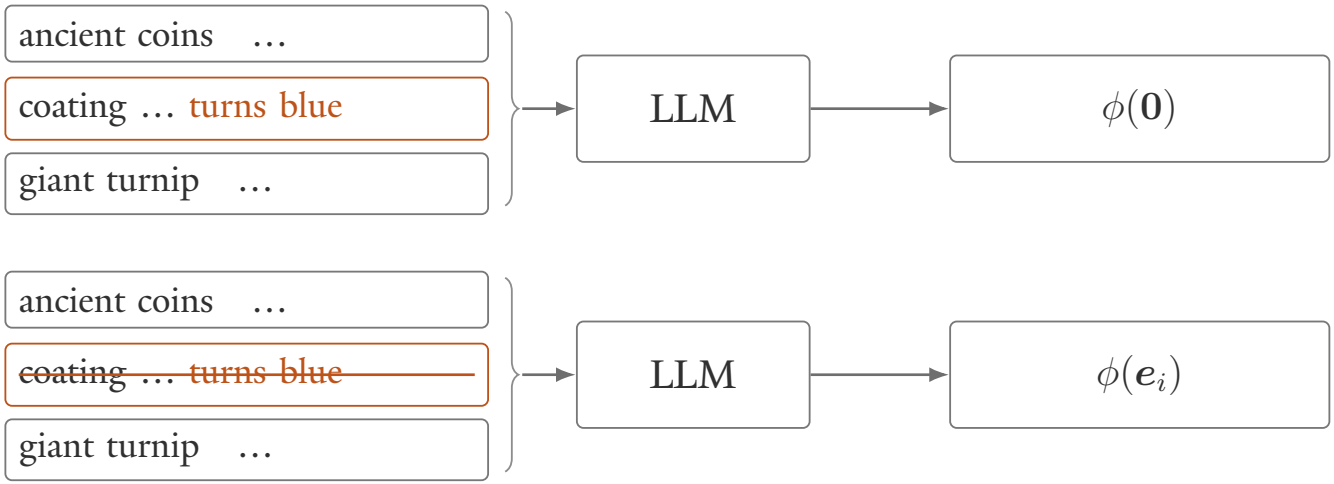
For $\boldsymbol{w} \in \{0,1\}^n$, $w_i = 1$ deletes chunk i.



- Estimate $\phi(\boldsymbol{e}_i) - \phi(\mathbf{0})$ for every $i \in [n]$.

Leave-One-Out Context Attribution

For $\boldsymbol{w} \in \{0,1\}^n$, $w_i = 1$ deletes chunk i.



- Estimate $\phi(\boldsymbol{e}_i) - \phi(\mathbf{0})$ for every $i \in [n]$.
- Extend binary masks to continuous deletion weights:

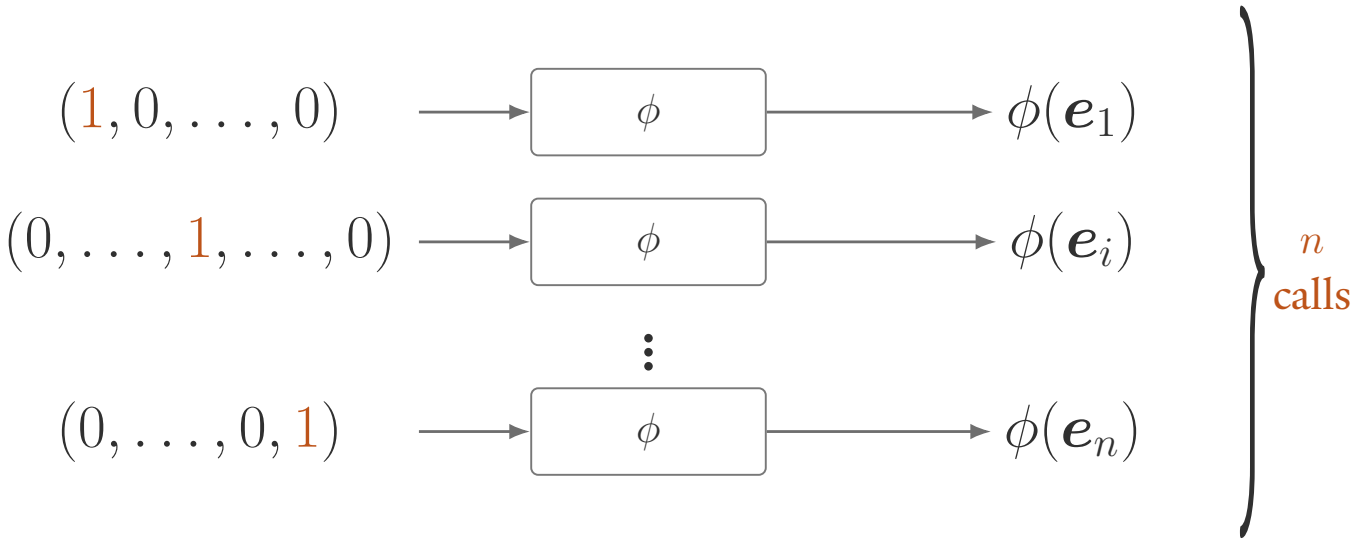
$$\phi : \{0,1\}^n \to \mathbb{R} \qquad \longrightarrow \qquad \phi : [0,1]^n \to \mathbb{R}.$$

The Cost of Context Attribution

- Evaluating ϕ requires a model run.

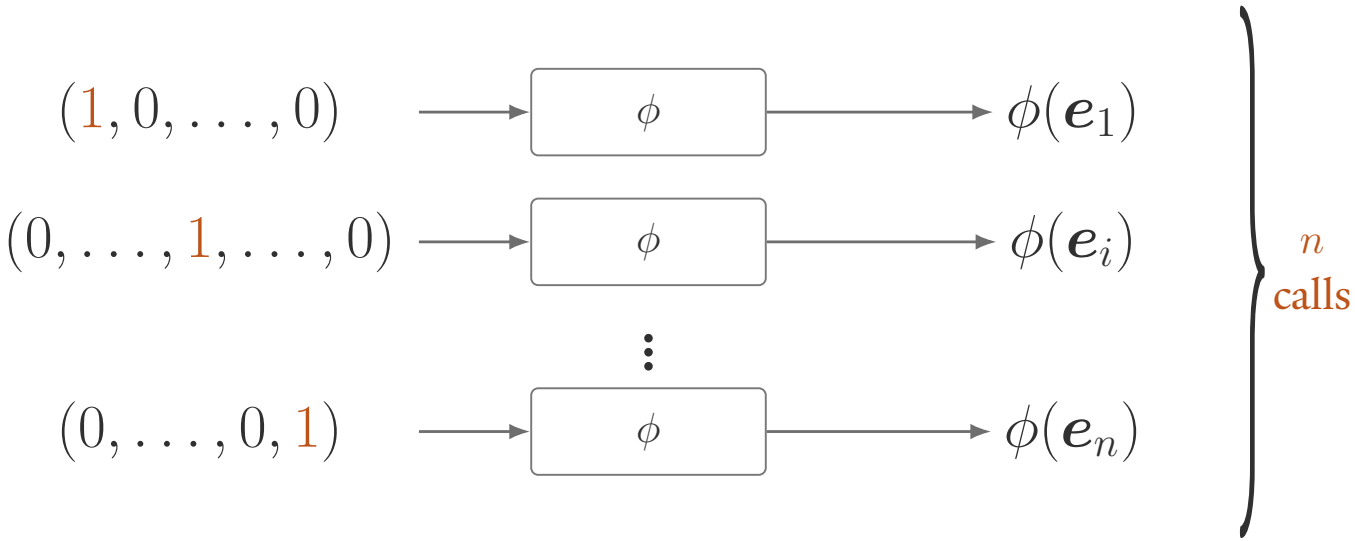
The Cost of Context Attribution

- Evaluating ϕ requires a model run.
- Direct algorithm: one evaluation per context chunk.



The Cost of Context Attribution

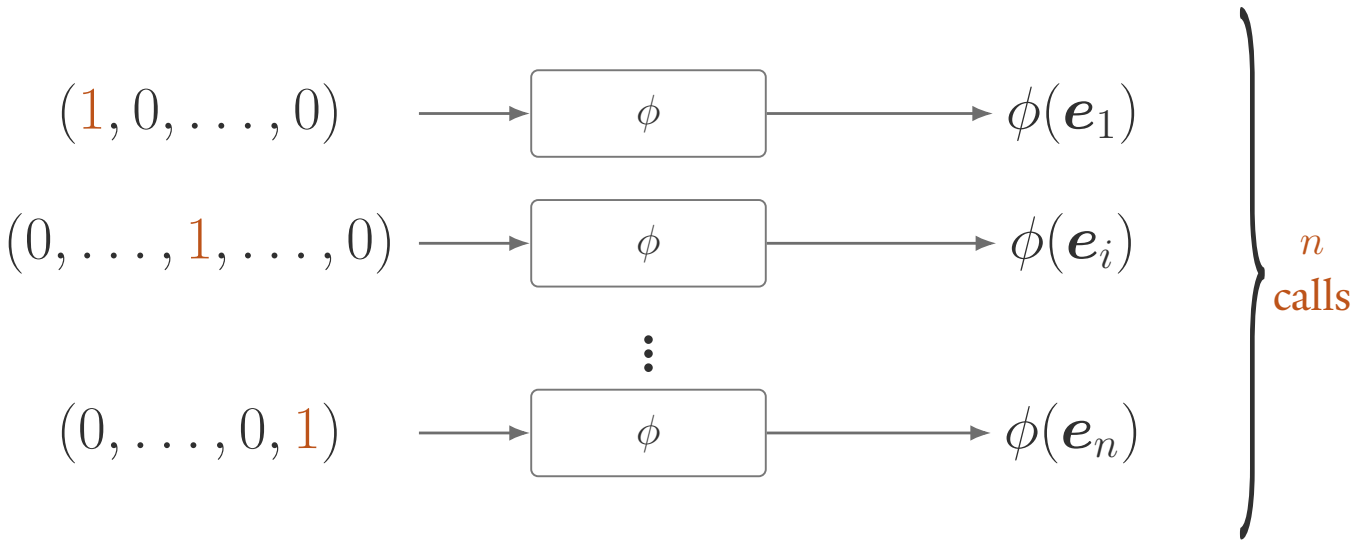
- Evaluating ϕ requires a model run.
- Direct algorithm: one evaluation per context chunk.



