Surrogate Fidelity: When Can Open LLMs Explain Closed Ones?

Philippe Chlenski [1] **Zachariah Carmichael** [1] **Ayush Warikoo** [1] **Chia-Tse Shao** [1] **Yingxiao Ye** [1]
Aobo Yang [1] **Vivek Miglani** [1] **Nehal Bandi** [1]

Abstract

Mechanistic interpretability (MI) requires full access to model internals, yet the APIs for most widely deployed language models at best expose log-probabilities over output tokens. This creates a surrogate problem: when do measurements made on open models allow us to make claims about a closed model? We evaluate *surrogate fidelity* at the prediction, attribution, and representation levels. For binary classification tasks, log-odds provide an API-compatible scalar readout of the model's representation space, and leave-one-out attributions provide insight into model behavior. Across eleven models spanning four families (Llama, Qwen, GPT, and Gemini), we find that prediction fidelity substantially overstates attribution fidelity: models that agree on *what* the answer is often disagree on *why*. We document an *access–validity inversion:* white-box signals like attention patterns and perturbation magnitudes are highly stable across models but only weakly predictive of causal attributions, which black-box input ablations capture by design. Mechanistic insight does not automatically transfer to closed targets, and prediction-level agreement is insufficient to warrant such transfer. Code and results are available at https://github.com/facebookresearch/surrogate.

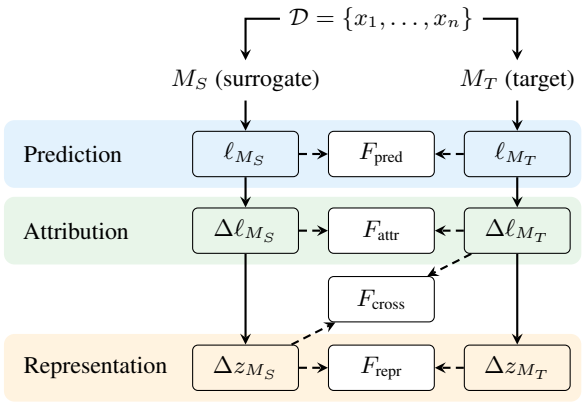


Figure 1. The surrogate fidelity evaluation pipeline. For each input corpus $\mathcal{D}$, we extract three signals from each model: prediction log-odds ℓ_M, ablation-based attributions $\Delta\ell_M$, and perturbation responses Δz. Prediction fidelity (F_{pred}) and attribution fidelity (F_{attr}) can be computed for any model pair, including closed-source targets accessible only via API. Representation fidelity (F_{repr}) and cross-level fidelity (F_{cross}) additionally require access to model internals.

1. Introduction

The models that most urgently need mechanistic interpretability (MI) are precisely those least amenable to it. Most classical MI tools, such as circuit analysis, sparse autoencoders, and activation patching, require full access to model internals. Meanwhile, most users of frontier large language models (LLMs) access text completions from behind APIs that, at best, offer log-probabilities for a handful of high-likelihood tokens. This creates a bifurcation: MI research is either conducted by frontier labs within the walled gardens of their own models, or by academic groups studying small open models and hoping their findings generalize. The broader community of practitioners, safety researchers, and auditors is left to wonder what can be learned about a model we cannot open from studying one we can.

This assumption that mechanistic insights transfer across models is implicit, pervasive, and under-examined in the literature. If it is wrong, then open-source MI work may be telling us less than we think about the frontier models we most need to understand. We argue that the field needs an open dialogue about cross-model fidelity, grounded in a shared convention for measurement that is realistic about the distinction between closed and open models.

We define a *surrogate fidelity metric* as a measure of how faithfully a measurement in a surrogate model predicts a measurement in a target model. A surrogate fidelity metric should be *appropriately scoped* (dataset- and claim-dependent), *discriminative* between different model pairs,

[1]Meta. Correspondence to: Philippe Chlenski <pac@fb.com>.

Mechanistic Interpretability Workshop at the 43rd International Conference on Machine Learning, Seoul, South Korea, 2026.
Copyright 2026 by the author(s).

and governed by the *principle of least privilege*, i.e. it should require no more access to the target model than the claim being evaluated demands. In particular, at least some surrogate fidelity measurements should be computable using top-K log-probabilities alone: a constraint consistent with the APIs of many frontier model providers.

To this end, we propose a hierarchy of increasingly strict tests as a concrete starting point for measuring surrogate fidelity, drawing on established MI intuitions and practices. *Prediction fidelity* for outputs; *attribution fidelity* for output-level responses to causal interventions (input ablations); *representation fidelity* for representation-level responses to the same interventions; and *cross-level fidelity* for predicting the target's attributions from the surrogate's internal representations. This progression from observational to interventional to mechanistic agreement echoes the causal hierarchy of Pearl (2009), grounding each level in a concrete, measurable test of surrogate faithfulness.

