Jacobian Scopes: token-level causal attributions in LLMs

Anonymous ACL submission

Abstract

Large language models (LLMs) make next-token predictions based on clues present in their context, such as semantic descriptions and in-context examples. Yet, elucidating which prior tokens most strongly influence a given prediction remains challenging due to the proliferation of layers and attention heads in modern architectures. We propose *Jacobian Scopes*, a suite of gradient-based, token-level causal attribution methods for interpreting LLM predictions. Grounded in perturbation theory and information geometry, Jacobian Scopes quantify how input tokens influence various aspects of a model's prediction, such as specific logits, the full predictive distribution, and model uncertainty (effective temperature). Through case studies spanning instruction understanding, translation, and in-context learning (ICL), we demonstrate how Jacobian Scopes reveal implicit political biases, uncover word- and phrase-level translation strategies, and shed light on recently debated mechanisms underlying in-context time-series forecasting. To facilitate exploration of Jacobian Scopes on custom text, we will open-source our implementations and provide a cloud-hosted interactive demo if the manuscript is accepted.

1 Introduction

Large Language Models (LLMs) exhibit unexpected and remarkable emergent abilities, such as multi-step (Huang and Chang, 2023) and in-context learning (ICL) (Dong et al., 2024; Garg et al., 2023). However, the mechanistic underpinning of these emergent abilities remains a deep and open question (Olsson et al., 2022; Reddy, 2024; Von Oswald et al., 2023).

This has fueled a growing endeavor, across academia and industry, to interpret and explain LLMs' inference process during such emergent thinking modes; as when writing (Wan et al., 2023; Guo et al., 2024) and reasoning (Wei et al., 2022; Wang et al., 2023).

In this work, we present a suite of gradient-based causal attribution methods called *Jacobian Scopes*. Jacobian Scopes explains LLM predictions in terms of influential input tokens. This is achieved by analyzing the Jacobian matrix, which captures locally linearized relation between the predictive distribution and input embeddings.

Jacobian Scopes come in three variants: Semantic, Fisher, and Temperature Scopes, each corresponding to a different *explanandum* (attribution objective): logit of a certain token, the full predictive distribution, and the model's uncertainty (effective temperature). A survey of related gradient-based attribution methods and their relation to Jacobian Scopes is provided in Appendix A.2.

We showcase the efficacy of our proposed Jacobian Scopes through a range of case studies, including natural language instruction such as system prompts and translation, as well as ICL tasks such as time-series forecasting (Gruver et al., 2023; Mirchandani et al., 2023). We further evaluate attribution faithfulness quantitatively on two benchmark datasets (LAMBADA and IWSLT2017) across two suites of leading LLMs: LLaMA-3.2 (Grattafiori et al., 2024) and Qwen2.5 (Qwen Team, 2025).

2 Methodology

Notations. We use bold lowercase letters (e.g. $\boldsymbol{x}$) for vectors and bold uppercase letters (e.g. $\boldsymbol{W}$) for matrices.

Let $\mathcal{H} := \mathbb{R}^{d_{\text{model}}}$ denote the model's hidden space. We represent a length-T input sequence as $\boldsymbol{X}_{1:T} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T)$, where $\boldsymbol{x}_t \in \mathcal{H}$. We view an auto-regressive LLM as a function that maps a sequence of inputs to the final-layer, post-layer-norm hidden state at the leading position:

$$\boldsymbol{y} := f(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_t, \ldots, \boldsymbol{x}_T) \in \mathcal{H}.$$

The corresponding logits $\boldsymbol{z}$ and predictive distribution $\boldsymbol{p}$ are given by $\boldsymbol{z} := \boldsymbol{W}\boldsymbol{y}$ and $\boldsymbol{p}(\cdot \mid \boldsymbol{X}_{1:T}) :=$

$\mathrm{softmax}(\boldsymbol{z}) \in \mathbb{R}^{|\mathcal{V}|}$, where $\boldsymbol{W} \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{model}}}$ denotes the unembedding matrix, and $\mathcal{V}$ is the vocabulary set of size $|\mathcal{V}|$.

Definition of influence score. We are interested in quantifying how perturbations to an input token embedding $\boldsymbol{x}_t$ affect the leading hidden state $\boldsymbol{y}$. We define the input-to-output Jacobian at position t as

$$\boldsymbol{J}_t := \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}_t} \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}.$$

Forming the full Jacobian $\boldsymbol{J}_t$ requires d_{model} backward passes, which can be expensive for modern LLMs. Fortunately, in practice we find it sufficient to compute the vector-Jacobian product, $\boldsymbol{v}^\intercal \boldsymbol{J}_t$, where $\boldsymbol{v} \in \mathcal{H}$ captures a property of interest in the predictive distribution, such as confidence (effective inverse temperature) or the embedding of a target token.

Geometrically, $\|\boldsymbol{v}^\intercal \boldsymbol{J}_t\|_2$ measures the largest displacement along the direction $\boldsymbol{v}$, that could be induced by an ε-norm perturbation to $\boldsymbol{x}_t$ (proven in Appendix A.8). We thus define the *influence* of token $\boldsymbol{x}_t$ on the leading output $\boldsymbol{y}$ as

$$\mathrm{Influence}_t := \|\boldsymbol{v}^\intercal \boldsymbol{J}_t\|_2 \,. \tag{1}$$

Influence computation via auto-diff. For the leading prediction $\boldsymbol{y}$, the influence scores at all input positions $t \leq T$ can be computed with a single backward pass by auto-differentiating the customized scalar loss $\mathcal{L}(\boldsymbol{X}_{1:T}) := \boldsymbol{v}^\intercal \boldsymbol{y}$:

$$\left\| \frac{\partial \mathcal{L}(\boldsymbol{X}_{1:T})}{\partial \boldsymbol{x}_t} \right\|_2 = \left\| \boldsymbol{v}^\intercal \frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}_t} \right\|_2 = \mathrm{Influence}_t. \tag{2}$$

In Section 3, we present three specific choices for the direction of interest $\boldsymbol{v}$, leading to *Semantic Scope*, *Fisher Scope* and *Temperature Scope*.

3 Example applications

This section discusses three formulations of Jacobian Scopes, each motivated by a specific attribution objective. All experiments are conducted on LLaMA-3.2 1B, unless otherwise specified.

3.1 Semantic Scope

Semantic Scope traces the input tokens that most prominently contribute to the predicted probability of a target vocabulary. Here, the direction of interest is $\boldsymbol{v} = \boldsymbol{w}_{\text{target}}$, which is the row of the unembedding matrix $\boldsymbol{W}$ corresponding to the target token. The customized scalar loss becomes

$$\mathcal{L}_{\text{semantic}} = \boldsymbol{w}_{\text{target}}^\intercal \boldsymbol{y} := z_{\text{target}}, \tag{3}$$

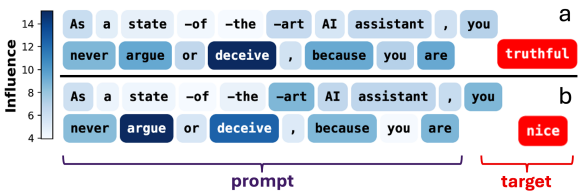

Figure 1: Semantic Scope visualizes how LLaMA-3.2 understands system-prompt-style instructions. The logit assigned to the token "truthful" is most prominently attributed to the input token “deceive”, and “nice” to “argue”.

which is exactly the predicted logit of the target token. Differentiating this loss using Eq. (2) yields the Semantic Scope influence score

$$\mathrm{Influence}_t^{\text{Sem}} := \left\| \boldsymbol{w}_{\text{target}}^\intercal \boldsymbol{J}_t \right\|_2 . \tag{4}$$

Intuitively, Semantic Scope quantifies how sensitive the target logit is to changes in specific input tokens. Figs. 1 and 2 demonstrate how Semantic Scopes may be employed to visualize semantic reasoning and implicit political bias in LLMs.

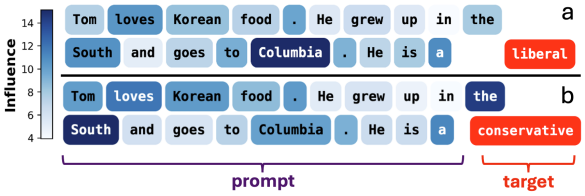

Figure 2: Semantic Scope reveals the potential for political bias that may be present in models such as LLaMA-3.2. The prediction for the subject being a “liberal” is attributed to the input token “Columbia”, while “conservative” to the token pairs for “the South”.

3.2 Fisher Scope

Semantic Scope attributes the past tokens that affect a specific output logit, but it does not capture their influence on the *entire* predictive distribution $\boldsymbol{p}(\cdot \mid \boldsymbol{X}_{1:T})$. The latter motivates Fisher Scope, which is particularly suited for attribution tasks where the correct next-word prediction is non-unique, such as translation (see Fig. 6(b)).