Can we avoid n model calls?

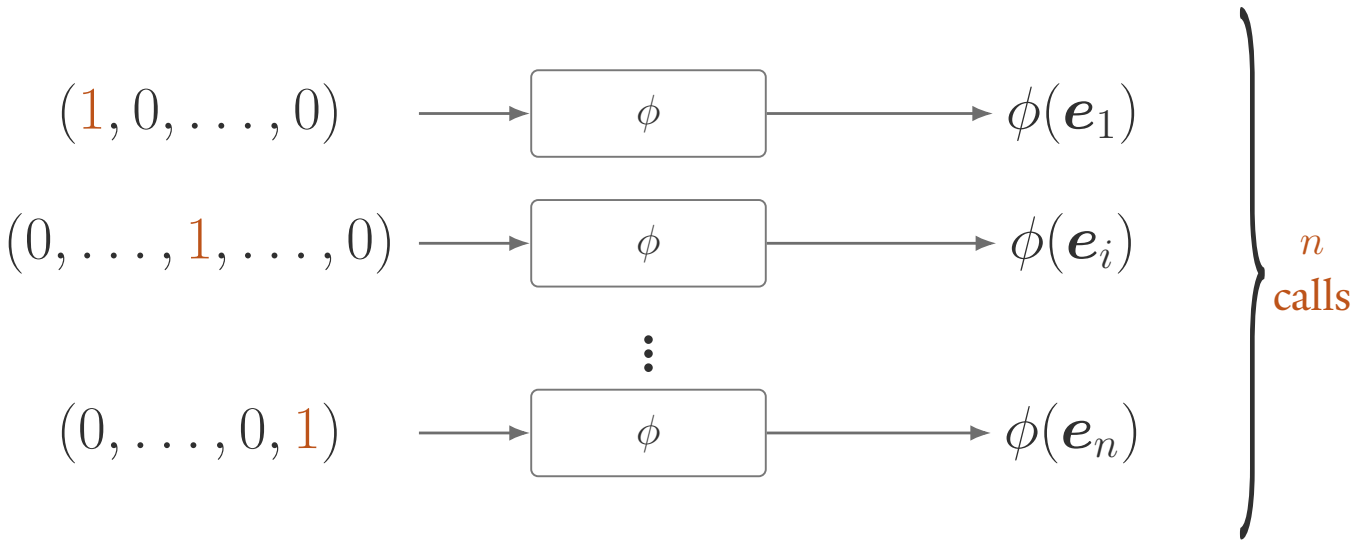
Modeling the Evaluation Function

- Model ϕ as an arithmetic circuit: a computation graph with addition and multiplication gates.



Modeling the Evaluation Function

- Model ϕ as an arithmetic circuit: a computation graph with addition and multiplication gates.
- For other gates, use local polynomial approximations.



Modeling the Evaluation Function

- Model ϕ as an arithmetic circuit: a computation graph with addition and multiplication gates.
- For other gates, use local polynomial approximations.
- Easily implemented on `Llama-3.2-1B-Instruct`.

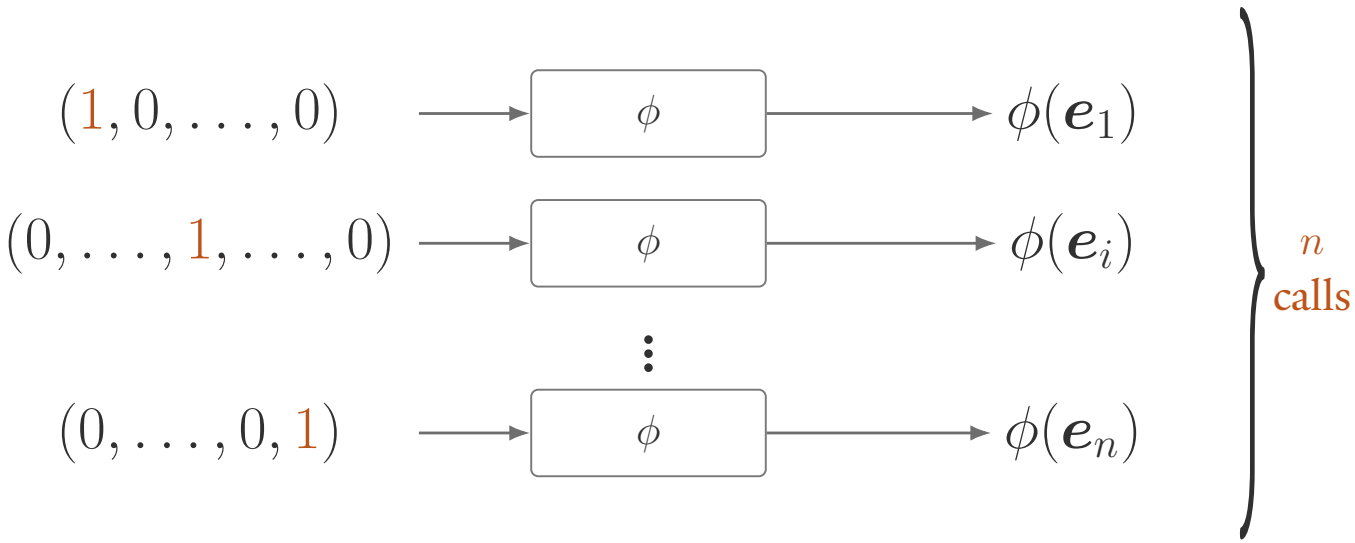


Table of Contents

Motivation

Problem Definition

Prior Works

Contributions

Proof Overview
- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

Prior Works

- **Heuristics.** Sparse surrogates, gradient scores, caching, and proxy models [CSGM24; QSFB24; LKR25].

Prior Works

- **Heuristics.** Sparse surrogates, gradient scores, caching, and proxy models [CSGM24; QSFB24; LKR25].
- **Data attribution.** Gunn [Gun26] sketches stable arithmetic circuits to estimate effects of deleting training data.

Prior Works

- **Heuristics.** Sparse surrogates, gradient scores, caching, and proxy models [CSGM24; QSFB24; LKR25].
- **Data attribution.** Gunn [Gun26] sketches stable arithmetic circuits to estimate effects of deleting training data.
- **Context LOO.** Adapt the sketch to singleton prompt deletions: $\widetilde{O}(\log(n)/\varepsilon^2)$ model calls.

Table of Contents

Motivation

Problem Definition

Prior Works

Contributions

Proof Overview
- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

Our Results

THEOREM 1 (INFORMAL)

Fix a "stable" circuit ϕ. There is an algorithm for LOO context attribution.

- **Accuracy.** With constant probability, all n LOO estimates have error $O(\varepsilon)$.

Theorems 5.1 and 5.3. The upper bound counts sketches and exact evaluations.

Our Results

THEOREM 1 (INFORMAL)

Fix a "stable" circuit ϕ. There is an algorithm for LOO context attribution.

- **Accuracy.** With constant probability, all n LOO estimates have error $O(\varepsilon)$.
- **Upper bound.** $\widetilde{O}\left(\min\left\{n, \frac{\log n}{\varepsilon}\right\}\right)$ model calls.

Theorems 5.1 and 5.3. The upper bound counts sketches and exact evaluations.

Our Results

THEOREM 1 (INFORMAL)

Fix a "stable" circuit ϕ. There is an algorithm for LOO context attribution.

- **Accuracy.** With constant probability, all n LOO estimates have error $O(\varepsilon)$.
- **Upper bound.** $\widetilde{O}\left(\min\left\{n, \frac{\log n}{\varepsilon}\right\}\right)$ model calls.
- **Lower bound.** Any algorithm* requires $\Omega(\min\{n, 1/\varepsilon\})$ calls.

*With the same access to the model.
Theorems 5.1 and 5.3. The upper bound counts sketches and exact evaluations.

Our Results

THEOREM 1 (INFORMAL)

Fix a "stable" circuit ϕ. There is an algorithm for LOO context attribution.

- **Accuracy.** With constant probability, all n LOO estimates have error $O(\varepsilon)$.
- **Upper bound.** $\widetilde{O}\left(\min\left\{n, \frac{\log n}{\varepsilon}\right\}\right)$ model calls.
- **Lower bound.** Any algorithm* requires $\Omega(\min\{n, 1/\varepsilon\})$ calls.

$$1/\varepsilon^2 \quad \longrightarrow \quad 1/\varepsilon$$

*With the same access to the model.
Theorems 5.1 and 5.3. The upper bound counts sketches and exact evaluations.

Remarks

- For fixed accuracy, number of calls grows as $O(\log n)$.

Remarks

- For fixed accuracy, number of calls grows as $O(\log n)$.
- Estimating a coin's bias to error ε can take $\Omega(1/\varepsilon^2)$ flips.

Remarks

- For fixed accuracy, number of calls grows as $O(\log n)$.
- Estimating a coin's bias to error ε can take $\Omega(1/\varepsilon^2)$ flips.
- Yet can estimate all n LOO values with only $\widetilde{O}(1/\varepsilon)$ model calls.