Our contributions. Our contributions are as follows:

1. We argue that the MI community's implicit assumption that mechanistic findings transfer across models deserves direct scrutiny, and propose *surrogate fidelity* as a conceptual framework for this dialogue.
2. We propose a hierarchy of surrogate fidelity metrics—prediction, attribution, representation, and cross-level—instantiated for binary classification benchmarks. Prediction and attribution require only top-K log-probabilities, while representation- and cross-level fidelity require open-weight model access.
3. Evaluating two open-weight families (Qwen 2.5, Llama 3) and two closed-API families (GPT, Gemini) across multiple scales (0.5B–70B parameters) on three binary classification benchmarks, we find a striking dissociation: models that share similar representations and predictions can nonetheless disagree sharply on causal attributions.
4. We identify an *access–validity inversion*: high-agreement white-box signals like attention and perturbation magnitude fail to predict more causally relevant black-box attributions.

2. Preliminaries

Throughout this section, all quantities are implicitly conditioned on a single input and its representation z in model M; we make these dependencies explicit in Section 3 when comparing across models and inputs.

2.1. Static geometry

We review the geometric relationship between a transformer's residual stream and its output probabilities, establishing the quantities on which our fidelity metrics operate. Given a representation $z \in \mathbb{R}^d$ in a transformer's residual stream, the logit for token t is the dot product

$$\text{logit}_t = z \cdot u_t, \tag{1}$$

where $u_t \in \mathbb{R}^d$ is the column of the unembedding matrix $W_U \in \mathbb{R}^{d \times V}$ corresponding to token t. The output probability is computed via softmax:

$$P(t) = \frac{\exp(z \cdot u_t)}{\sum_{i=1}^{V} \exp(z \cdot u_i)}. \tag{2}$$

Thus, the log-probability of any token is $\log P(t) = z \cdot u_t - C$, where $C = \log \sum_{s=1}^{V} \exp(z \cdot u_s)$ is the log of the softmax normalization constant. Without access to the full vocabulary of logits, C is not recoverable from a handful of top-K log-probabilities.

Since C is shared across all logits, when we take the *difference* of two log-probabilities, C cancels out:

$$\begin{aligned} \log P(t) - \log P(s) &= (z \cdot u_t - C) - (z \cdot u_s - C) \\ &= z \cdot (u_t - u_s). \end{aligned} \tag{3}$$

Defining the *readout (log-odds) direction* $v = u_+ - u_-$ for a binary classification task with positive and negative class tokens, the log-odds reduce to a single dot product:

$$\ell(x) = z \cdot v. \tag{4}$$

While individual log-probabilities are nonlinear functions of z, the log-odds are *exactly linear* in the representation. This is the key observation that makes our fidelity framework possible: log-odds are both API-accessible (as a difference of two log-probabilities) and geometrically interpretable (as a projection onto a readout direction in representation space). Figure 2(a–b) illustrates this progression from individual logits to the log-odds direction.

Residual stream decomposition. The residual stream of a transformer is a running sum: after layer l, the state is $z^{(l)} = z^{(0)} + \sum_{k=1}^{l}(a^{(k)} + m^{(k)})$, where $z^{(0)}$ is the token embedding and $a^{(k)}, m^{(k)} \in \mathbb{R}^d$ are the outputs of the attention and MLP sublayers at layer k. Because log-odds are linear in z, this additivity extends directly to the prediction: projecting any intermediate state $z^{(l)} \cdot v$ yields a valid log-odds estimate at depth l — the *logit lens* (nostalgebraist, 2020) — while decomposing the final projection term by term yields *direct logit attribution* (DLA) (Elhage et al., 2021). Writing out both relationships:

$$z^{(L)} = z^{(0)} + \sum_{l=1}^{L} \left(a^{(l)} + m^{(l)} \right) \tag{5}$$

$$\ell(x) = z^{(0)} \cdot v + \sum_{l=1}^{L} \left(a^{(l)} \cdot v + m^{(l)} \cdot v \right). \tag{6}$$

Prediction level:
Log-probs are nonlinear in z; log-odds are linear.

Attribution level:
Differences in projections are projections of differences.