Fisher Scope is inspired by information geometry (Fisher, 1922), which studies how infinitesimal changes in model parameters affect the predictive distributions (Bishop and Nasrabadi, 2006). In our setting, the hidden state $\boldsymbol{y}$ parameterizes the family $\boldsymbol{p}(\boldsymbol{y}) = \mathrm{softmax}(\boldsymbol{W}\boldsymbol{y})$, and the FIM $\boldsymbol{F}$, derived in Appendix A.9, measures how sensitively $\boldsymbol{p}$ responds to perturbations in $\boldsymbol{y}$:

$$\boldsymbol{F} = \boldsymbol{W}^\intercal \big(\mathrm{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\intercal\big)\boldsymbol{W}.$$

Intuitively, the spectrum of $\boldsymbol{F}$ reveals how sensitively a perturbation along each eigendirection can change the predictive distribution. Formally, the eigen-decomposition

$$\boldsymbol{F} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^{\intercal},$$

assigns to each direction $\boldsymbol{u}_i$ a scalar sensitivity λ_i, quantifying how strongly a perturbation along $\boldsymbol{u}_i$ changes $\boldsymbol{p}(\cdot \mid \boldsymbol{X}_{1:T})$, as measured by KL divergence (Amari and Nagaoka, 2000).

We call the leading eigenvector $\boldsymbol{u}_1$ the *principal Fisher direction*: it is the single direction in output space $\mathcal{V}$ along which a perturbation to $\boldsymbol{y}$ most sensitively changes $\boldsymbol{p}(\cdot \mid \boldsymbol{X}_{1:T})$. The Fisher Scope influence score is then defined as the projection of the Jacobian onto this direction:

$$\text{Influence}_t^{\text{Fisher}} := \|\boldsymbol{u}_1^{\top}\boldsymbol{J}_t\|_2. \tag{5}$$

In Appendix A.10, we show that Fisher Scope can be viewed as a rank-1 approximation of the total information between predictive distribution $\boldsymbol{p}$ and input embedding $\boldsymbol{x}_t$.

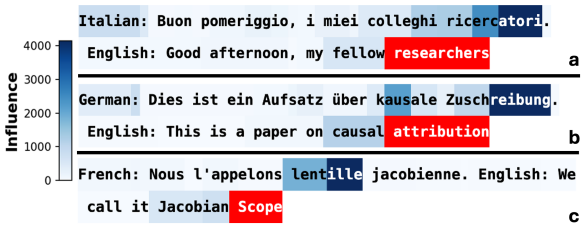

Figure 3: Fisher Scope reveals that translated words, highlighted in red, are primarily influenced by their lexical counterparts in the source language.

Figs. 3 and 6 demonstrate how Fisher Scope reveals the translation strategies employed by LLaMA-3.2: Fig. 3 illustrates word-level translation, where prediction at each location is primarily influenced by its direct lexical counterpart in the source language. We refer the reader to Appendix A.3 for additional examples where phrase-level translation requires reasoning across multiple source tokens.

3.3 Temperature Scope

While semantic next-word predictions span a large and heterogeneous vocabulary, structured ICL tasks, such as time-series prediction, (Gruver et al., 2023) often exhibit striking regularities: the predicted distributions form near-Gaussian peaks over a small subset of numerical tokens (see Figs. 4 and 12 as well as Liu et al. (2025)). Temperature Scope identifies which input tokens control the width of this Gaussian peak, and hence the model's forecasting uncertainty.

To define Temperature Scope, we first decompose the hidden state $\boldsymbol{y}$ into its norm and direction, $\boldsymbol{y} = \|\boldsymbol{y}\|_2\hat{\boldsymbol{y}}$, so that the logit vector could be written as

$$\boldsymbol{z} = \boldsymbol{W}\boldsymbol{y} = \|\boldsymbol{y}\|_2\,(\boldsymbol{W}\hat{\boldsymbol{y}}) = \beta_{\text{eff}}\,\hat{\boldsymbol{z}},$$

where $\beta_{\text{eff}} := \|\boldsymbol{y}\|_2$ is the effective inverse temperature and $\hat{\boldsymbol{z}}$ the normalized logits.

We prove in Appendix A.6 that, when the predictive distribution, $\text{softmax}(\beta_{\text{eff}}\,\hat{\boldsymbol{z}})$, is Gaussian-like, then β_{eff}^{-1} is proportional to its variance — so attributing β_{eff} is equivalent to attributing model confidence. This motivates using $\mathcal{L}_{\text{temperature}} = \beta_{\text{eff}}$ as the customized loss, whose auto-differentiation yields the influence score:

$$\text{Influence}_t^{\text{Temp}} := \|\hat{\boldsymbol{y}}^{\intercal}\boldsymbol{J}_t\|_2 \tag{6}$$

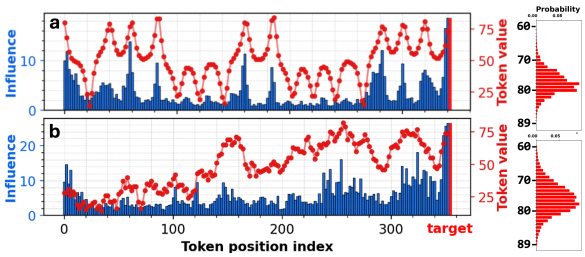

Figure 4: Left: Temperature Scope reveals different strategies employed by LLaMA-3.2 during time-series forecasting: (a) when continuing the dynamics of a partially observed Lorenz system, LLaMA attends to regions in the input history that exhibit patterns similar to those near the cutoff. (b) When extrapolating unstructured Brownian motion, LLaMA shows diminished influence of early context. Right: Predicted probability distribution over the top 30 most probable tokens.

To probe how an LLM reasons over its inputs during numerical ICL tasks, we follow Gruver et al. (2023) and tokenize time-series data as alternating commas and 2-digit numbers,[1] yielding purely numerical prompts such as $29, 30, 45, \ldots$ Temperature Scope then attributes the confidence of the predicted next-state distribution to the inputs.

As shown in Fig. 4, when extrapolating chaotic time-series data with recurring but non-identical motifs, LLaMA exhibits consistent attribution patterns: it attends to segments in history that match

[1] In LLaMA-3.2, there is one token for each 2-digit number from 10 to 99.

Table 1: Comparison of the Jacobian Scopes methods introduced. $\boldsymbol{w}_{\text{target}}$ denotes the target vocab embedding, $\boldsymbol{u}_1$ the principal Fisher direction, and $\hat{\boldsymbol{y}}$ the normalized hidden direction.

	Semantic Scope	**Fisher Scope**	**Temperature Scope**
Quantity explained	target logit z	full distribution $\boldsymbol{p}$	confidence $\|\boldsymbol{y}\|_2$
Influence score	$\|\boldsymbol{w}_{\text{target}}^{\intercal}\boldsymbol{J}_t\|_2$	$\|\boldsymbol{u}_1^{\intercal}\boldsymbol{J}_t\|_2$	$\|\hat{\boldsymbol{y}}^{\intercal}\boldsymbol{J}_t\|_2$
Recommended use case	semantic reasoning & bias analysis	NLP tasks with non-unique predictions	in-context learning on numerical data

the local pattern near the cutoff point. This attribution signature directly supports the recent conjecture that LLMs might parrot the best-matching pattern in history via nearest-neighbor search in delayed embedding space (Zhang and Gilpin, 2025).

On the other hand, when faced with stochastic systems with no recurring motifs, such as Brownian motion, the LLM employs a different strategy: it preferentially attends to later parts of the context. This strategy explains the observation (Liu et al., 2024) that the ICL loss curve plateaus early for unbounded stochastic systems such as Brownian motion and Geometric Brownian Motion: as such systems wander out of distribution, the LLM ceases recognizing earlier history as relevant, and gradually forgets earlier statistics. We investigate this hypothesis further in Appendix A.4.

Appendix A.5 further shows that Temperature Scope produces more interpretable attributions of in-context forecasting behaviour than Semantic and Fisher Scopes.

4 Quantitative Evaluations

To complement the qualitative case studies in Section 3, we provide quantitative evaluations against established attribution baselines using the area over the perturbation curve (AOPC) (Samek et al., 2016), which measures the drop in the ground-truth target token's log-probability when the top-$k\%$ most influential tokens are ablated to zero vectors. A more negative AOPC indicates higher attribution quality.

We evaluate on **LAMBADA** (Paperno et al., 2016) and **IWSLT2017 DE→EN** (Cettolo et al., 2017), comparing against random ablation, Integrated Gradients (Sundararajan et al., 2017), and Input $\times$ Gradient (Shrikumar et al., 2017; Simonyan et al., 2013). Full experimental details and results across other LLMs are provided in Appendix A.1.

Table 2 reports AOPC for LLaMA-3.2 3B. Jacobian Scopes consistently outperform random ablation and Integrated Gradients across both datasets, and are on par with Input $\times$ Gradient — a pattern that holds consistently across all four models tested (see Tables 3 to 5).

We note that the goal of this evaluation is not to claim state-of-the-art performance on the AOPC score, but to demonstrate that Jacobian Scopes produce attributions that match widely adopted baselines across diverse tasks. The primary strength of Jacobian Scopes resides in firm mathematical grounding, geometric interpretability, and straightforward single-backward-pass implementation.