Table of Contents

Motivation

Problem Definition

Prior Works

Contributions

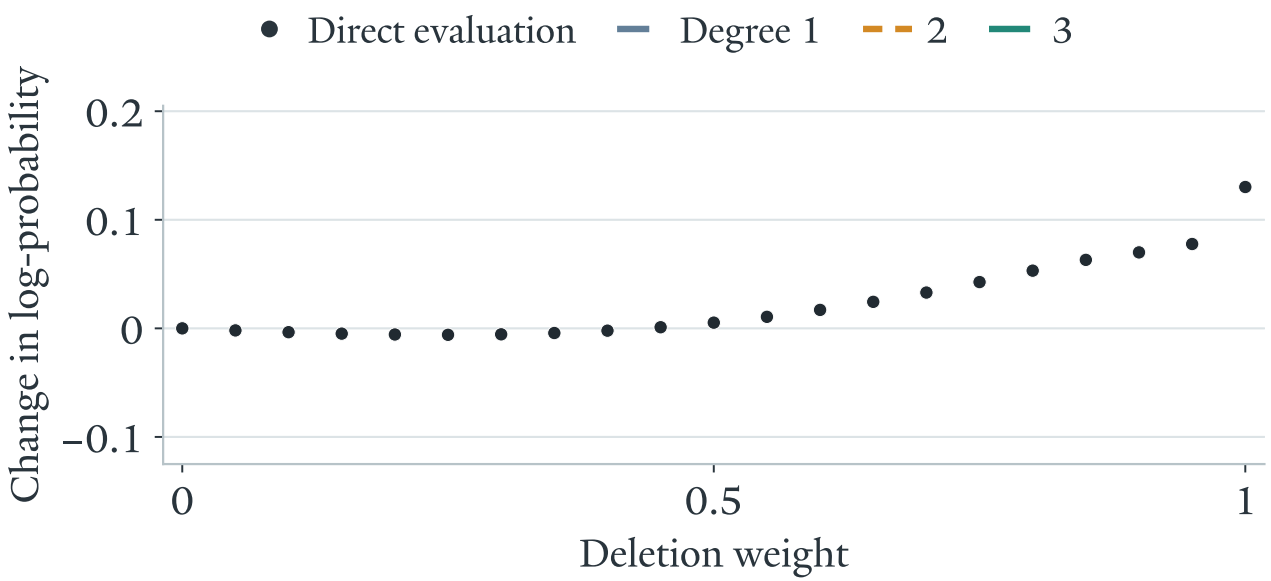
Proof Overview
- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

Taylor Expansion of the Measurement

Around the full context $\mathbf{0}$, for $\boldsymbol{w} \in [0,1]^n$,

$$\phi(\boldsymbol{w}) = \phi(\mathbf{0}) + \nabla\phi(\mathbf{0})^\top \boldsymbol{w} + \tfrac{1}{2}\boldsymbol{w}^\top \nabla^2 \phi(\mathbf{0})\boldsymbol{w} + \cdots .$$



Chunk 6
"giant turnip …"
$\phi(z\boldsymbol{e}_6) - \phi(\mathbf{0})$

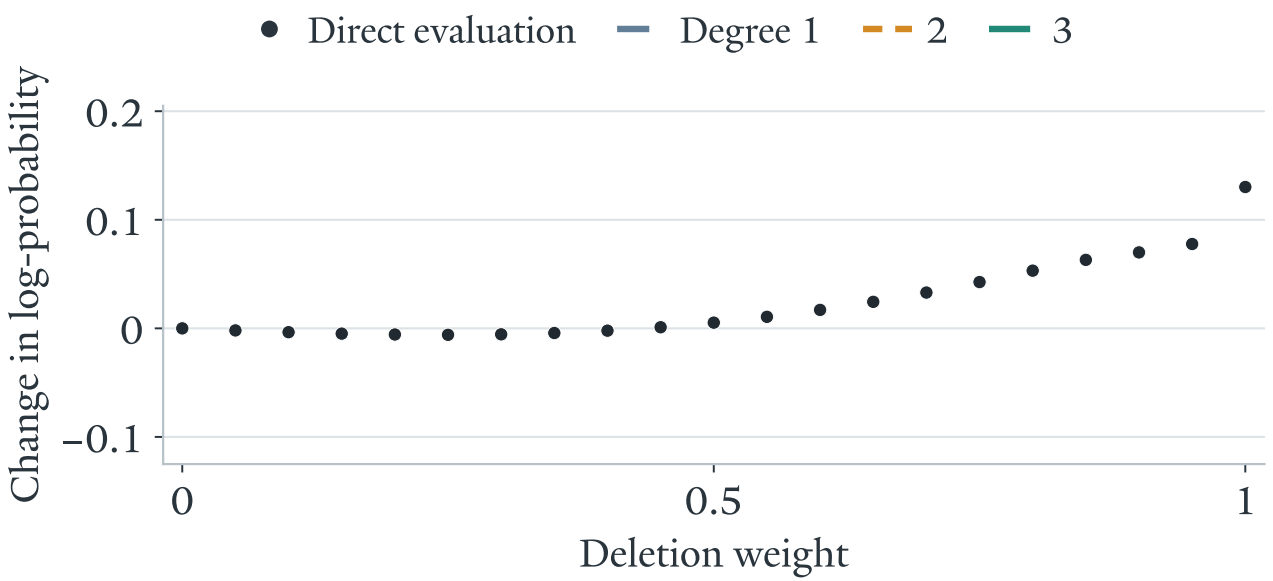
Taylor Expansion of the Measurement

Around the full context $\mathbf{0}$, for $\boldsymbol{w} \in [0,1]^n$,

$$\phi(\boldsymbol{w}) = \phi(\mathbf{0}) + \nabla\phi(\mathbf{0})^\top \boldsymbol{w} + \tfrac{1}{2}\boldsymbol{w}^\top \nabla^2\phi(\mathbf{0})\boldsymbol{w} + \cdots .$$

Let $H := \nabla^2\phi(\mathbf{0})$. For deletion i,

$$\phi(\boldsymbol{e}_i) = \phi(\mathbf{0}) + \partial_i\phi(\mathbf{0}) + \tfrac{1}{2}H_{ii} + \cdots .$$



Chunk 6
"giant turnip …"
$\phi(z\boldsymbol{e}_6) - \phi(\mathbf{0})$

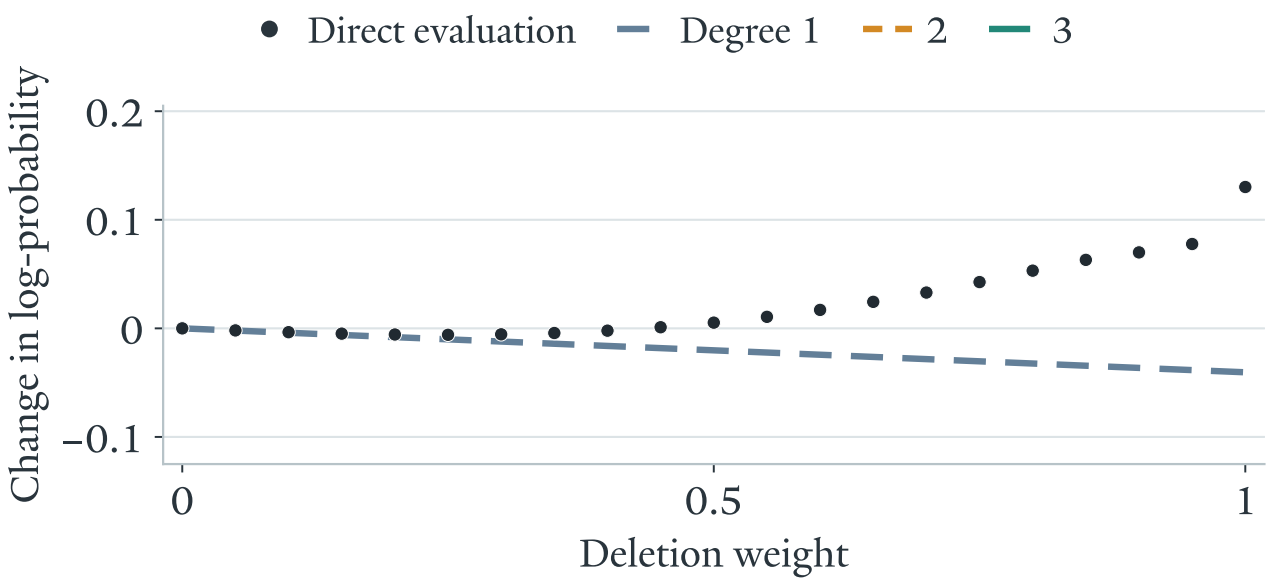
Taylor Expansion of the Measurement

Around the full context $\mathbf{0}$, for $\boldsymbol{w} \in [0,1]^n$,

$$\phi(\boldsymbol{w}) = \phi(\mathbf{0}) + \nabla\phi(\mathbf{0})^\top \boldsymbol{w} + \tfrac{1}{2}\boldsymbol{w}^\top \nabla^2\phi(\mathbf{0})\boldsymbol{w} + \cdots .$$

Let $H := \nabla^2\phi(\mathbf{0})$. For deletion i,

$$\phi(\boldsymbol{e}_i) = \phi(\mathbf{0}) + \partial_i\phi(\mathbf{0}) + \tfrac{1}{2}H_{ii} + \cdots .$$



Chunk 6
"giant turnip …"
$\phi(z\boldsymbol{e}_6) - \phi(\mathbf{0})$

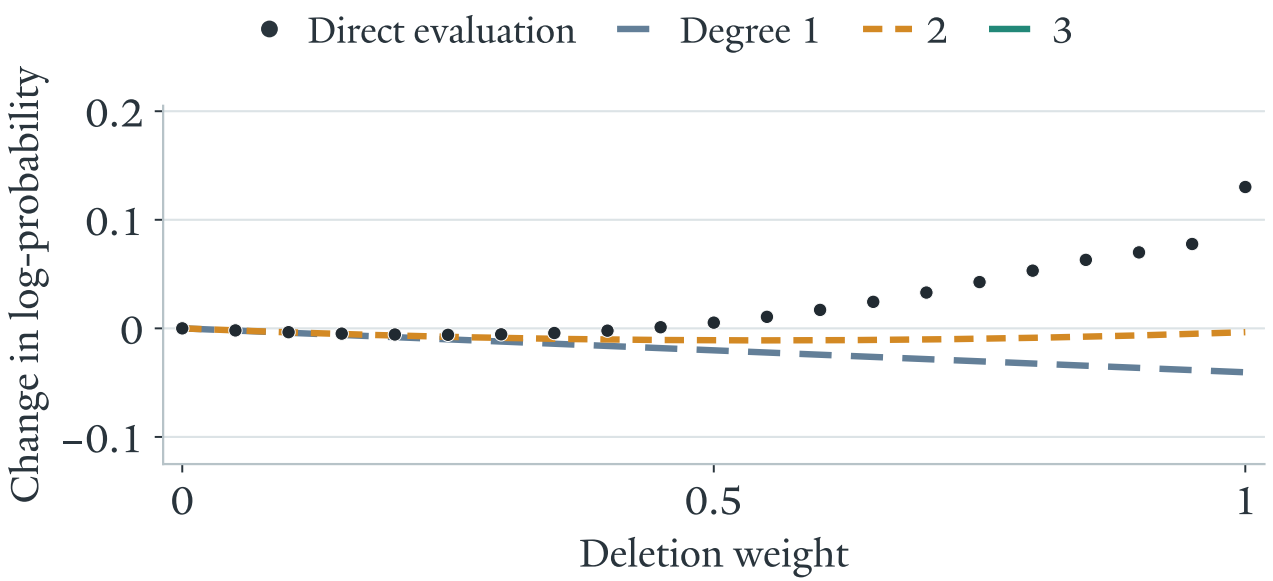
Taylor Expansion of the Measurement

Around the full context $\mathbf{0}$, for $\boldsymbol{w} \in [0,1]^n$,

$$\phi(\boldsymbol{w}) = \phi(\mathbf{0}) + \nabla\phi(\mathbf{0})^\top \boldsymbol{w} + \tfrac{1}{2}\boldsymbol{w}^\top \nabla^2\phi(\mathbf{0})\boldsymbol{w} + \cdots .$$