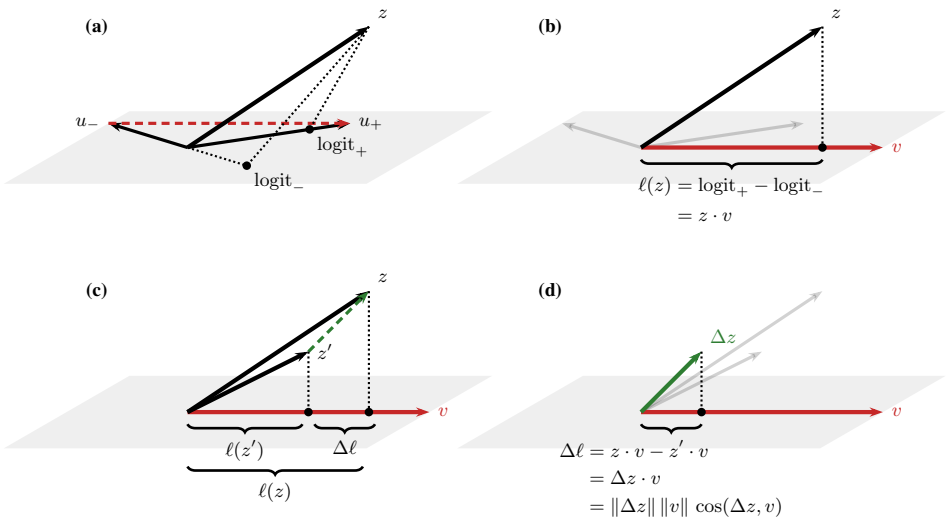


Figure 2. The geometric intuition behind our evaluations. **(a)** Logits are computed from a representation z by projection onto unembedding vectors u_- and u_+; however, these projections are not recoverable from the top-K log-probabilities exposed by most LLM inference providers. **(b)** The log-odds are equivalent to the difference in logits, which in turn equals the projection of z onto the difference vector $v = u_+ - u_-$. This quantity can be derived from log-probabilities alone. **(c)** Given a base representation z and a modified representation z', we define the attribution as the difference in log-odds derived from the two representations. This is equal to the difference in projection lengths. **(d)** The difference in projections can also be thought of as a projection of the difference, i.e. $\Delta z \cdot v$. Thus, for any Δz, the attribution depends on its magnitude, $\|\Delta z\|$, and its alignment with v, $\cos(\Delta z, v)$.

The first line is the residual stream; the second is the same equation projected onto v. The logit lens reads the partial sums $z^{(l)} \cdot v$ cumulatively across depth; DLA reads the individual terms $a^{(l)} \cdot v$ and $m^{(l)} \cdot v$ as per-component contributions to the final log-odds. We investigate the DLA of predictions and attributions in Section C.

2.2. Intervention and attribution

For an intervention[1] that changes the representation from z to z', the resulting change in log-odds is:

$$\Delta\ell = \ell(z) - \ell(z') = \Delta z \cdot v, \tag{7}$$

where $\Delta z = z - z'$ is the perturbation in the residual stream. We call $\Delta\ell$ the *attribution* of the intervention. Concretely, attribution measures the causal effect of changing the input on ℓ; it is not a complete internal mechanism.

Because attribution is a dot product, it decomposes into three factors:

$$\Delta\ell = \|\Delta z\| \, \|v\| \, \cos(\Delta z, v). \tag{8}$$

Since $\|v\|$ is fixed for a given model and task, attribution is governed by two properties of the perturbation: its *magnitude* $\|\Delta z\|$ and its *alignment* $\cos(\Delta z, v)$ with the log-odds direction. A large perturbation orthogonal to v produces zero attribution; a small perturbation aligned with v can shift the prediction substantially. Figure 2(c–d) illustrates this geometry.

Eqs. 7–8 are exact when z and z' are the post-normalization representations that the unembedding matrix sees.

The role of LayerNorm. In closed models, $\Delta\ell$ only reveals information about the post-normalization representations z and z'. In practice, however, interventions perturb the *pre-normalization* residual stream $\hat{z}$. For a normalization $z \to \hat{z} \,/\, f(\hat{z})$ that acts by scalar rescaling[2], the post-normalization perturbation decomposes further as:

$$\Delta z = \underbrace{\frac{\Delta\hat{z}}{f(\hat{z})}}_{\text{direct}} + \underbrace{\hat{z}' \, \frac{f(\hat{z}') - f(\hat{z})}{f(\hat{z}) \, f(\hat{z}')}}_{\text{indirect}}. \tag{9}$$

The *direct* term is the pre-normalization perturbation $\Delta\hat{z}$ rescaled by the baseline normalization factor, i.e. the contribution that would remain if $\hat{z}'$ was normalized by the original norm $\|\hat{z}\|$. The *indirect* term arises because changing $\hat{z} \to \hat{z}'$ also changes the normalization denominator, effectively rescaling the entire post-normalization representation z'. This term is nonzero whenever $f(\hat{z}) \neq f(\hat{z}')$ and $\hat{z}' \cdot v \neq 0$, even if $\Delta\hat{z}$ itself is orthogonal to v.