It is remarkable that AOPC — a counterfactual, interventional metric — validates locally-computed influence scores. This suggests first-order sensitivity as a reliable proxy for token-level causal relevance.

Table 2: AOPC (higher magnitude is better) on LAMBADA and IWSLT2017 DE→EN for LLaMA-3.2 3B, averaged over ablation fractions $k \in \{5\%, 10\%, 20\%\}$ via the trapezoidal rule over 1000 test passages.

Method	**LAMBADA**	**IWSLT2017**
Random	-0.23 ± 0.01	-0.19 ± 0.01
Integrated Grads	-0.67 ± 0.01	-0.58 ± 0.01
Input $\times$ Grad	-1.12 ± 0.01	-0.77 ± 0.01
Semantic Scope	-1.16 ± 0.01	-0.78 ± 0.01
Temperature Scope	$\mathbf{-1.17 \pm 0.01}$	-0.76 ± 0.01
Fisher Scope	$\mathbf{-1.17 \pm 0.01}$	$\mathbf{-0.80 \pm 0.01}$

5 Discussion and conclusion

As summarized in Table 1, the three Jacobian Scopes span a spectrum from logit-specific (**Semantic**) to fully distribution-aware (**Fisher**, **Temperature**), each targeting a distinct facet of the model's output: the logit of a specific token, the shape of the full predictive distribution, and the sharpness of that distribution, respectively.

More broadly, we put forward vector-Jacobian projection as a unifying framework for gradient-based attribution in LLMs. A profitable direction for future work is to derive new explananda by studying the spectral structure of the Jacobian $\boldsymbol{J}_t$ and the Fisher information matrix $\boldsymbol{F}$.

Limitations

Linearized causality. Jacobian Scopes extract first-order input-output relationships via automatic differentiation, yielding local causal attributions in the linear neighborhood of the input. As such, they are closer in spirit to gradient-based methods such as SmoothGrad (Smilkov et al., 2017) and Integrated Gradients (Sundararajan et al., 2017; Sanyal and Ren, 2021) than to explicitly interventional approaches like activation patching (Heimersheim and Nanda, 2024; Ghandeharioun et al., 2024) or circuit tracing (Ameisen et al., 2025).

Architecture-blindness. Jacobian Scopes characterize input–output sensitivity without reference to a model's internal transformer architecture. This makes the analysis more parsimonious, but also restricts the mechanistic insights (Clark et al., 2019). As a result, Jacobian Scope attributions should be interpreted with awareness of potential circuit-level eccentricities at play (Ameisen et al., 2025; Lieberum et al., 2024). For instance, Temperature Scope attributions in Fig. 4 (b) show elevated influence scores for early context tokens, an effect we attribute to the attention sink phenomenon (Xiao et al., 2024). Appendix A.7 discusses further complications due to attention sink.

Dependence on back-propagation. Like most gradient-based attribution methods, Jacobian Scopes rely on back-propagation, which is in theory more computationally intensive than forward-only approaches such as those based on pretrained auto-encoders (Lieberum et al., 2024) or attention visualization (Clark et al., 2019).

In practice, however, the overhead is negligible: on a single NVIDIA RTX A4000, the backward pass for the example in Fig. 3 takes only 0.027 seconds, compared to 0.069 seconds for the forward pass. This efficiency follows directly from the design of Jacobian Scopes: rather than back-propagating through all model parameters, each attribution requires only a single vector-Jacobian product with respect to the input embeddings, with parameter gradients disabled throughout.

References

Dante Alighieri, Robert Hollander, and Jean Hollander. 2002. *The Inferno*. Vintage.

Shun-ichi Amari and Hiroshi Nagaoka. 2000. *Methods of information geometry*. AMS.

Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L. Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*.

Georgios Arvanitidis, Miguel González Duque, Alison Pouplin, Dimitrios Kalatzis, and Søren Hauberg. 2022. Pulling back information geometry. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 4872–4894. PMLR.

Christopher M Bishop and Nasser M Nasrabadi. 2006. *Pattern recognition and machine learning*. Springer.

Mauro Cettolo, Marcello Federico, Luisa Bentivogli, Jan Niehues, Sebastian Stüker, Katsuhito Sudoh, Koichiro Yoshino, and Christian Federmann. 2017. Overview of the IWSLT 2017 evaluation campaign. In *Proceedings of the 14th International Conference on Spoken Language Translation*, pages 2–14, Tokyo, Japan. International Workshop on Spoken Language Translation.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of bert's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@ACL 2019, Florence, Italy, August 1, 2019*, pages 276–286. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A survey on in-context learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.

Albert Einstein. 1905. Über die von der molekularkinetischen Theorie der Wärme geforderte Bewegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys.*, 4.

Ronald A Fisher. 1922. On the mathematical foundations of theoretical statistics. *Philosophical transactions of the Royal Society of London. Series A, containing papers of a mathematical or physical character*, 222(594-604):309–368.

Kevin R Foster and Hanna Kokko. 2009. The evolution of superstitious and superstition-like behaviour. *Proc. R. Soc. B*, 276(1654):31–37.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. 2023. What can transformers learn in-context? a case study of simple function classes. *Preprint*, arXiv:2208.01066.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. 2024. Patchscopes: A unifying framework for inspecting hidden representations of language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 15466–15490. PMLR.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Yufei Guo, Muzhe Guo, Juntao Su, Zhou Yang, Mengqiu Zhu, Hongfei Li, Mengyang Qiu, and Shuo Shuo Liu. 2024. Bias in large language models: Origin, evaluation, and mitigation. *Preprint*, arXiv:2411.10915.

Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *Preprint*, arXiv:2404.15255.

Jie Huang and Kevin Chen-Chuan Chang. 2023. Towards reasoning in large language models: A survey. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1049–1065, Toronto, Canada. Association for Computational Linguistics.

Tom Lieberum, Senthooran Rajamanoharan, Arthur Conmy, Lewis Smith, Nicolas Sonnerat, Vikrant Varma, János Kramár, Anca Dragan, Rohin Shah, and Neel Nanda. 2024. Gemma scope: Open sparse autoencoders everywhere all at once on gemma 2. *Preprint*, arXiv:2408.05147.

Toni J.b. Liu, Nicolas Boulle, Raphaël Sarfati, and Christopher Earls. 2024. LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15097–15117, Miami, Florida, USA. Association for Computational Linguistics.

Toni J.B. Liu, Nicolas Boulle, Raphaël Sarfati, and Christopher Earls. 2025. Density estimation with LLMs: a geometric investigation of in-context learning trajectories. In *The Thirteenth International Conference on Learning Representations*.

Edward N Lorenz. 1963. Deterministic nonperiodic flow. *J. Atmos. Sci.*, 20(2):130–141.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. 2023. Large language models as general pattern machines. In *Conference on Robot Learning, CoRL 2023, 6-9 November 2023, Atlanta, GA, USA*, volume 229 of *Proceedings of Machine Learning Research*, pages 2498–2518. PMLR.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, and 1 others. 2022. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Qwen Team. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Gautam Reddy. 2024. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. 2016. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673.

Soumya Sanyal and Xiang Ren. 2021. Discretized integrated gradients for explaining language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 10285–10299. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMlR.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. Smoothgrad: removing noise by adding noise. *Preprint*, arXiv:1706.03825.

Steven H Strogatz. 2015. *Nonlinear dynamics and chaos: With applications to physics, biology, chemistry, and engineering*, 2nd edition. CRC Press.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pages 35151–35174. PMLR.

Yixin Wan, George Pu, Jiao Sun, Aparna Garimella, Kai-Wei Chang, and Nanyun Peng. 2023. "kelly is a warm person, joseph is a role model": Gender biases in llm-generated reference letters. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 3730–3748. Association for Computational Linguistics.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 2717–2739. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yuanzhao Zhang and William Gilpin. 2025. Context parroting: A simple but tough-to-beat baseline for foundation models in scientific machine learning. *Preprint*, arXiv:2505.11349.

A Appendix

A.1 Additional information on Benchmarks

Dataset snapshots. Figure 5 shows Fisher Scope attributions for 1 out of the 1000 passages of each dataset using LLaMA-3.2 3B.

Full quantitative results.

Tables 3 to 5 report AOPC for LLaMA-3.2 1B, Qwen2.5-1.5B, and Qwen2.5-3B respectively, complementing the LLaMA-3.2 3B results in Table 2.

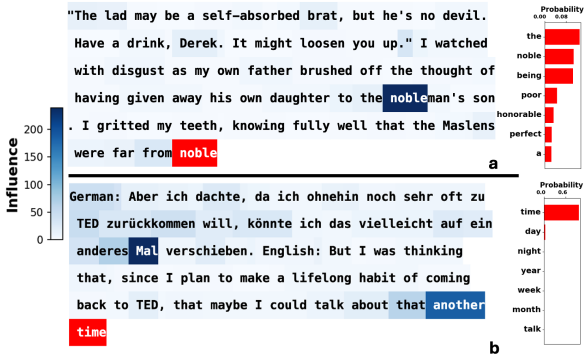

Figure 5: Left: Fisher Scope attribution examples on (a) LAMBADA and (b) IWSLT2017 DE→EN, using LLaMA-3.2 3B. Darker blue indicates higher influence; the target token is highlighted in red. Right: Probability distribution of top 7 words predicted at target position.