Let $H := \nabla^2\phi(\mathbf{0})$. For deletion i,

$$\phi(\boldsymbol{e}_i) = \phi(\mathbf{0}) + \partial_i\phi(\mathbf{0}) + \tfrac{1}{2}H_{ii} + \cdots .$$



Chunk 6
"giant turnip …"
$\phi(z\boldsymbol{e}_6) - \phi(\mathbf{0})$

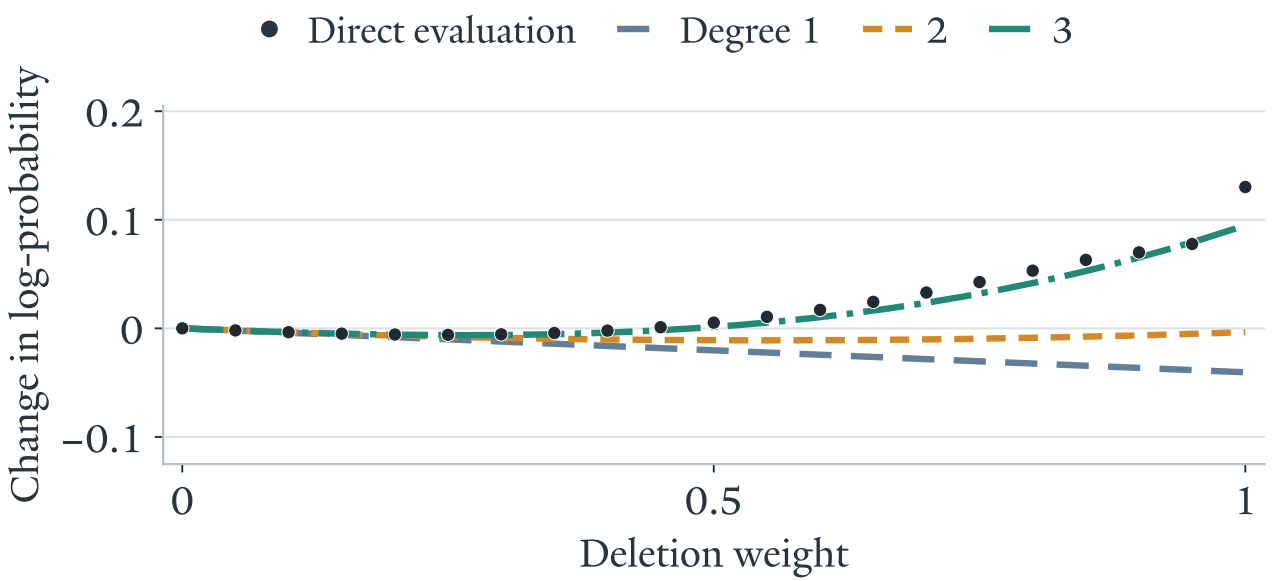
Taylor Expansion of the Measurement

Around the full context $\mathbf{0}$, for $\boldsymbol{w} \in [0,1]^n$,

$$\phi(\boldsymbol{w}) = \phi(\mathbf{0}) + \nabla\phi(\mathbf{0})^\top \boldsymbol{w} + \tfrac{1}{2}\boldsymbol{w}^\top \nabla^2\phi(\mathbf{0})\boldsymbol{w} + \cdots .$$

Let $H := \nabla^2\phi(\mathbf{0})$. For deletion i,

$$\phi(\boldsymbol{e}_i) = \phi(\mathbf{0}) + \partial_i\phi(\mathbf{0}) + \tfrac{1}{2}H_{ii} + \cdots .$$



Chunk 6
"giant turnip …"
$\phi(z\boldsymbol{e}_6) - \phi(\mathbf{0})$

Strategy: Estimate Taylor Expansion

High-level strategy follows [Gun26]

1. Predict each LOO value from its Taylor terms up to some degree d.

Strategy: Estimate Taylor Expansion

High-level strategy follows [Gun26]

1. Predict each LOO value from its Taylor terms up to some degree d.
2. Random projections to avoid computing n^d degree-d coefficients.
   - Estimate Taylor expansion from *sketch* of coefficients.

Strategy: Estimate Taylor Expansion

High-level strategy follows [Gun26]

1. Predict each LOO value from its Taylor terms up to some degree d.
2. Random projections to avoid computing n^d degree-d coefficients.
   - Estimate Taylor expansion from *sketch* of coefficients.
3. "Stability" controls truncation error and sketch variance.

Main Structural Property: Stability

DEFINITION 2 (STABILITY, INFORMAL)

The Taylor coefficient tensors decay in aggregate with degree:

$$\underbrace{\|\nabla\phi(\mathbf{0})\|_2}_{\text{degree } 1} \lesssim 4^{-1}, \quad \underbrace{\tfrac{1}{2}\|\nabla^2\phi(\mathbf{0})\|_F}_{\text{degree } 2} \lesssim 4^{-2}, \quad \ldots, \quad \underbrace{\tfrac{1}{r!}\|\nabla^r\phi(\mathbf{0})\|_F}_{\text{degree } r} \lesssim 4^{-r}.$$

Main Structural Property: Stability

DEFINITION 2 (STABILITY, INFORMAL)

The Taylor coefficient tensors decay in aggregate with degree:

$$\underbrace{\|\nabla\phi(\mathbf{0})\|_2}_{\text{degree } 1} \lesssim 4^{-1}, \quad \underbrace{\tfrac{1}{2}\|\nabla^2\phi(\mathbf{0})\|_F}_{\text{degree } 2} \lesssim 4^{-2}, \quad \ldots, \quad \underbrace{\tfrac{1}{r!}\|\nabla^r\phi(\mathbf{0})\|_F}_{\text{degree } r} \lesssim 4^{-r}.$$

1. Truncating at degree $d = \log_4(1/\varepsilon)$ gives $O(\varepsilon)$ error.

Main Structural Property: Stability

DEFINITION 2 (STABILITY, INFORMAL)

The Taylor coefficient tensors decay in aggregate with degree:

$$\underbrace{\|\nabla\phi(\mathbf{0})\|_2}_{\text{degree } 1} \lesssim 4^{-1}, \quad \underbrace{\tfrac{1}{2}\|\nabla^2\phi(\mathbf{0})\|_F}_{\text{degree } 2} \lesssim 4^{-2}, \quad \ldots, \quad \underbrace{\tfrac{1}{r!}\|\nabla^r\phi(\mathbf{0})\|_F}_{\text{degree } r} \lesssim 4^{-r}.$$

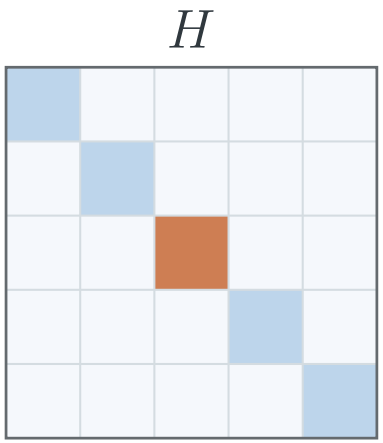
1. Truncating at degree $d = \log_4(1/\varepsilon)$ gives $O(\varepsilon)$ error.
2. Stability also bounds the variance of the sketches.

Warm-Up: The Hessian

- Estimate each Taylor degree, then combine the estimates.

Warm-Up: The Hessian

- Estimate each Taylor degree, then combine the estimates.
- At degree two, estimate the diagonal of $H = \nabla^2 \phi(\mathbf{0})$.

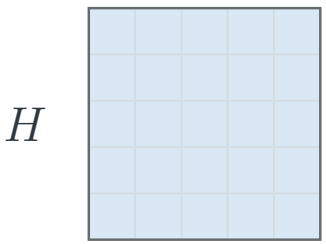


Key Innovation: Sketching Slices

Let σ_i^2 be the variance of one sketch's estimate of H_{ii}.

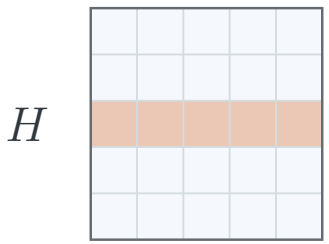
- **Forward** [Gun26]: sketch H. The variance bound uses $\|H\|_F^2$.

Forward



$$\sigma_i^2 \lesssim \|H\|_F^2$$

Backward–forward



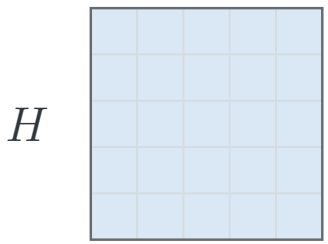
$$\sigma_i^2 \lesssim \|H_{i,:}\|_2^2$$

Key Innovation: Sketching Slices

Let σ_i^2 be the variance of one sketch's estimate of H_{ii}.

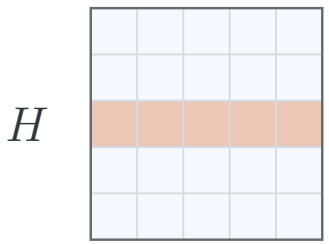
- **Forward** [Gun26]: sketch H. The variance bound uses $\|H\|_F^2$.
- **Backward–forward** (ours): sketch each row individually. For query i, the bound depends only on row i.