[1] Our framework is intervention-agnostic. For our experiments, we use prompt-level leave-one-out ablations because they are compatible with black-box API access.

[2] For RMSNorm (used in Qwen 2.5 and Llama 3), $f(\hat{z}) = \|\hat{z}\|/\sqrt{d}$.

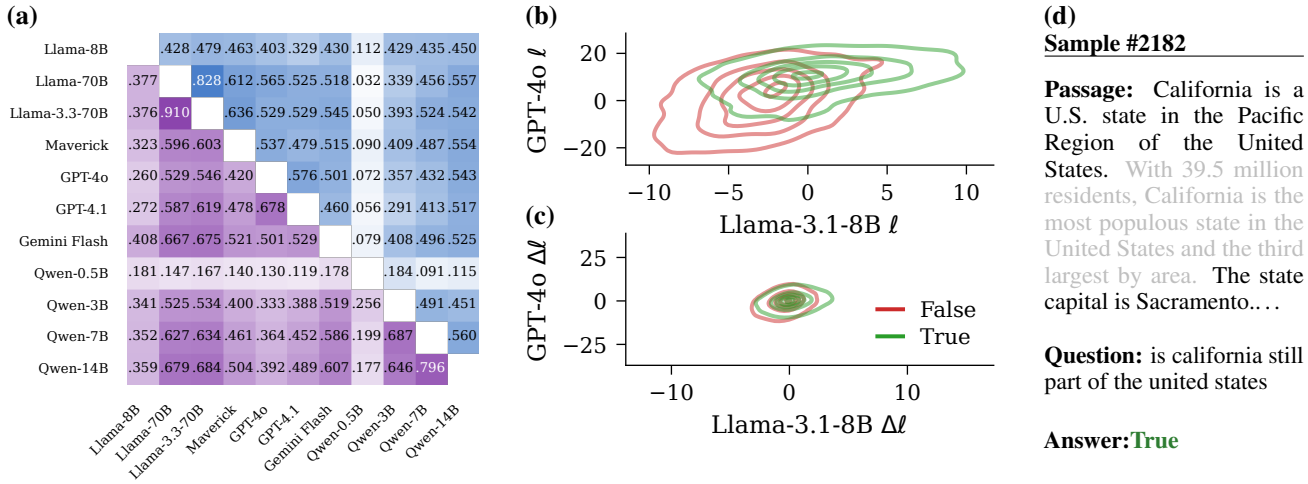


Figure 3. **(a)** Pairwise F_{pred} and F_{attr} heatmaps for eleven models across four families. **(b)** Log-odds contour plot for a representative cross-family pair (Llama 3-8B vs. GPT-4o), stratified by ground truth answer (**True** vs **False**). **(c)** Ablation contour plot for the same pair. **(d)** Example BoolQ prompt with a sentence-level ablation (greyed-out text).

3. Method

We formalize surrogate fidelity at three levels of increasing depth: do two models make similar *predictions*, produce similar *attributions* under the same perturbations, and share *internal representations*? Each level imposes a stricter test of whether mechanistic insights transfer across model families.

Prediction and attribution fidelity require only log-probabilities and are therefore computable for any model pair, including closed targets; representation fidelity additionally requires access to model internals.

Let M_S and M_T denote surrogate and target models, and $\mathcal{D} = \{x_1, \ldots, x_n\}$ an evaluation corpus. For each model M, the quantities defined in Section 2 — log-odds $\ell_M(x)$, attributions $\Delta\ell_M(x, i)$, perturbation norms $\|\Delta z_M(x, i)\|$, and alignments $\gamma_M(x, i) = \cos(\Delta z_M, v_M)$ — are stacked over all prompts and segments to form vectors (denoted in boldface). We additionally write $\alpha_M(x, i)$ for attention-based importance scores. Throughout, agreement is measured by r^2 with bootstrapped 95% confidence intervals.

Prediction fidelity measures agreement on outputs:

$$F_{\text{pred}}(M_S, M_T; \mathcal{D}) = r^2(\boldsymbol{\ell}_{M_S},\ \boldsymbol{\ell}_{M_T})\,. \tag{10}$$

Attribution fidelity measures agreement on causal importance under leave-one-out ablation:

$$F_{\text{attr}}(M_S, M_T; \mathcal{D}) = r^2(\boldsymbol{\Delta\ell}_{M_S},\ \boldsymbol{\Delta\ell}_{M_T})\,. \tag{11}$$

High F_{pred} with low F_{attr} signals that models agree on outputs while disagreeing on which inputs drive them, a critical failure mode for surrogate-based interpretability.