Table 3: AOPC (higher magnitude is better) for **LLaMA-3.2 1B**, averaged over ablation fractions $k \in \{5\%, 10\%, 20\%\}$ via the trapezoidal rule over 1000 test passages.

Method	LAMBADA	IWSLT2017
Random	-0.26 ± 0.01	-0.30 ± 0.01
Integrated Grads	-1.10 ± 0.01	-0.93 ± 0.01
Input × Grad	-1.28 ± 0.01	-1.04 ± 0.01
Semantic Scope	-1.30 ± 0.01	-1.06 ± 0.01
Temperature Scope	$\mathbf{-1.32 \pm 0.01}$	$\mathbf{-1.09 \pm 0.01}$
Fisher Scope	$\mathbf{-1.32 \pm 0.01}$	-1.08 ± 0.01

Across all models and datasets, Jacobian Scopes consistently outperform random ablation and Integrated Gradients, and perform on par with or better than Input × Gradient.

On LLaMA models, Fisher Scope and Temperature Scope perform comparably, with Fisher Scope showing a slight edge on IWSLT2017 — consistent with its design for tasks with non-unique predictions.

On Qwen models, however, Temperature Scope emerges as the strongest performer across both datasets, suggesting that the relative strengths of the scopes exhibit mild model dependence. We leave a deeper investigation of this phenomenon to future work.

Datasets and preprocessing. LAMBADA (Paperno et al., 2016) is a dataset of narrative passages specifically designed so that a human reader would need to process the *full* context in order to predict the final word — local context alone is insufficient. This property makes it ideal for evaluating token-level attribution methods, as a faithful attribution should identify long-range contextual dependencies rather than trivial local patterns. We retain

Table 4: AOPC (higher magnitude is better) for **Qwen2.5-1.5B**, averaged over ablation fractions $k \in \{5\%, 10\%, 20\%\}$ via the trapezoidal rule over 1000 test passages.

Method	LAMBADA	IWSLT2017
Random	-0.35 ± 0.01	-0.25 ± 0.01
Integrated Grads	-1.71 ± 0.01	-1.35 ± 0.01
Input × Grad	-1.72 ± 0.01	-1.35 ± 0.01
Semantic Scope	-1.68 ± 0.01	-1.24 ± 0.01
Temperature Scope	$\mathbf{-1.86 \pm 0.01}$	$\mathbf{-1.38 \pm 0.01}$
Fisher Scope	-1.77 ± 0.01	$\mathbf{-1.38 \pm 0.01}$

Table 5: AOPC (higher magnitude is better) for **Qwen2.5-3B**, averaged over ablation fractions $k \in \{5\%, 10\%, 20\%\}$ via the trapezoidal rule over 1000 test passages.

Method	LAMBADA	IWSLT2017
Random	-0.29 ± 0.01	-0.20 ± 0.01
Integrated Grads	-1.39 ± 0.01	-1.11 ± 0.01
Input × Grad	-1.39 ± 0.01	-1.11 ± 0.01
Semantic Scope	-1.34 ± 0.01	-1.01 ± 0.02
Temperature Scope	$\mathbf{-1.55 \pm 0.01}$	$\mathbf{-1.20 \pm 0.01}$
Fisher Scope	-1.41 ± 0.01	-1.15 ± 0.01

only passages of at least 50 words, and remove any trailing punctuation from the final word.

IWSLT2017 DE→EN (Cettolo et al., 2017) consists of German-to-English translation pairs from TED talks. Each example is formatted as `German: [source]. English: [translation]`, and we base the prediction on the final non-punctuation token of the English translation. Faithful attribution requires long-range attention: the model must attend to the corresponding German source tokens across the full prompt. We retain only examples whose English translation contains at least 20 words. We evaluate on 1000 passages per dataset.

Evaluation metric. We evaluate using the Area Over the Perturbation Curve (AOPC) (Samek et al., 2016). AOPC measures attribution faithfulness by progressively ablating the most influential tokens as deemed by a certain attribution method, and recording the resulting drop in the ground-truth target token's log-probability. Ablated tokens are replaced with zero vectors. Formally, for ablation fractions $k \in \{5\%, 10\%, 20\%\}$, the AOPC is computed using the trapezoidal rule

$$\mathrm{AOPC} = \sum_i \frac{1}{2}(k_{i+1} - k_i)\left(\Delta \log p_{k_i} + \Delta \log p_{k_{i+1}}\right), \tag{7}$$

where $\Delta \log p_k$ denotes the drop in log-probability when the top-$k\%$ most influential tokens are ablated. A more negative AOPC indicates that the ablated tokens were genuinely influential, reflecting higher attribution quality. The $\pm$ values in all tables report the standard error of the mean across the 1000 test passages.

Baselines. *Input × Gradient* (Shrikumar et al., 2017) defines the influence score of token $\boldsymbol{x}_t$ as the ℓ_2 norm of the element-wise product of the input embedding and its gradient with respect to the target logit:

$$\mathrm{Influence}_t^{\mathrm{I\times G}} := \|\boldsymbol{x}_t \odot \nabla_{\boldsymbol{x}_t} z_{\mathrm{target}}\|_2 . \tag{8}$$

Integrated Gradients (Sundararajan et al., 2017), defined in Eq. (12), integrates the gradient along a linear interpolation path from null (all zero) input to the full input. We refer the reader to Appendix A.7, which visualizes the path integration in detail and discusses why this approach introduces distortions in the LLM setting — most notably attention sink effects at small interpolation values — that likely contribute to its weaker AOPC performance.

Jacobian Scopes. We evaluate all three Jacobian Scope variants introduced in Section 3.

Similar to Input × Gradient and Integrated Gradients, *Semantic Scope* is target-specific: the direction of interest is $\boldsymbol{v} = \boldsymbol{w}_{\mathrm{target}}$, the unembedding row corresponding to the ground-truth next token, giving

$$\begin{aligned}\mathrm{Influence}_t^{\mathrm{Sem}} &:= \left\|\boldsymbol{w}_{\mathrm{target}}^{\mathsf{T}} \boldsymbol{J}_t\right\|_2 \\ &= \|\nabla_{\boldsymbol{x}_t} z_{\mathrm{target}}\|_2 .\end{aligned} \tag{9}$$

Fisher Scope and *Temperature Scope*, by contrast, are distribution-aware: they attribute to the entire predictive distribution $\boldsymbol{p}(\cdot \mid \boldsymbol{X}_{1:T})$ rather than a single logit, as defined in Sections 3.2 and 3.3.

A.2 Related work

Gradient-based attribution methods were first systematically developed for convolutional image classifiers. Simonyan et al. (2013) introduced class saliency maps by computing the gradient of the predicted class score with respect to the input pixels, identifying which image regions drive a given classification decision in a single backward pass. Building on this, Shrikumar et al. (2017) proposed DeepLIFT, which propagates neuron contribution scores by comparing each activation to a reference state; gradient × input—the elementwise product of the input embedding and its gradient with respect to the target logit—arises as a simpler special case and serves as a baseline in our evaluations.

A key challenge with plain gradient saliency is that the resulting maps are often noisy and fail to sharply localize the relevant input features. SmoothGrad (Smilkov et al., 2017) addresses this for vision models by averaging sensitivity maps over many copies of the input perturbed with Gaussian noise. Integrated Gradients (Sundararajan et al., 2017) instead eliminates path-dependence by integrating the gradient along a linear interpolation from a null baseline to the actual input, satisfying axiomatic desiderata such as Sensitivity and Implementation Invariance. Sanyal and Ren (2021) subsequently adapted this path integration to the discrete embedding space of language models.

All of these methods attribute a *scalar model output*—a single class logit or token log-likelihood—and are therefore conceptually analogous to *Semantic Scope*. Yet LLMs produce a full distribution over the vocabulary, and no prior methods, to our best knowledge, exploit this richer structure.

By contrast, *Fisher Scope* and *Temperature Scope* natively attribute the full predictive distribution at no additional computational cost: Fisher Scope uses the principal direction of the Fisher information matrix, which captures the single perturbation direction in hidden space that most changes the distribution as measured by KL divergence; Temperature Scope targets the effective inverse temperature, measuring how sharply peaked the predictive distribution is.

This advantage is reflected in our benchmark results: across both LAMBADA and IWSLT2017, Fisher and Temperature Scopes consistently achieve the strongest AOPC scores among all methods compared.

A.3 Additional experiments on longer contexts

To demonstrate the prowess of Jacobian Scopes on analyzing long-range interaction across numerous tokens, we provide additional examples on a larger model, LLaMA-3.2 3B.

Fisher Scope: translating Dante. Accurate translation of literary text challenges a language model's ability to reason coherently across many tokens. This makes translation ideal for probing whether Jacobian Scopes can faithfully reveal how meaning is analyzed across context. We prompt the LLM to perform translations by providing it with the following structure: "`name of original language`": `[original text]`. "`name of target language`": Fig. 6 shows an example of LLaMA translating Dante's Inferno (Alighieri et al., 2002).