Forward



$$\max_i \sigma_i^2 \lesssim \|H\|_F^2$$

Backward–forward



$$\sum_i \sigma_i^2 \lesssim \|H\|_F^2$$

$$\sum_i \|H_{i,:}\|_2^2 = \|H\|_F^2$$

Why $1/\varepsilon$ Sketches Suffice

- For query i, $O(\sigma_i^2/\varepsilon^2)$ sketches suffice for error ε

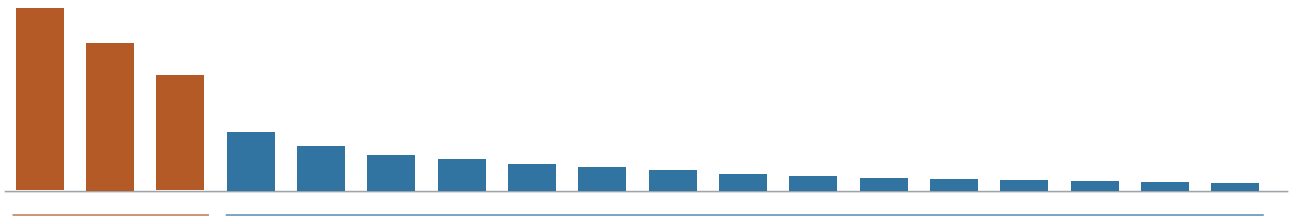
Why $1/\varepsilon$ Sketches Suffice

- For query i, $O(\sigma_i^2/\varepsilon^2)$ sketches suffice for error ε
- Stability gives $\sum_i \sigma_i^2 = O(\|H\|_F^2) = O(1)$

Why $1/\varepsilon$ Sketches Suffice

- For query i, $O(\sigma_i^2/\varepsilon^2)$ sketches suffice for error ε
- Stability gives $\sum_i \sigma_i^2 = O(\|H\|_F^2) = O(1)$
- By Markov's inequality,

$$\left|\{i : \sigma_i^2 > \varepsilon\}\right| \leq \frac{1}{\varepsilon} \sum_i \sigma_i^2 = O(1/\varepsilon).$$

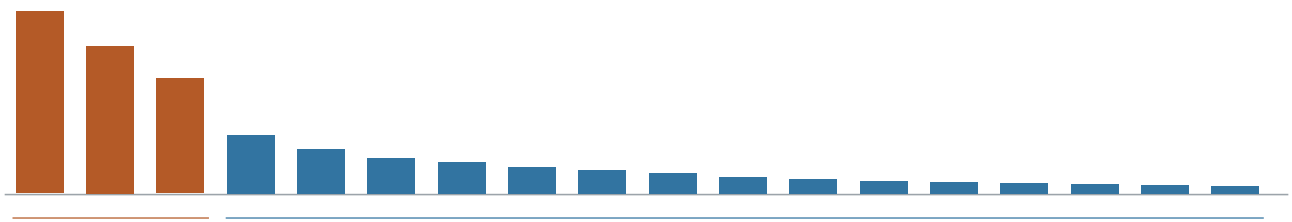


Why $1/\varepsilon$ Sketches Suffice

- For query i, $O(\sigma_i^2/\varepsilon^2)$ sketches suffice for error ε
- Stability gives $\sum_i \sigma_i^2 = O(\|H\|_F^2) = O(1)$
- By Markov's inequality,

$$\left|\{i : \sigma_i^2 > \varepsilon\}\right| \leq \frac{1}{\varepsilon} \sum_i \sigma_i^2 = O(1/\varepsilon).$$

- Rank by estimated variance. Evaluate the top $O(1/\varepsilon)$ exactly.



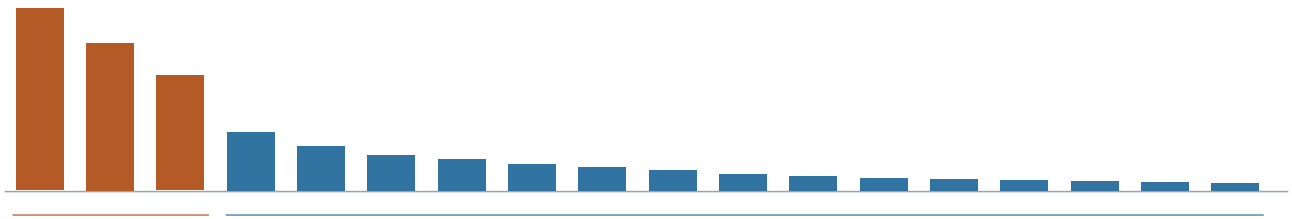
Union bound over n queries: $\widetilde{O}(\log(n)/\varepsilon)$ calls.

Why $1/\varepsilon$ Sketches Suffice

- For query i, $O(\sigma_i^2/\varepsilon^2)$ sketches suffice for error ε
- Stability gives $\sum_i \sigma_i^2 = O(\|H\|_F^2) = O(1)$
- By Markov's inequality,

$$\left|\{i : \sigma_i^2 > \varepsilon\}\right| \leq \frac{1}{\varepsilon} \sum_i \sigma_i^2 = O(1/\varepsilon).$$

- Rank by estimated variance. Evaluate the top $O(1/\varepsilon)$ exactly.
- For the rest, $\sigma_i^2 = O(\varepsilon)$, so $O(1/\varepsilon)$ shared sketches suffice.



Union bound over n queries: $\widetilde{O}(\log(n)/\varepsilon)$ calls.

Table of Contents

Motivation

Problem Definition

Prior Works

Contributions

Proof Overview

- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

Gunn's Sketch

Target: $H_{ii} = \langle H, \boldsymbol{e}_i \boldsymbol{e}_i^\top \rangle$.

Gunn's Sketch

$$\text{Target:} \qquad H_{ii} = \langle H, \boldsymbol{e}_i \boldsymbol{e}_i^\top \rangle.$$

Draw a standard complex Gaussian direction $\boldsymbol{g}$.

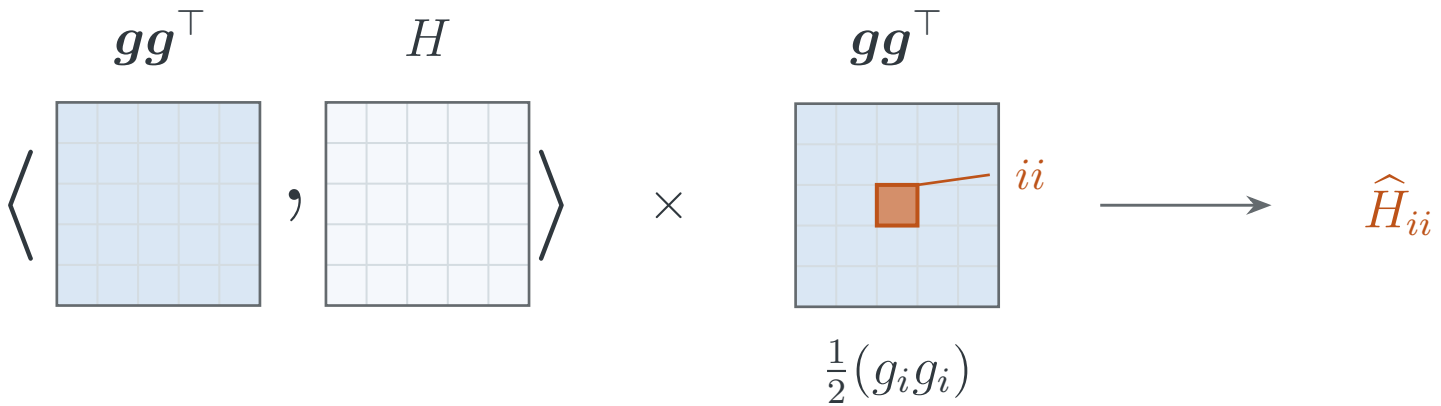
$$\mathbb{E}_{\boldsymbol{g}} \left[\langle \boldsymbol{g}\boldsymbol{g}^\top, H \rangle \cdot \langle \tfrac{1}{2} \boldsymbol{e}_i \boldsymbol{e}_i^\top, \boldsymbol{g}\boldsymbol{g}^\top \rangle \right] = \mathbb{E}_{\boldsymbol{g}} \left[\langle \boldsymbol{g}\boldsymbol{g}^\top, H \rangle \, \frac{g_i g_i}{2} \right] = H_{ii}.$$

Gunn's Sketch

Target: $H_{ii} = \langle H, \boldsymbol{e}_i \boldsymbol{e}_i^\top \rangle$.

Draw a standard complex Gaussian direction $\boldsymbol{g}$.

$$\mathbb{E}_{\boldsymbol{g}}\left[\langle \boldsymbol{g}\boldsymbol{g}^\top, H\rangle \cdot \langle \tfrac{1}{2}\boldsymbol{e}_i \boldsymbol{e}_i^\top, \boldsymbol{g}\boldsymbol{g}^\top\rangle\right] = \mathbb{E}_{\boldsymbol{g}}\left[\langle \boldsymbol{g}\boldsymbol{g}^\top, H\rangle \frac{g_i g_i}{2}\right] = H_{ii}.$$



- Reuse sketch across LOO queries.

Directional Derivatives via Forward Differentiation

Consider $\phi(z\boldsymbol{g})$ as a function of the scalar variable z:

$$\phi(z\boldsymbol{g}) = \phi(\mathbf{0}) + z\,\nabla\phi(\mathbf{0})^\top \boldsymbol{g} + \frac{z^2}{2}\boldsymbol{g}^\top H\boldsymbol{g} + \cdots .$$

Directional Derivatives via Forward Differentiation

Consider $\phi(z\boldsymbol{g})$ as a function of the scalar variable z:

$$\phi(z\boldsymbol{g}) = \phi(\mathbf{0}) + z\,\nabla\phi(\mathbf{0})^\top \boldsymbol{g} + \frac{z^2}{2}\boldsymbol{g}^\top H\boldsymbol{g} + \cdots .$$

- Forward differentiation: evaluate the circuit in $\mathbb{C}[z]/(z^{d+1})$.

Directional Derivatives via Forward Differentiation

Consider $\phi(z\boldsymbol{g})$ as a function of the scalar variable z:

$$\phi(z\boldsymbol{g}) = \phi(\mathbf{0}) + z\,\nabla\phi(\mathbf{0})^\top \boldsymbol{g} + \frac{z^2}{2}\boldsymbol{g}^\top H\boldsymbol{g} + \cdots .$$

- Forward differentiation: evaluate the circuit in $\mathbb{C}[z]/(z^{d+1})$.
- One pass returns the vector of Taylor coefficients:

$$\Big(\phi(\mathbf{0}),\ \nabla\phi(\mathbf{0})^\top \boldsymbol{g},\ \tfrac{1}{2}\boldsymbol{g}^\top H\boldsymbol{g},\ \ldots,\ [z^d]\phi(z\boldsymbol{g})\Big).$$

Directional Derivatives via Forward Differentiation

Consider $\phi(z\boldsymbol{g})$ as a function of the scalar variable z:

$$\phi(z\boldsymbol{g}) = \phi(\mathbf{0}) + z\,\nabla\phi(\mathbf{0})^\top \boldsymbol{g} + \frac{z^2}{2}\boldsymbol{g}^\top H\boldsymbol{g} + \cdots .$$

- Forward differentiation: evaluate the circuit in $\mathbb{C}[z]/(z^{d+1})$.
- One pass returns the vector of Taylor coefficients:

$$\Big(\phi(\mathbf{0}),\ \nabla\phi(\mathbf{0})^\top \boldsymbol{g},\ \tfrac{1}{2}\boldsymbol{g}^\top H\boldsymbol{g},\ \ldots,\ [z^d]\phi(z\boldsymbol{g})\Big).$$

Read the Hessian projection from the z^2 entry.