For open-weight models, we can additionally define representation-level metrics, referred to generically as F_{repr}. The decomposition in Equation (8) motivates two representation fidelity metrics measuring agreement in perturbation magnitude and alignment respectively:

$$F_{\text{mag}}(M_S, M_T; \mathcal{D}) = r^2(\|\boldsymbol{\Delta z}_{M_S}\|,\ \|\boldsymbol{\Delta z}_{M_T}\|)\,, \tag{12}$$

$$F_{\text{align}}(M_S, M_T; \mathcal{D}) = r^2(\boldsymbol{\gamma}_{M_S},\ \boldsymbol{\gamma}_{M_T})\,. \tag{13}$$

Attention patterns have also been used as a proxy for mechanistic importance (Abnar & Zuidema, 2020); we include *attention fidelity* as a third representation-level metric:

$$F_{\text{attn}}(M_S, M_T; \mathcal{D}) = r^2(\boldsymbol{\alpha}_{M_S},\ \boldsymbol{\alpha}_{M_T})\,. \tag{14}$$

For attention, we consider several attention aggregation strategies (mean, max, rollout; see Section 4).

Finally, *cross-level fidelity* asks whether the surrogate's internals predict the target's attributions:

$$F_{\text{cross}}(M_S, M_T; \mathcal{D}) = r^2(\boldsymbol{\alpha}_{M_S},\ \boldsymbol{\Delta\ell}_{M_T})\,. \tag{15}$$

For closed M_T, F_{cross} quantifies whether mechanistic analysis of an open M_S yields actionable insight about M_T.

Readout-compatible linear transfer. Our use of Pearson r^2 is motivated by a readout-compatible notion of representation transfer. Concretely, for idealized models M_S, M_T related through a readout-compatible linear transformation, i.e. $z_T \approx Az_S$ and $A^T v_T \approx cv_S$,

$$z_T{\cdot}v_T \approx Az_S{\cdot}v_T = z_S{\cdot}(A^T v_T) \approx c(z_S{\cdot}v_S) = c\ell_S, \tag{16}$$

i.e. the log-odds of M_S and M_T are in a perfect linear relation. The same holds for $\Delta\ell$, yielding $\Delta\ell_T \approx c\Delta\ell_S$.

Thus, r^2 measures the fraction of variation explained by M_S under the simplest readout-compatible linear calibration. Low r^2 in a single dimension rules out high-fidelity transfer under such a readout-compatible A, though it leaves open weaker forms of representational similarity.

Analogous correlation-based measures exist in higher dimensions: the RV coefficient (Robert & Escoufier, 1976) for multivariate readouts and centered kernel alignment (CKA) (Kornblith et al., 2019) for full representations. This lineage grounds our use of r^2 as a readout-level probe into overall representational similarity and motivates our multi-class extension in Section H.

4. Evaluation Setting

This is a summary of our experimental setup; full details can be found in the Appendix, Section A.

4.1. Benchmarks

We focus our initial investigation on binary classification tasks: when a model's output space is $\{y^+, y^-\}$, the log-odds $\ell_M(x)$ collapse to a single scalar per prompt, providing a compact and interpretable summary of the model's internal decision boundary.

Binary classification is practical under API constraints: the top-K most likely tokens are typically constrained to a small, predictable set of label variants. In more open-ended settings, we expect significantly more missing values.

We evaluate on three binary classification benchmarks spanning distinct reasoning capabilities:

- **BoolQ** (Clark et al., 2019): Boolean question answering over Wikipedia passages (validation split; n=3,270). Tests reading comprehension with yes/no questions.
- **ANLI** (Nie et al., 2020): Adversarial natural language inference (test splits R1/R2: n=1,000 each, R3: n=1,200). Progressively harder rounds constructed to fool strong models, providing a difficulty gradient.
- **WinoGrande** (Sakaguchi et al., 2021): Debiased coreference resolution (validation split; n=1,267). Binary choice between candidate referents. Also serves as our primary testbed for ablation experiments.

Beyond our core binary classification experiments, we also extend our evaluation framework to two other benchmarks:

- **LAMBADA** (Paperno et al., 2016): Word prediction requiring broad discourse context (test split; n=5,153). We score completion log-probability rather than log-odds, providing a robustness check across scoring methods.
- **RACE** (Lai et al., 2017): Multiple-choice reading comprehension (test split; n=4,934, 4 choices per question). We report RV coefficients on pairwise log-odds.