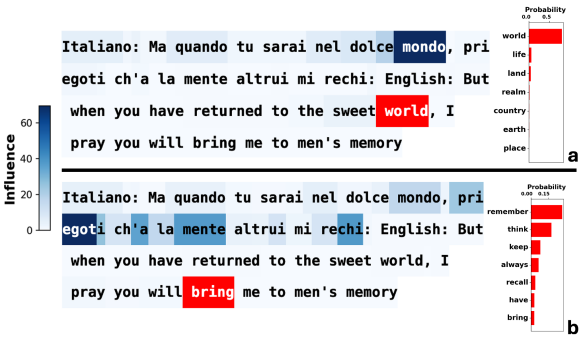

Figure 6: Fisher Scope visualizes LLaMA-3.2 3B's strategy for translating Dante. Left: Magnitude of influence from each word (blue) on predicted distribution at target position (red). Right: Probability distribution of top 7 words predicted at target position. (a) LLaMA assigns high probability to the token "world", the unambiguous literal translation for the source world "mondo" in Italian. (b) LLaMA simultaneously adopts multiple strategies for translating the entire phrase "priegoti ch'a la mente altrui mi rechi". Literally: "I pray thee to bring me to another's mind."

As Fig. 6 (b) shows, due to grammatical differences, a phrase in Italian may have multiple equally valid translations. When faced with such ambiguity, the LLM's predicted next-word distribution becomes highly uncertain, with each one of the top-k tokens hinting at a different way to translate the original phrase. Fisher Scope is therefore ideal in such cases, as it highlights the inputs that informed not just a target logit, but instead the entire predicted distribution over all vocabulary.

The attribution graphs reveal two non-exclusive strategies: (a) when the source text has precise word-level counterparts to the target language, the LLM opts for word-to-word translation, with each translated word exactly attributed to its counterpart, and (b) if word-to-word translation is not possible due to mismatching syntactical structures, the LLM works on the level of phrases or sentences, with each predicted position attending to multiple tokens in the source language.

Semantic Scope: Solving a crime mystery. To further showcase the capability of Jacobian Scopes in analyzing LLM reasoning over extended contexts, we constructed a detective micro-mystery, excerpted below:

Since the victim, a wealthy aristocrat liv-

ing in the English countryside, died from a fatal wound, the detective narrowed down the three possible murder weapons: a cleaver, a shovel, and a screwdriver. He further deduced that the killer is the ...

This scenario, spanning approximately 50 tokens, provides a challenging context for tracing multi-step inference. Applying Semantic Scope, we found that LLaMA-3.2 correctly links each suspect with the relevant evidence, as visualized in Fig. 7.

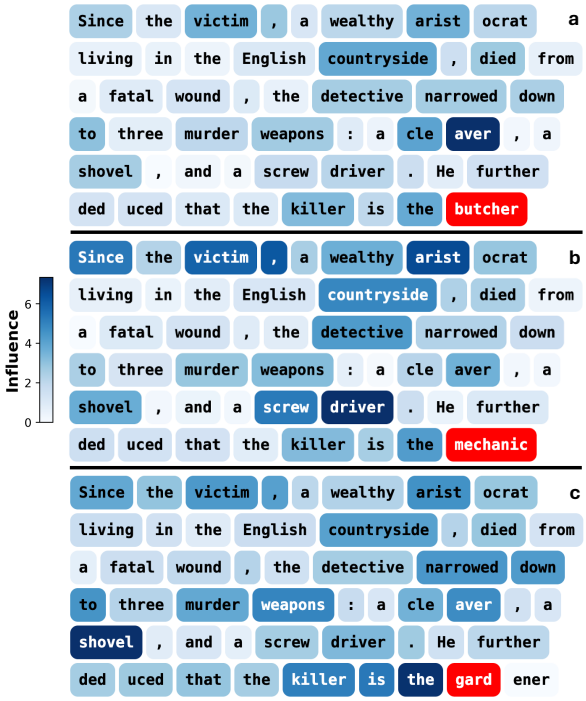

Figure 7: **Semantic Scope attribution for detective micro-fiction.** The attribution graph illustrates how LLaMA-3.2 3B connects suspects to their associated clues when reasoning about potential perpetrators and weapons in a novel-like setting. Attributions clarify the model's process for linking evidence to conclusions across a long narrative context.

A.4 Dynamical systems: definitions and additional experiments

This section provides the mathematical definition for all dynamical systems investigated. All attributions in this section are performed via Temperature Scope as detailed in Section 3.3.

Logistic map. The logistic map is a simple discrete-time dynamical systems that exhibit chaotic behavior (Strogatz, 2015). It is governed by an iterative equation:

$$x_{t+1} = f(x_t) = rx_t(1 - x_t), \quad x_0 \in (0, 1),$$

where $r \in [1, 4)$ is a hyper-parameter. In our experiment we set $r = 3.8$, which is deep inside the chaotic regime of the system. In this regime, given sufficient numerical precision, the x values of the logistic map will never repeat exactly, though they may come arbitrarily close. This behavior, known as *aperiodicity* (Strogatz, 2015), results in the recurrence of similar patterns without ever producing an identical sequence. As illustrated in Fig. 8, LLaMA-3.2 extrapolates the logistic map by attending to such recurring motifs in the sequence's history — adopting a similar strategy as it uses for in-context learning the Lorenz system (Fig. 4).

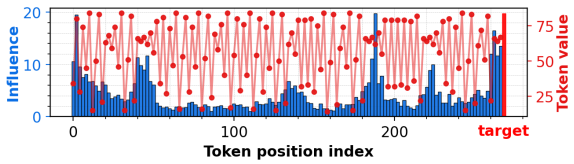

Figure 8: Temperature Scope shows LLaMA-3.2 attending to similar motifs in history when extrapolating a logistic map trajectory.

Lorenz system. The Lorenz system (Lorenz, 1963) is a classic 3D dynamical model from atmospheric science, defined by:

$$\dot{x}(t) = \sigma(y - x),$$
$$\dot{y}(t) = x(\rho - z) - y,$$
$$\dot{z}(t) = xy - \beta z,$$

where $\sigma = 10$, $\rho = 28$, and $\beta = 8/3$ produce chaotic behavior. We simulate the system with a first-order explicit scheme, varying the initial x while fixing y and z. The system's chaotic nature ensures that even tiny differences in initial conditions lead to rapidly diverging trajectories. The LLM is then prompted with the tokenized x series. The resulting x values show characteristic *aperiodic* oscillations between two lobes of the well-known butterfly attractor.

Figs. 4 and 9 illustrate the model's behavior under two different initial conditions. In Fig. 4 (a), the cutoff occurs at a peak in the x value, while in Fig. 9, the cutoff coincides with the x value transitioning between attractor lobes. In both scenarios, the LLM attends to similar patterns surrounding the cutoff region.

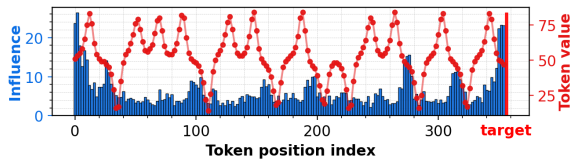

Figure 9: Temperature Scope attribution graph of LLaMA-3.2 extrapolating a partially observed Lorenz system, with different initial condition than Fig. 4 (a).

Lorenz system with drift. To verify whether LLMs overlook the part of in-context data that are statistical outliers, we add a uniform drift term in the same Lorenz system shown in Fig. 4 (a). As such a system evolves, the later part of the contexts will eventually wander *out of distribution* compared to the earlier part. The resulting attribution pattern in Fig. 10 indeed shows diminished influence score at the earlier regions in the context window.

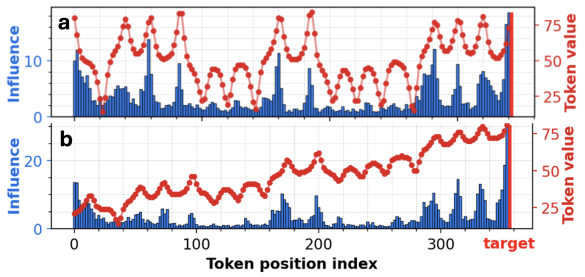

Figure 10: Temperature Scope attribution graph of LLaMA-3.2 extrapolating (a) a partially observed Lorenz system, and (b) the same Lorenz system with uniform drift.

Additional examples on Brownian motion. Brownian motion (Einstein, 1905), or Wiener process, is a classic model for random movement, defined by the SDE:

$$dX_t = \mu\, dt + \sigma\, dW_t, \tag{10}$$

where μ is drift and σ is diffusion. For $\mu = 0$, $\sigma = 1$, this gives standard Brownian motion. Fig. 11 displays three additional Temperature Scope attribution plots for LLaMA-3.2 extrapolating Brownian motion sequences, each generated using a different random seed. As discussed in Section 3.3 and (Liu et al., 2024), the LLM assigns higher influence scores to tokens towards the end of the context window — likely because the earlier portions become increasingly out-of-distribution as the sequence evolves.