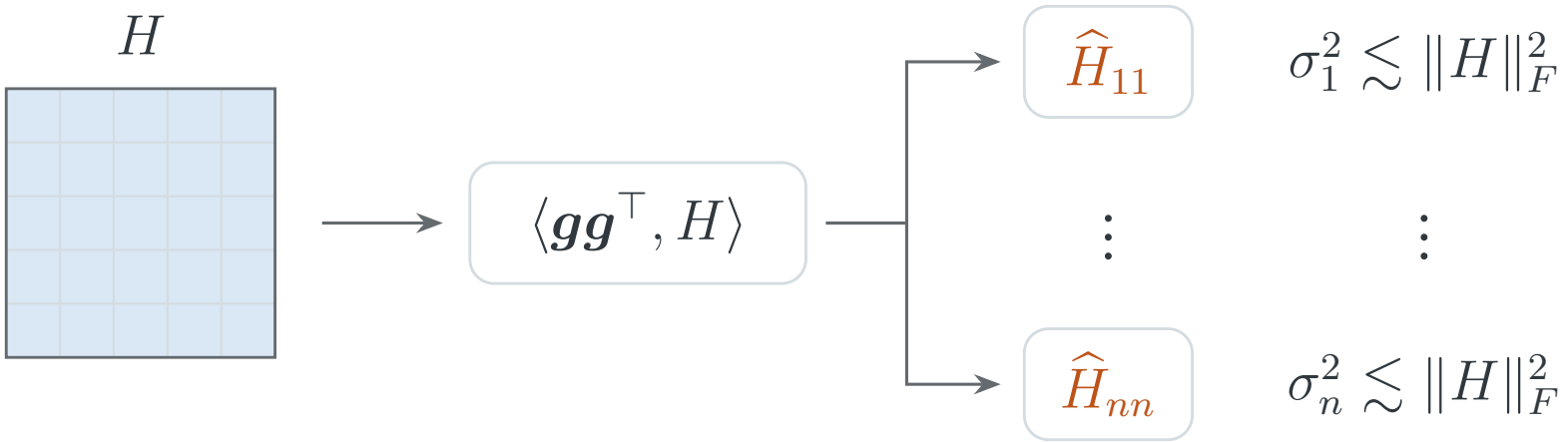
Forward Sketch

- One pass computes all degrees through $d = O(\log(1/\varepsilon))$.
- Cost per pass: $O(\mathrm{size}(\phi)\, d \log d)$ operations.

Forward differentiation [Wen64]. Forward sketch adapted from [Gun26].

Forward Sketch

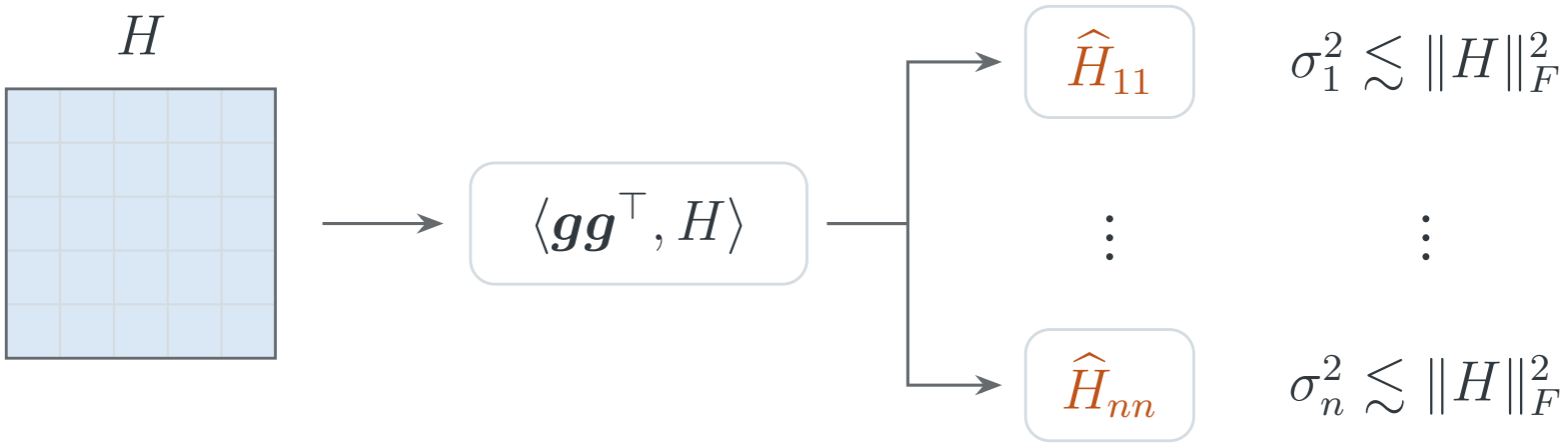
- One pass computes all degrees through $d = O(\log(1/\varepsilon))$.
- Cost per pass: $O(\mathrm{size}(\phi)\, d \log d)$ operations.
- Variance bound $\sigma_i^2 \lesssim \|H\|_F^2$.
- Each query's variance bound uses the whole Hessian.



Forward differentiation [Wen64]. Forward sketch adapted from [Gun26].

Forward Sketch

- One pass computes all degrees through $d = O(\log(1/\varepsilon))$.
- Cost per pass: $O(\mathrm{size}(\phi)\, d \log d)$ operations.
- Variance bound $\sigma_i^2 \lesssim \|H\|_F^2$.
- Each query's variance bound uses the whole Hessian.



Mean estimation and a union bound: $\widetilde{O}(\log(n)/\varepsilon^2)$ calls.

Forward differentiation [Wen64]. Forward sketch adapted from [Gun26].

Table of Contents

Motivation

Problem Definition

Prior Works

Contributions

Proof Overview

- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

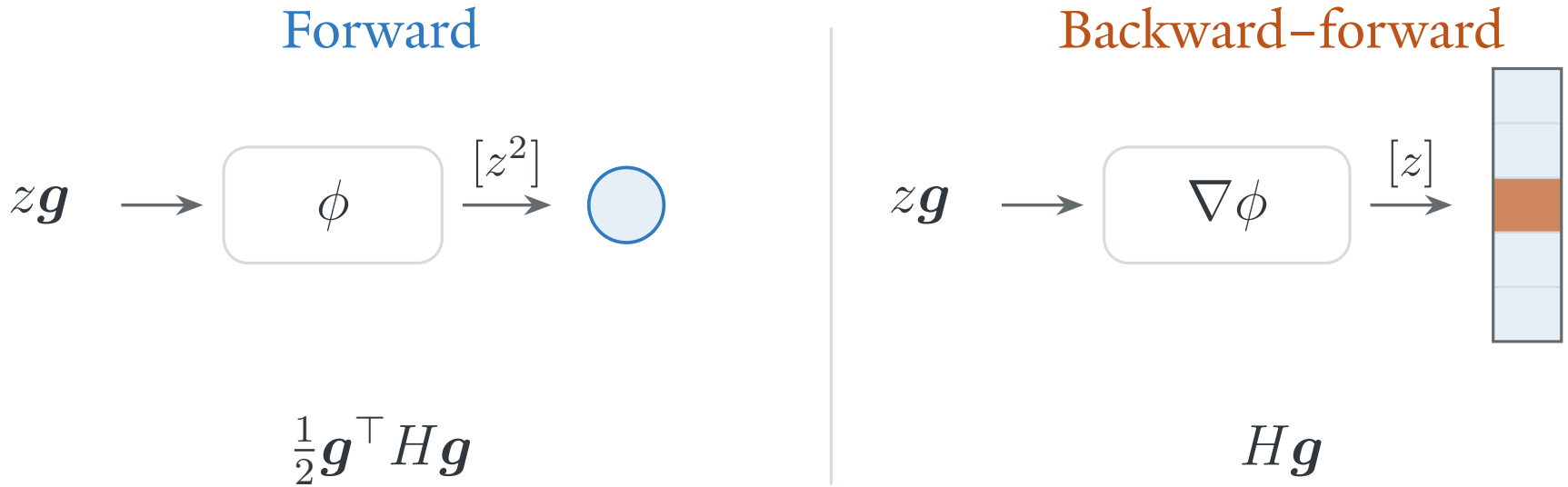
Key Idea: More Informative Sketches for Free

- Sketch gradient $\nabla\phi$ instead of just ϕ

Constant-factor overhead relative to a forward sketch. [BS83; RHW86]

Key Idea: More Informative Sketches for Free

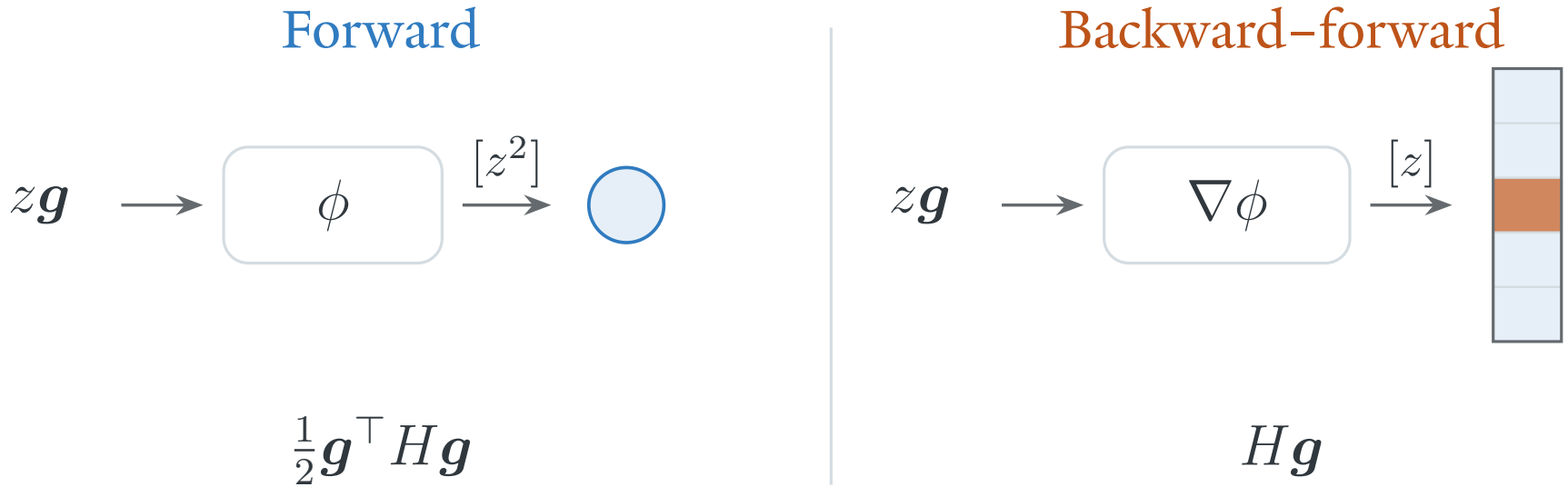
- Sketch gradient $\nabla\phi$ instead of just ϕ
- Apply forward differentiation to the backpropagation circuit!