4.2. Models

We compare models across four families and multiple scales:

- **Llama** (Touvron et al., 2023): Llama 3.1-8B, Llama 3.1-70B, Llama 3.3-70B, Llama 4 Maverick
- **Qwen** (Bai et al., 2023): Qwen-0.5B, Qwen-3B, Qwen-7B, Qwen-14B
- **GPT** (Achiam et al., 2023): GPT-4o, GPT-4.1
- **Gemini** (Comanici et al., 2025): Gemini 2.5 Flash Lite

The Llama and Qwen families provide multiscale open-weight models, enabling representation-level extraction (for F_{repr} and F_{cross}) and investigation into how surrogate fidelity scales with model size. GPT and Gemini models are accessible only through black-box APIs exposing top-K log-probabilities, restricting us to F_{pred} and F_{attr} for these targets. Notably, the Claude and GPT-5 families, which do not expose any log-probabilities to end users, are excluded from this evaluation.

4.3. Ablation protocol

For attribution fidelity, we segment each prompt and remove individual segments to produce ablated inputs $\tilde{x}_i$. We evaluate at two granularities:

- **Sentence-level**: Each sentence in the prompt constitutes a segment. Coarser but less expensive, as each prompt yields few ablations.
- **Word-level**: Each whitespace-delimited token is a segment. Finer-grained but requires substantially more model queries per prompt.

Sentence-level ablations are applied to all benchmarks; word-level ablations are applied to a downsampled subset (n=10,000) of each benchmark due to cost constraints.

5. Results

We first evaluate prediction and sentence-level attribution fidelity on the BoolQ validation set across all eleven models. We report pairwise r^2 values in Table 1 and show the full pairwise fidelities and cross-family contour plots for F_{pred} and F_{attr} in Figure 3, as well as an example BoolQ prompt.

5.1. Prediction fidelity

Prediction fidelity is high within families but drops substantially across families. Predictably, Llama 3.1-70B and Llama 3.3-70B have high mutual F_{pred} (0.910) and exhibit similar F_{pred} profiles generally. The same is true for Qwen-7B and Qwen-14B (0.796) and GPT-4o and GPT-4.1 (0.678). By contrast, the highest cross-family F_{pred} is between Llama 3.1-70B and Qwen 2.5-14B at 0.684; cross-family ranges are generally slightly lower.

Table 1. Cross-model agreement (Pearson r^2) on BoolQ sentence-level (n = 27,516 segments / 3,270 prompts). Each cell is min/**median**/max over the relevant model-pair set. F_{xxx} measures agreement on a single signal; F_{cross} measures how well a signal from model M_S predicts ablation response in model M_T. Open→Open = within the 5 open instruct models; Open→Closed = open×closed-source pairs; All→All = all 11 models. Cells marked — cannot be filled because representation-level signal is missing for closed models.

	Metric	Open→Open r^2	Open→Closed r^2	All→All r^2
Black-box	F_{pred}	.177 / **.652** / .795	.108 / **.682** / .845	.108 / **.709** / .956
	F_{attr}	.091 / **.432** / .560	.032 / **.420** / .557	.032 / **.460** / .828
Representation-level	$F_{\text{attn}}^{\text{mean}}$	.762 / **.905** / .984	—	—
	F_{mag}	.641 / **.855** / .903	—	—
	$F_{\text{attn}}^{\text{rollout}}$	.443 / **.848** / .938	—	—
	$F_{\text{attn}}^{\text{max}}$	.641 / **.842** / .910	—	—
	F_{align}	.058 / **.200** / .294	—	—
Cross-level	$F_{\text{align}\to\text{attr}}$	.060 / **.259** / .381	.032 / **.262** / .362	.032 / **.259** / .381
	$F_{\text{mag}\to\text{attr}}$	.000 / **.074** / .229	.000 / **.122** / .282	.000 / **.109** / .282
	$F_{\text{attn}\to\text{attr}}^{\text{mean}}$	.000 / **.004** / .035	.000 / **.003** / .024	.000 / **.003** / .035
	$F_{\text{attn}\to\text{attr}}^{\text{rollout}}$	.000 / **.003** / .069	.000 / **.002** / .013	.000 / **.002** / .069
	$F_{\text{attn}\to\text{attr}}^{\text{max}}$	.000 / **.002** / .032	.000 / **.002** / .019	.000 / **.002** / .032

Notably, Qwen-0.5B is an outlier on all fronts, with low prediction fidelity to any models including larger models in the Qwen family. This is consistent with its uniquely small size among the evaluated models, and is contextualized by the generally low confidence and poor log-probability separation reported in Figure 8 in the Appendix.

5.2. Attribution fidelity

For each model and each BoolQ prompt, we computed sentence-level ablations and measured the attribution fidelity across model pairs.

At this level, attribution agreement is strikingly low across all model pairs, even within-family. The median within-family F_{attr} is only 0.432, and even the closest pair (Qwen-7B vs Qwen-14B) only achieves a F_{attr} score of 0.560.