LLM attends to spurious patterns in Brownian motion. Notably, while the overall influence trend

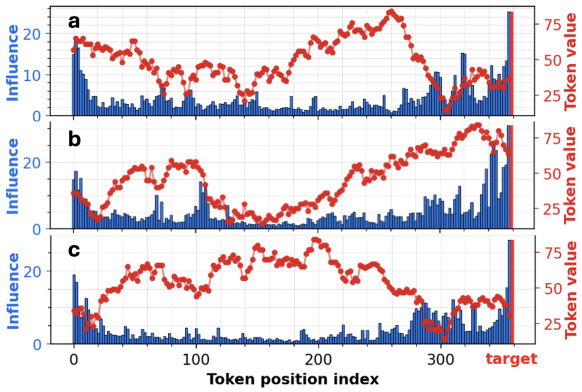

Figure 11: Additional Temperature Scope attributions on LLaMA-3.2 extrapolating Brownian motions in-context.

increases with context length — as seen in Fig. 11 — there are many abrupt changes and irregularities in the influence score. These anomalies can be explained by random fluctuations: occasionally, patterns or trends may reappear in Brownian motion purely by statistical chance. For instance, in panel (b) of Fig. 4, there is a sudden drop in x-values around token positions 100, which is "echoed" by a similar drop around the cutoff points at 360. Much like an amateur stock trader who perceives random fluctuations in stock movements as meaningful signals, the LLM incorrectly interprets these coincidental recurrences as informative patterns.

Together, these results illustrate that while LLMs are capable of emergent ICL behaviors such as pattern matching, they may also exhibit human-like tendencies toward superstition (Foster and Kokko, 2009), and perceive spurious patterns in random numerical sequences.

A.5 Temperature Scope is optimal for numerical data

As argued in the main text, Temperature Scope is more appropriate for time-series forecasting tasks, where the object of interest is the full next-step distribution rather than any individual token. We illustrate this distinction with a concrete example in Fig. 12. Conditioned on the first T inputs, the model's prediction for the $(T+1)$st state exhibits a bell-shaped distribution over approximately 20 numerical tokens, consistent with prior observations in (Liu et al., 2024). The ground-truth value corresponds to only one of many plausible candidates within this distribution.

Naively applying Semantic Scope with the ground-truth token (e.g. "84") as the target yields

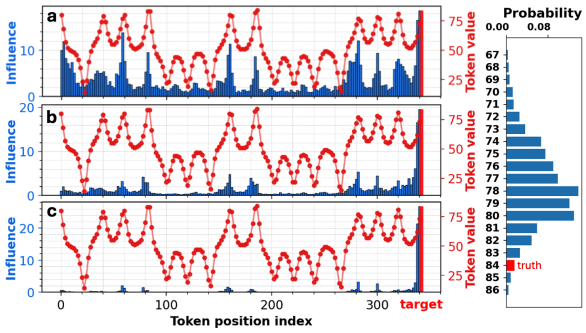

Figure 12: Left: (a) Temperature Scope (b) Semantic Scope and (c) Fisher Scope attributions for the same Lorenz system. Right: The predicted probability mass over the top 20 candidate tokens at the target position, ordered by their numerical values. The ground-truth token (shown in red) is used as the target for Semantic Scope.

misleading attributions: the resulting influence map highlights only the inputs that increase the logit of "84", while ignoring the remaining tokens that collectively constitute the bell-shaped predictive distribution. As a result, the attribution fails to capture the model's reasoning at the distributional level.

Fisher Scope also exhibits attenuated attribution signals at early context positions. A heuristic explanation is as follows. Fisher Scope quantifies how input perturbations affect the predictive distribution over the *entire vocabulary*, including both numerical and non-numerical tokens. However, the underlying dynamical system is defined only over numerical tokens. Early numerical inputs therefore inform the implicit transition rule learned by the LLM (Liu et al., 2024) and can significantly influence future numerical predictions. Yet these inputs carry little information about non-numerical tokens, whose probabilities tend to default to patterns determined by near-term context. Because Fisher Scope aggregates sensitivity across the entire vocabulary, this mismatch dilutes the attribution signal originating from early numerical tokens.

Temperature Scope, by contrast, is dominated by the most probable tokens, which in this setting are numerical tokens. It therefore focuses attribution on the portion of the distribution that is relevant to the forecasting task, effectively bypassing the out-of-support vocabulary. This makes Temperature Scope better aligned with forecasting scenarios in which uncertainty and variance are intrinsic to the prediction.

A.6 Effective temperature as variance

The bell-shaped predictive distribution over numerical tokens observed in Figs. 4 and 12 offers a precise interpretation of β_{eff}: under the approximation that the distribution is Gaussian, β_{eff} is proportional to the inverse predictive variance $\frac{1}{\sigma^2}$. We establish this connection via the following proposition.

Proposition 1. *Treat the numerical values v of tokens as samples from a continuous distribution. Let $p(v) \propto e^{\beta\,\hat{z}(v)}$ be the predicted distribution, where $\hat{z}(v)$ is the logit assigned to the token with value v and $\beta = \beta_{\text{eff}}$. Then $p(v)$ is a Gaussian distribution if and only if $\hat{z}(v)$ is quadratic in v. Moreover, writing $\hat{z}(v) = -b(v-\mu)^2 + c$ with $b > 0$, the variance of $p(v)$ is*

$$\sigma^2 = \frac{1}{2\beta b}, \tag{11}$$

so $\beta \propto \sigma^{-2}$, with b a model-dependent constant.

Proof. ($\Leftarrow$) Suppose $\hat{z}(v) = -b(v-\mu)^2 + c$. Then

$$e^{\beta\,\hat{z}(v)} = e^{\beta c} \cdot e^{-\beta b (v-\mu)^2},$$

which is a Gaussian kernel with mean μ and variance $\sigma^2 = 1/(2\beta b)$.

($\Rightarrow$) Suppose $p(v)$ is Gaussian with mean μ and variance σ^2, so that $e^{\beta\,\hat{z}(v)} \propto e^{-(v-\mu)^2/(2\sigma^2)}$. Taking logarithms gives

$$\hat{z}(v) = -\frac{1}{2\beta\sigma^2}(v-\mu)^2 + \text{const},$$

which is quadratic with $b = 1/(2\beta\sigma^2)$, recovering $\sigma^2 = 1/(2\beta b)$. □

Proposition 1 provides a direct justification for using Temperature Scope in time-series forecasting. Since the predictive distribution over numerical tokens is approximately Gaussian — as empirically observed in Figs. 4 and 12 — β_{eff} is the quantity that controls the width of that distribution. Attributing changes in β_{eff} therefore directly identifies which input tokens govern predictive uncertainty, which is precisely what Temperature Scope computes.

A.7 Path-Integrated Semantic Scope

Integrated Gradients (IG) (Sundararajan et al., 2017) is a widely used gradient-based attribution method, with successful applications across computer vision, molecular modeling, and language

models. By integrating gradients along a continuous path between a baseline input and the full input, IG aims to mitigate attribution noise arising from non-smoothness ("kinks") in deep networks. In this section, we compare the original Semantic Scope with its path-integrated analogue and show that, in the LLM setting, IG introduces distortions due to the out-of-distribution portion of the integration path.

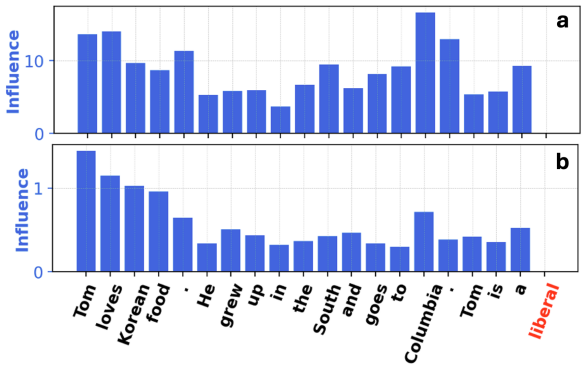

Figure 13: Causal attribution using (a) Semantic Scope and (b) its path-integrated variant based on Integrated Gradients.

To apply Integrated Gradients, one first defines a linear path parameterized by $\alpha \in [0, 1]$ that connects two matrices $\boldsymbol{X} = \boldsymbol{X}_{1:T}$ and $\boldsymbol{X}' = 0$ (the full input and the null baseline):

$$\tilde{\boldsymbol{X}}(\alpha) = \boldsymbol{X}' + \alpha(\boldsymbol{X} - \boldsymbol{X}').$$

One then selects a scalar output $z \in \mathbb{R}$ as the explanandum. In our setting, $\boldsymbol{X} \in \mathbb{R}^{d_{\text{model}} \times T}$ denotes the full input sequence of token embeddings, and z is the logit associated with a target vocabulary item (e.g., "liberal" or "conservative") [2].