Constant-factor overhead relative to a forward sketch. [BS83; RHW86]

Key Idea: More Informative Sketches for Free

- Sketch gradient $\nabla\phi$ instead of just ϕ
- Apply forward differentiation to the backpropagation circuit!



One scalar $\longrightarrow$ Entire vector.

Constant-factor overhead relative to a forward sketch. [BS83; RHW86]

Sketching Slices of the Hessian

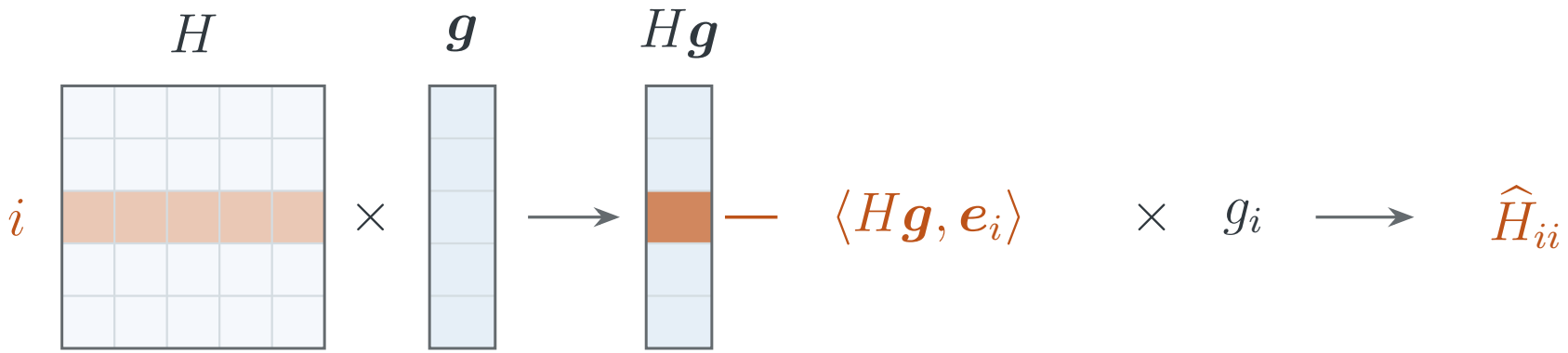
One backward–forward sketch gives the vector $H\boldsymbol{g}$.

$$\mathbb{E}_{\boldsymbol{g}}[\langle H\boldsymbol{g}, \boldsymbol{e}_i\rangle \cdot \langle \boldsymbol{e}_i, \boldsymbol{g}\rangle] = \mathbb{E}_{\boldsymbol{g}}[\langle H\boldsymbol{g}, \boldsymbol{e}_i\rangle\, g_i] = H_{ii}.$$

Sketching Slices of the Hessian

One backward–forward sketch gives the vector $H\boldsymbol{g}$.

$$\mathbb{E}_{\boldsymbol{g}}[\langle H\boldsymbol{g}, \boldsymbol{e}_i\rangle \cdot \langle \boldsymbol{e}_i, \boldsymbol{g}\rangle] = \mathbb{E}_{\boldsymbol{g}}[\langle H\boldsymbol{g}, \boldsymbol{e}_i\rangle\, g_i] = H_{ii}.$$



$(H\boldsymbol{g})_i = H_{i,:}\boldsymbol{g}$, so only row i enters this estimate.
Hence $\sigma_i^2 \lesssim \|H_{i,:}\|_F^2$ and $\sum_i \sigma_i^2 \lesssim \|H\|_F^2$.

Table of Contents

Motivation

Problem Definition

Prior Works

Contributions

Proof Overview
- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

Query Complexity Lower Bound

- Oracle Model: Access ϕ up to exact first-order $(\phi(\boldsymbol{w}), \nabla\phi(\boldsymbol{w}))$ and higher order directional derivatives $D^r_{\boldsymbol{g}}\phi(\boldsymbol{w}), r \geq 3$.

Query Complexity Lower Bound

- Oracle Model: Access ϕ up to exact first-order $(\phi(\boldsymbol{w}), \nabla\phi(\boldsymbol{w}))$ and higher order directional derivatives $D^r_{\boldsymbol{g}}\phi(\boldsymbol{w}), r \geq 3$.
- Consider $\phi(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w}$ for random matrix A.

Query Complexity Lower Bound

- Oracle Model: Access ϕ up to exact first-order $(\phi(\boldsymbol{w}), \nabla\phi(\boldsymbol{w}))$ and higher order directional derivatives $D^r_{\boldsymbol{g}}\phi(\boldsymbol{w}), r \geq 3$.
- Consider $\phi(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w}$ for random matrix A.
- One derivative-oracle query reduces to two matrix–vector products: $A\boldsymbol{w}$ and $A\boldsymbol{g}$.

Query Complexity Lower Bound

- Oracle Model: Access ϕ up to exact first-order $(\phi(\boldsymbol{w}), \nabla\phi(\boldsymbol{w}))$ and higher order directional derivatives $D^r_{\boldsymbol{g}}\phi(\boldsymbol{w}), r \geq 3$.
- Consider $\phi(\boldsymbol{w}) = \frac{1}{2}\boldsymbol{w}^\top A\boldsymbol{w}$ for random matrix A.
- One derivative-oracle query reduces to two matrix–vector products: $A\boldsymbol{w}$ and $A\boldsymbol{g}$.
- A diagonal-estimation lower bound [ACDH+24] requires $\Omega(\min\{n, 1/\varepsilon\})$ matrix–vector queries.

Table of Contents

Motivation

Problem Definition

Prior Works

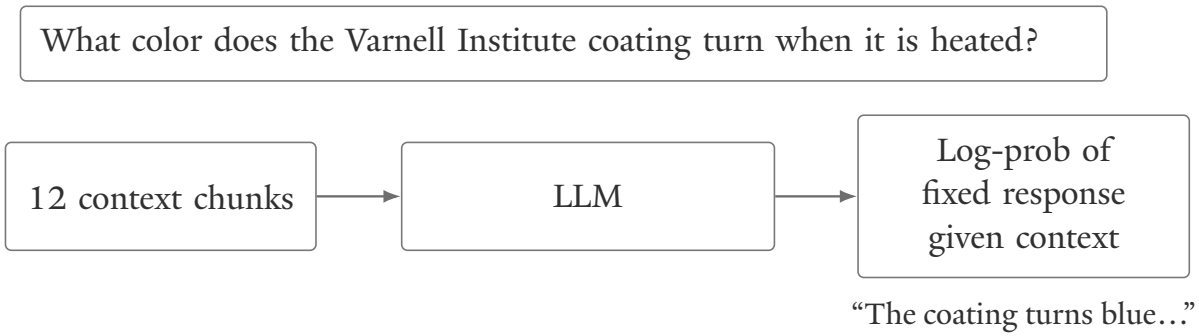
Contributions

Proof Overview
- High-Level Strategy
- Forward Sketches
- Backward–Forward Sketches
- Lower Bound

Experiments

Experimental Setup

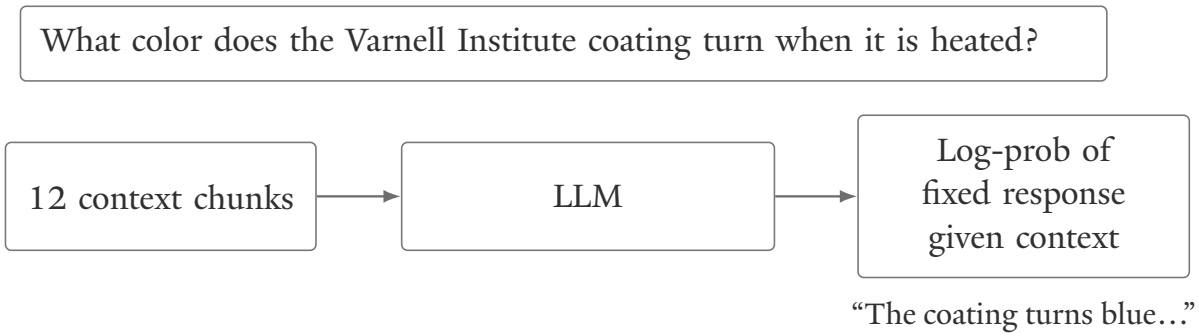
- `Llama-3.2-1B-Instruct`, 12 chunks, one fixed response.



One H100 GPU per trial. 64-bit arithmetic. Forward differentiation with JAX `jet`.