Unlike F_{pred}, where r^2 closely tracks Spearman ρ, r^2 values for F_{attr} are slightly higher than Spearman ρ (see Figure 9 in Appendix), as they are significantly more stable in the presence of near-zero attributions.

Word-level validation. To rule out that the low ablation agreement is an artifact of sentence-level coarseness, we repeat the analysis at word level, using leave-one-out *single deletion* (Nauta et al., 2023) of individual whitespace tokens. We subsample 10,000 (prompt, word) ablation pairs per model from a pool of 476,154 possible BoolQ word ablations using a shared random seed.

Per Table 2, switching from sentence- to word-level ablation leaves mean and rollout attention fidelity mostly unchanged but sharply reduces F_{attr} (from .460 to .175), driven almost entirely by the collapse of F_{align} (.200 → .012). The attribution signal fractures at finer resolution; the representational signals do not.

5.3. Representational fidelity

In this section, we focus both on decomposing F_{attr} according to the components of attribution (norm of perturbation, angle between perturbation and readout direction, and LayerNorm-induced rescaling) and investigating the role attention could play in surrogate model predictions.

Decomposing attributions. Figure 4 shows that models' responses to perturbations are stable in magnitude (high F_{mag}) but not direction (low F_{align}). This suggests that low F_{attr} is driven primarily by disagreement in the direction of the perturbation-response vector Δz. The maximum cosine similarity between Δz and v tops out around 0.4, with most cosine similarities landing near zero. Using unsigned cosine similarity does not improve F_{align}.

While most perturbations have negligible effects on log-odds from the perspective of the projection of Δz onto z, it is also possible for perturbations to affect a model's log-odds through the LayerNorm's rescaling of the overall vector, as rescaling a vector by α also rescales ℓ by α.

We find that roughly 15–23% of each attribution is due to LayerNorm rescaling, with the rest explainable by the unnormalized change in representation. Interestingly, the direction of LayerNorm rescaling is essentially uncorrelated with attribution, meaning that while this component of the model architecture can contribute to lower F_{attr} scores, it is unlikely to bias it in a specific direction. For models ≥3B parameters, there is no clear trend in terms of the LayerNorm's contribution to $\Delta\ell$, with Llama 3-8B sitting in the middle of the Qwen results across all metrics.

Attention: stable but limited. In contrast to ablation attributions, attention-based scores exhibit strong agreement across the board. Mean attention correlations range from

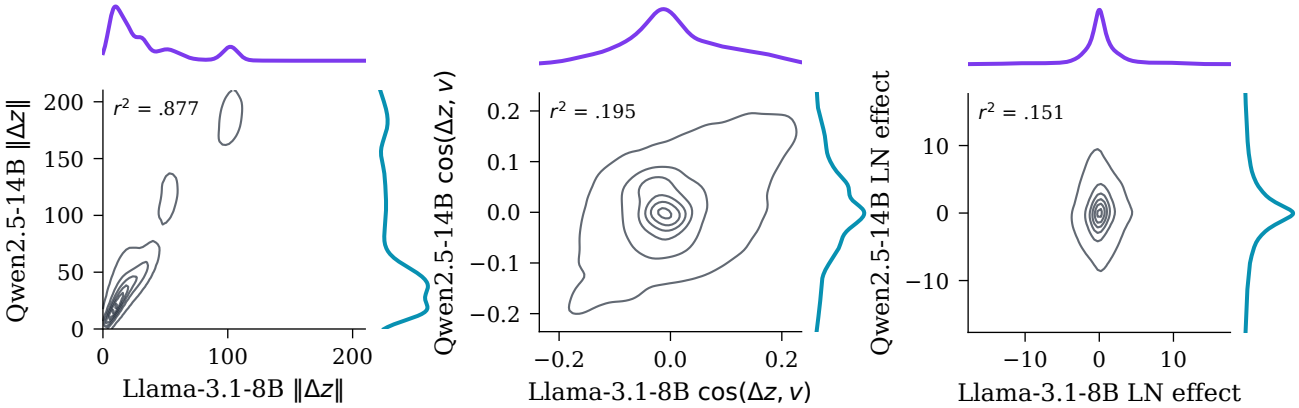


Figure 4. Perturbations affect attribution through their norm (**left**), alignment with the log-odds direction v (**middle**), and indirectly via LayerNorm (**right**). We plot joint and marginal distributions of each, for a representative model pair.

$r^2 = 0.762$ to 0.984, and attention rollout, which accounts for residual connections across layers (Abnar & Zuidema, 2020), achieves median $r^2 = 0.848$. Even Qwen-0.5B, which was an outlier for prediction and attribution fidelity, attends to the same tokens as its larger siblings.