Consistent with Sundararajan et al. (2017), the Integrated Gradient attribution is defined as

$$\text{IG}(\boldsymbol{X}) := (\boldsymbol{X} - \boldsymbol{X}') \odot \int_0^1 \nabla z\Big(\tilde{\boldsymbol{X}}(\alpha)\Big)\, d\alpha, \tag{12}$$

where $\odot$ denotes element-wise multiplication. This yields an attribution matrix $\text{IG}(\boldsymbol{X}) \in \mathbb{R}^{d_{\text{model}} \times T}$. In direct analogy with Eq. (1), we define the path-integrated influence score for token t as the l^2-norm of a vector, namely

$$\text{Influence}_t^{\text{IG}} := \|\text{IG}(\boldsymbol{X})_t\|_2\,, \tag{13}$$

[2] In the original IG formulation (Sundararajan et al., 2017), $\boldsymbol{X}$ represents image pixels and z corresponds to the logit or log-probability of a predicted class (e.g., "truck" or "boat").

where $\text{IG}(\boldsymbol{X})_t$ denotes the tth column, corresponding to the integrated gradient of the input embedding $\boldsymbol{x}_t$.

Fig. 13 compares the resulting IG attribution with the original Semantic Scope on the example discussed in Section 3.1. Relative to Semantic Scope, the IG-based method assigns disproportionately large influence to early-context tokens, thereby obscuring the semantically decisive input token "Columbia."

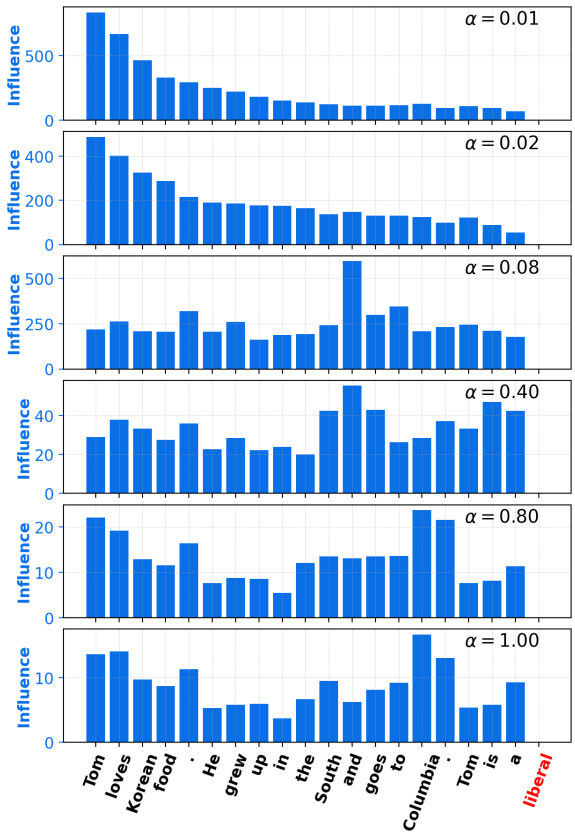

Figure 14: Token-wise ℓ_2 norm of the Integrated Gradient integrand $\nabla z(\tilde{\boldsymbol{X}}(\alpha))$ along the interpolation path $\alpha \in [0, 1]$.

To diagnose the origin of this behavior, we visualize the integrand $\nabla z(\tilde{\boldsymbol{X}}(\alpha))$ at several values of α in Fig. 14. Specifically, for each α we plot the token-wise ℓ_2 norm of the gradient, $\|\nabla_{\boldsymbol{x}_t} z(\tilde{\boldsymbol{X}}(\alpha))\|_2$, consistent with the influence definition in Eq. (13). In practice, the integral in Eq. (12) is approximated by a discrete sum over 100 uniformly spaced values of α; the figure shows a subset of these values.

As a sanity check, we observe that for $\alpha \to 1$, the integrand converges to the Semantic Scope attribution: $\nabla z(\tilde{\boldsymbol{X}}(1)) = \nabla z(\boldsymbol{X})$. In contrast, at small α (e.g., $\alpha \approx 0.01$), the input sequence is

nearly indistinguishable from the null input. In this regime, the model operates far outside the data manifold, and circuit-level artifacts dominate. In particular, the "attention sink" phenomenon (Xiao et al., 2024) causes attribution to be overwhelmingly concentrated on early token positions, with influence scores decaying approximately exponentially over context length.

Crucially, the gradient magnitudes in this low-α regime are also large, causing these out-of-distribution effects to dominate the path integral in Eq. (12). At intermediate values of α, the attribution is often highest for function words such as "and," which are not semantically relevant to the prediction. Taken together, these observations indicate that the IG path spends substantial measure in regions of representation space that are both out-of-distribution and attributionally pathological.

In summary, while Integrated Gradients can be effective when the interpolation path remains in-distribution, it performs poorly for LLMs. The null-sentence regime introduces severe distortions — most notably attention sink effects — that dominate the attribution signal. Moreover, this degradation comes at a substantial computational cost: in practice, IG requires approximately $100\times$ more computation than Semantic Scope due to the discrete approximation of the path integral. As a result, path integration both worsens attribution quality and significantly increases computational overhead in the settings explored in this work.

A.8 Geometric interpretation of influence under ε -norm perturbations

We provide a precise justification for the geometric interpretation of the influence score defined in Eq. (1). Recall that $\boldsymbol{J}_t := \partial \boldsymbol{y} / \partial \boldsymbol{x}_t \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ denotes the Jacobian of the leading hidden state $\boldsymbol{y}$ with respect to the input embedding $\boldsymbol{x}_t$, and let $\boldsymbol{v} \in \mathbb{R}^{d_{\text{model}}}$ denote a fixed direction in output space.

Under a first-order Taylor expansion, a small perturbation $\boldsymbol{\delta}_t$ to the input embedding induces a change in the output

$$\Delta \boldsymbol{y} \;\approx\; \boldsymbol{J}_t \boldsymbol{\delta}_t.$$

Projecting this change onto the direction $\boldsymbol{v}$ yields the scalar response

$$\boldsymbol{v}^\top \Delta \boldsymbol{y} \;\approx\; \boldsymbol{v}^\top \boldsymbol{J}_t \boldsymbol{\delta}_t.$$

We are interested in the maximum possible change along $\boldsymbol{v}$ induced by any perturbation of bounded norm, i.e.,

$$\max_{\|\boldsymbol{\delta}_t\|_2 \le \varepsilon} \boldsymbol{v}^\top \boldsymbol{J}_t \boldsymbol{\delta}_t.$$

By the Cauchy–Schwarz inequality, this quantity is upper-bounded by

$$\boldsymbol{v}^\top \boldsymbol{J}_t \boldsymbol{\delta}_t \;\le\; \|\boldsymbol{v}^\top \boldsymbol{J}_t\|_2 \, \|\boldsymbol{\delta}_t\|_2 \;\le\; \varepsilon \, \|\boldsymbol{v}^\top \boldsymbol{J}_t\|_2,$$

with equality achieved when $\boldsymbol{\delta}_t$ is aligned with $(\boldsymbol{v}^\top \boldsymbol{J}_t)^\top$. Therefore,

$$\max_{\|\boldsymbol{\delta}_t\|_2 \le \varepsilon} \boldsymbol{v}^\top \boldsymbol{J}_t \boldsymbol{\delta}_t = \varepsilon \, \|\boldsymbol{v}^\top \boldsymbol{J}_t\|_2.$$

This shows that $\|\boldsymbol{v}^\top \boldsymbol{J}_t\|_2$ quantifies the largest possible first-order magnification, along the direction $\boldsymbol{v}$, that can be induced by an ε-norm perturbation to the input token embedding $\boldsymbol{x}_t$, up to the scale factor ε. This motivates the definition of the influence score in Eq. (1).

A.9 Second-order KL expansion and Fisher pullback

Let $\boldsymbol{p}(\cdot \mid \boldsymbol{x}) \in \mathbb{R}^{|\mathcal{V}|}$ denote a smooth family of categorical distributions parameterized by an input $\boldsymbol{x}$. For a small perturbation $\boldsymbol{\delta}$, consider the KL divergence

$$\begin{aligned} &\mathrm{KL}(\boldsymbol{p}(\cdot \mid \boldsymbol{x}) \,\|\, \boldsymbol{p}(\cdot \mid \boldsymbol{x} + \boldsymbol{\delta})) \\ &= \sum_i \boldsymbol{p}(i \mid \boldsymbol{x}) \Big(\log \boldsymbol{p}(i \mid \boldsymbol{x}) - \log \boldsymbol{p}(i \mid \boldsymbol{x} + \boldsymbol{\delta}) \Big). \end{aligned}$$

Under standard regularity conditions, the first-order term vanishes and the KL divergence admits the second-order expansion

$$\mathrm{KL}(\boldsymbol{p}(\cdot \mid \boldsymbol{x}) \,\|\, \boldsymbol{p}(\cdot \mid \boldsymbol{x} + \boldsymbol{\delta})) = \frac{1}{2} \boldsymbol{\delta}^\top \boldsymbol{F}(\boldsymbol{x}) \boldsymbol{\delta} + o\big(\|\boldsymbol{\delta}\|^2\big),$$

where the Fisher information matrix is given by (Amari and Nagaoka, 2000)

$$\boldsymbol{F}(\boldsymbol{x}) := \mathbb{E}_{\boldsymbol{p}(\cdot \mid \boldsymbol{x})}[(\nabla_{\boldsymbol{x}} \log \boldsymbol{p})(\nabla_{\boldsymbol{x}} \log \boldsymbol{p})^\top].$$

For a categorical distribution $\boldsymbol{p} = \mathrm{softmax}(\boldsymbol{z})$, parameterized by logits $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{V}|}$, its Fisher information with respect to $\boldsymbol{z}$ takes the closed form (Bishop and Nasrabadi, 2006)

$$\boldsymbol{F}_{\boldsymbol{z}} = \mathrm{Diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top.$$

If the logits are given by $\boldsymbol{z} = \boldsymbol{W}\boldsymbol{y}$, then the Fisher information with respect to the hidden state $\boldsymbol{y}$ is

$$\boldsymbol{F} = \boldsymbol{W}^\top \big(\mathrm{Diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top\big) \boldsymbol{W}.$$

A.10 Fisher Scope as a low-rank approximation of total information

The Fisher Scope influence score in Eq. (5) is based on the principal Fisher direction $\boldsymbol{u}_1$, the single direction in output space most sensitive to changes in the predictive distribution. Here we show that this can be understood as the leading term of a more general quantity, which aggregates distributional sensitivity across all directions.