Experimental Setup

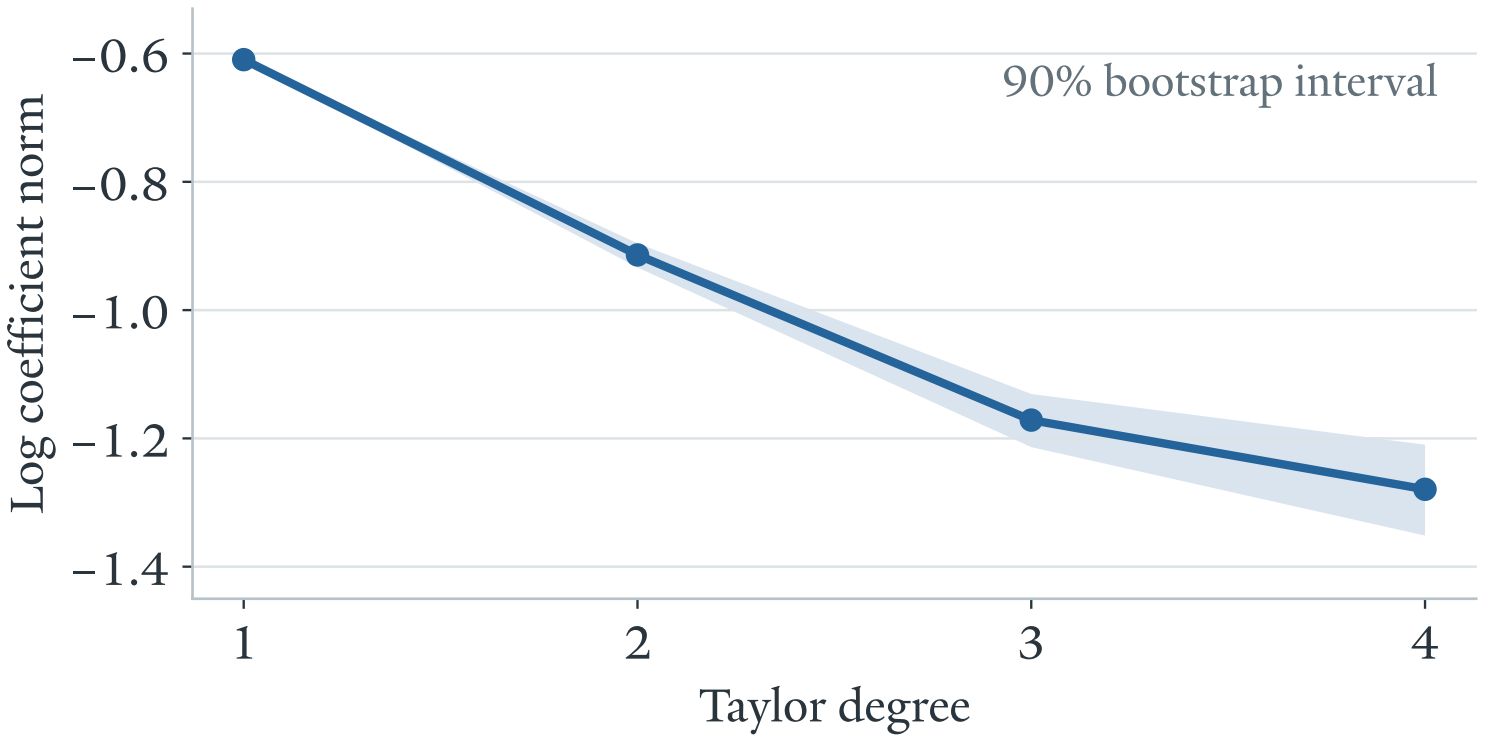
- `Llama-3.2-1B-Instruct`, 12 chunks, one fixed response.
- 8 sketches per trial, 128 independent trials.



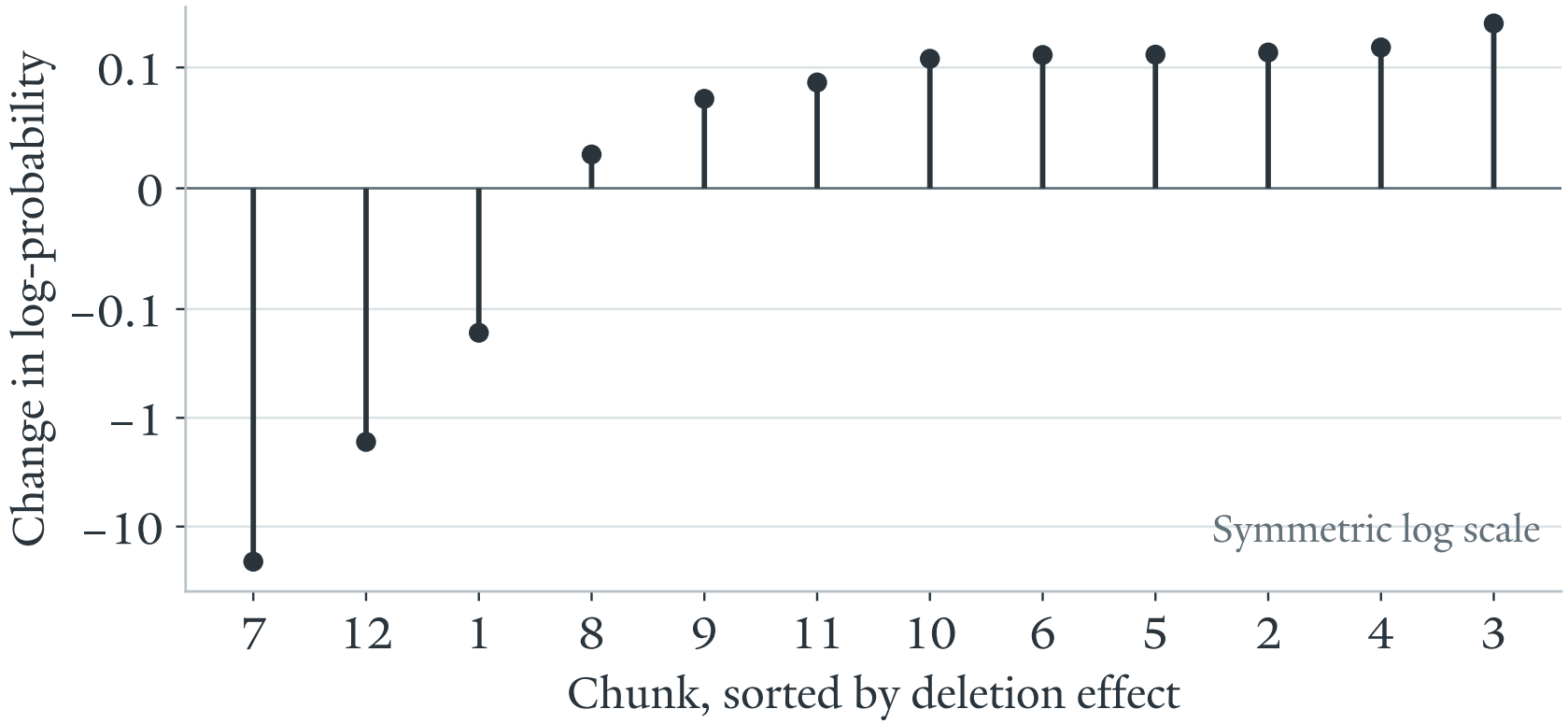
One H100 GPU per trial. 64-bit arithmetic. Forward differentiation with JAX `jet`.

Do the Coefficients Exhibit Stability?

Norms of Taylor coefficients decrease with degree.

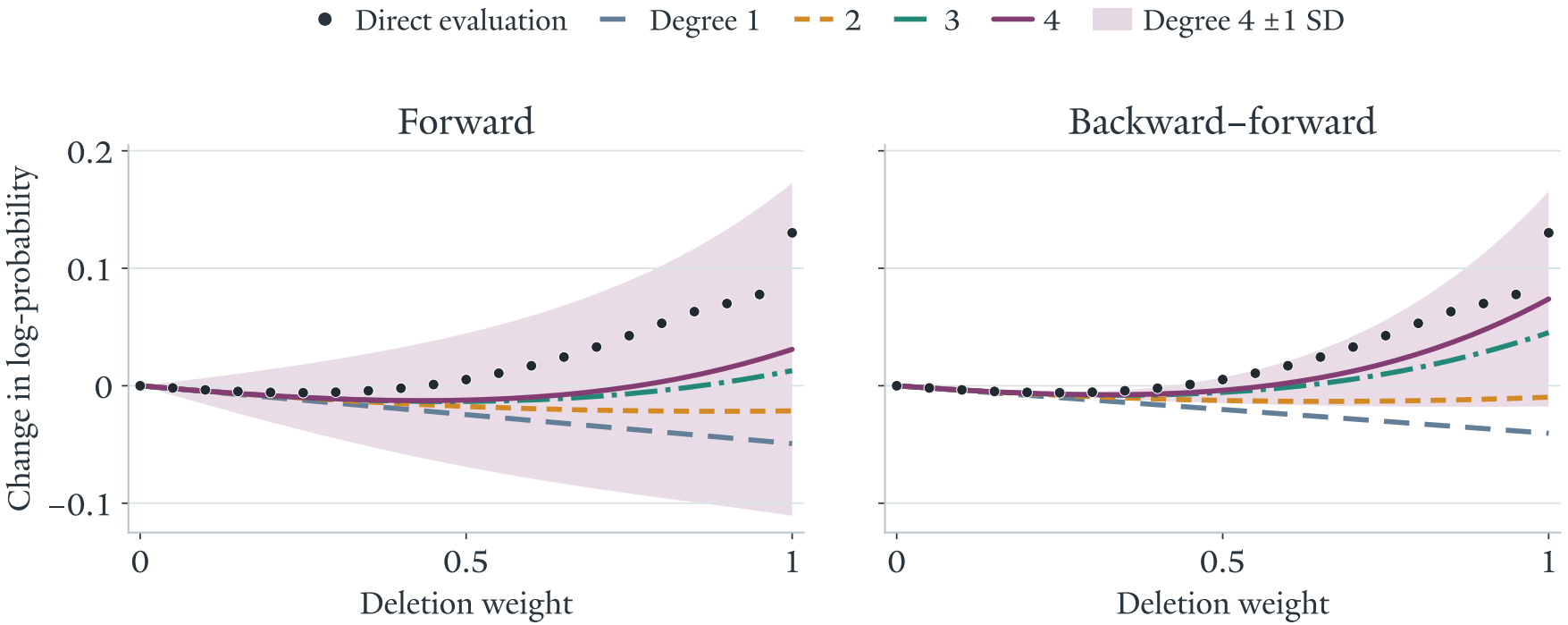


Leave-One-Out Values



Variance With Eight Sketches

Chunk 6: “giant turnip …”



Conclusion

- Sketching $\nabla\phi$ reduces variance vs sketching ϕ.

Conclusion

- Sketching $\nabla\phi$ reduces variance vs sketching ϕ.
- Under stability, exactly evaluating a few high-variance queries reduces the number of sketches needed.

Conclusion

- Sketching $\nabla\phi$ reduces variance vs sketching ϕ.
- Under stability, exactly evaluating a few high-variance queries reduces the number of sketches needed.
- $\widetilde{O}(\log(n)/\varepsilon)$ model calls for leave-one-out context attribution.