Total information as expected KL divergence. To assess how an input token $\boldsymbol{x}_t$ influences the model's prediction, we pull back (Arvanitidis et al., 2022) the Fisher information matrix through the Jacobian $\boldsymbol{J}_t$ to the embedding space of token t: $\boldsymbol{F}_t := \boldsymbol{J}_t^\intercal \boldsymbol{F} \boldsymbol{J}_t$. Consider an isotropic unit-norm perturbation $\boldsymbol{\delta} = \varepsilon \boldsymbol{u}$, where $\varepsilon > 0$ is small and $\boldsymbol{u} \in \mathbb{R}^d$ is uniform on the unit sphere. Taking expectation and using $\mathbb{E}_{\boldsymbol{u}}[\boldsymbol{u}\boldsymbol{u}^\intercal] = \frac{1}{d}\boldsymbol{I}$ yields

$$\begin{aligned}
&\mathbb{E}_{\boldsymbol{u}}[\mathrm{KL}(\boldsymbol{p}(\cdot \mid \boldsymbol{x}) \,\|\, \boldsymbol{p}(\cdot \mid \boldsymbol{x} + \varepsilon \boldsymbol{u}))] \\
&= \frac{1}{2}\,\varepsilon^2\, \mathbb{E}_{\boldsymbol{u}}[\boldsymbol{u}^\intercal \boldsymbol{F}_t\, \boldsymbol{u}] + o(\varepsilon^2) \\
&= \frac{1}{2}\,\varepsilon^2\, \mathrm{tr}\Big(\boldsymbol{F}_t\, \mathbb{E}_{\boldsymbol{u}}[\boldsymbol{u}\boldsymbol{u}^\intercal]\Big) + o(\varepsilon^2) \\
&= \frac{1}{2}\,\varepsilon^2\, \frac{1}{d}\, \mathrm{tr}(\boldsymbol{F}_t) + o(\varepsilon^2). \qquad (14)
\end{aligned}$$

Thus, up to the universal factor $\varepsilon^2/(2d)$, the trace $\mathrm{tr}(\boldsymbol{F}_t)$ governs the expected KL divergence induced by an isotropic perturbation of the input token $\boldsymbol{x}_t$. We therefore refer to $\mathrm{tr}(\boldsymbol{F}_t)$ as the *total information* at position t.

Fisher Scope as rank-1 approximation. Since $\boldsymbol{F}$ admits the eigen-decomposition $\boldsymbol{F} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^\intercal$, the total information can be written exactly as

$$\mathrm{tr}(\boldsymbol{F}_t) = \sum_{i=1}^{d_{\mathrm{model}}} \lambda_i \|\boldsymbol{u}_i^\intercal \boldsymbol{J}_t\|_2^2. \qquad (15)$$

In practice, $\boldsymbol{F}$ admits a good low-rank approximation $\boldsymbol{F} \approx \boldsymbol{U}_r \boldsymbol{\Lambda}_r \boldsymbol{U}_r^\intercal$, where $\boldsymbol{U}_r \in \mathbb{R}^{d_{\mathrm{model}} \times r}$ contains the top r eigenvectors and $\boldsymbol{\Lambda}_r$ the corresponding eigenvalues. This yields the approximation

$$\mathrm{tr}(\boldsymbol{F}_t) \approx \sum_{i=1}^{r} \lambda_i \|\boldsymbol{u}_i^\intercal \boldsymbol{J}_t\|_2^2, \qquad (16)$$

where each term $\|\boldsymbol{u}_i^\intercal \boldsymbol{J}_t\|_2$ requires only a single backward pass, analogous to Eq. (2). The full total information thus requires r backward passes, compared to d_{model} passes needed to form $\boldsymbol{J}_t$ explicitly.

The Fisher Scope influence score in Eq. (5) corresponds to the $r = 1$ case of Eq. (16), retaining only the leading term $\lambda_1 \|\boldsymbol{u}_1^\intercal \boldsymbol{J}_t\|_2^2$ and dropping the scalar λ_1 since it is constant across token positions t.

Low-rankness of $\boldsymbol{F}$. Figure 15 provides three complementary pieces of evidence that the $r = 1$ approximation is justified in practice. The top panel shows the eigenvalue spectrum of $\boldsymbol{F}$ across 10 LAMBADA prompts on a log-log scale: a consistent power-law decay spanning 6 orders of magnitude, indicating that the leading eigenvalue dominates all subsequent ones.

The middle panel confirms this quantitatively via the cumulative explained variance $\sum_{i=1}^{r} \lambda_i / \mathrm{tr}(\boldsymbol{F})$, averaged over 1000 LAMBADA passages: at $r = 1$ the mean explained fraction is already $\approx 37\%$, rising to $\approx 65\%$ by $r = 10$ and approaching 1 only beyond $r \approx 1000$.

The bottom panel provides direct empirical justification, extending the LLaMA-3.2 1B results in Table 3: rank-1 Fisher Scope ($r = 1$) already outperforms Input $\times$ Gradient and Semantic Scope on LAMBADA, and matches Temperature Scope. Increasing the rank to $r \in \{4, 16, 64, 256\}$ yields only marginal further gains in AOPC, while multiplying the computational cost by r.

Together, these results justify both the privileged status of $\boldsymbol{u}_1$ and the choice of $r = 1$ in Fisher Scope.

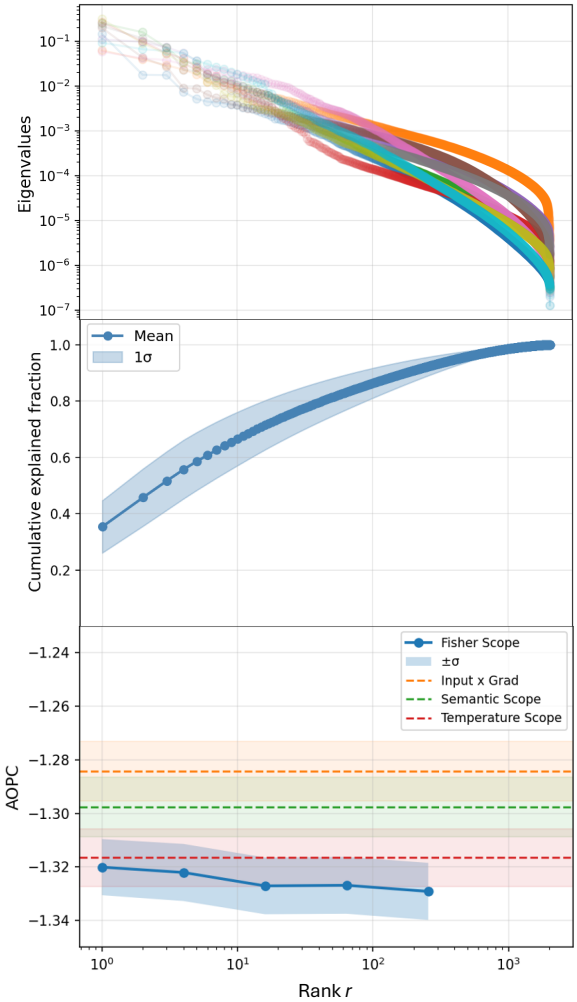

Figure 15: Justification for the rank-1 approximation used in Fisher Scope, evaluated on LLaMA-3.2 1B and LAMBADA. **Top:** Eigenvalue spectrum of $\boldsymbol{F}$ across 10 prompts on a log-log scale, exhibiting a consistent power-law decay spanning 6 orders of magnitude. **Middle:** Mean cumulative explained variance $\sum_{i=1}^{r} \lambda_i / \mathrm{tr}(\boldsymbol{F})$ averaged over 1000 passages (shaded band: $\pm 1\sigma$); $\boldsymbol{u}_1$ alone accounts for $\approx$37% of total sensitivity (1σ: 25–50%). **Bottom:** AOPC as a function of rank $r \in \{1, 4, 16, 64, 256\}$, with baselines shown as dashed lines. Rank-1 Fisher Scope already outperforms Input $\times$ Gradient and Semantic Scope; increasing r yields only marginal improvement at r-fold computational cost.