Despite high within-method consistency, the three attention aggregation strategies are not interchangeable. Mean and max attention correlate moderately ($r^2 \approx 0.473$–0.969), but rollout attention is almost uncorrelated with these ($r^2 \approx 0.045$–0.215). This methodological sensitivity underscores the importance of specifying the aggregation scheme when reporting attention-based results.

The contrast between high representational fidelity and low attribution fidelity reveals a dissociation: models can share attention patterns while diverging in causal structure. Attention consistency (often used as informal evidence that "models are doing the same thing") is a weak proxy for the mechanistic agreement that MI practitioners actually need.

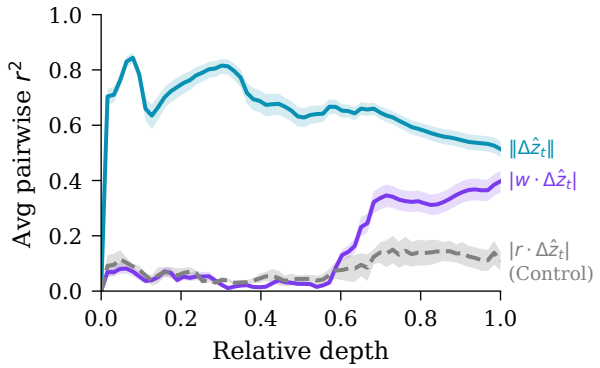


Figure 5. Bootstrap-CI pair-averaged per-layer $F^{(l)}_{\text{mag}}$ and $F^{(l)}_{\text{attr}}$ versus relative depth, with a random-direction control overlay (gray dashed). Both $F^{(l)}_{\text{attr}}$ and the control rise late.

Per-layer fidelity. We track the per-layer variants of $F^{(l)}_{\text{mag}}$ and $F^{(l)}_{\text{attr}}$ of the metrics defined in Equations (11) and (12) in Figure 5. A clean *depth dissociation* emerges: high representation-level agreement across most depths, but low attribution-level agreement for the first $\sim$55% of depth with a subsequent sharp rise. This pattern could be a generic property of late-layer perturbations, so we include a random direction control—replacing v with a per-model random unit direction r matched in norm to $\|v\|$. The random-direction control rises late as well—the v-specific premium over the random baseline at the readout layer is therefore $\approx$0.29 in r^2. We read this as a partial qualification of our claim that low F_{attr} reflects models "disagreeing on which inputs drive their predictions." See Section C for extended experiments, analysis, and experiment setup.

5.4. Additional experiments

We report our results for the ANLI, WinoGrande, LAMBADA, and RACE benchmarks in Table 2.

Across binary classification benchmarks, the qualitative ordering is stable: F_{pred} generally exceeds F_{attr}, attention is stable, and F_{mag} is more stable than F_{align}. Surprisingly, we do not find a monotonic relationship between ANLI difficulty (R1 < R2 < R3) and most fidelity statistics, suggesting that fidelity measurements may be insensitive to task difficulty. WinoGrande fidelity is lower across F_{pred} and F_{attr}, and shows higher F_{cross} for mean- and max-pooled attention. LAMBADA results show unusually low F_{pred} and F_{attr}, suggesting that the change in last-token log-probabilities may behave differently from log-odds.

While r^2 and RV scores are not directly comparable, the RACE results mostly resemble binary classification. Interestingly, F_{attr} is modestly larger than F_{pred} (0.524 vs 0.462), consistent with Figure 25, which shows that perturbation RV scores *exceed* representation RV scores in high

Table 2. Median cross-benchmark agreement on all benchmarks: we show r^2 for all benchmarks except RACE, a multiple-choice dataset which therefore requires an RV coefficient. It is generally expected that RV coefficients are higher than r^2, but the ordering of RV coefficients is meaningful. The first group of rows aggregates over open and closed models; the second group is only within open models; and in the final group, open models predict all models' attributions.

Metric	BoolQ	ANLI R1	ANLI R2	ANLI R3	WinoGrande	BoolQ word	LAMBADA	RACE
F_{pred}	.709	.354	.282	.343	.225	.708	.108	.462
F_{attr}	.460	.280	.223	.241	.150	.175	.115	.524
$F_{\text{attn}}^{\text{rollout}}$	.848	.852	.853	.854	.850	.857	.690	.811
$F_{\text{attn}}^{\text{mean}}$	.905	.953	.951	.954	.871	.881	.561	.913
$F_{\text{attn}}^{\text{max}}$	.842	.939	.937	.937	.808	.627	.332	.842
F_{mag}	.855	.867	.875	.874	.671	.524	.342	.771
F_{align}	.200	.078	.071	.060	.018	.012	.024	.100
$F_{\text{align}\to\text{attr}}$	.259	.128	.113	.114	.051	.026	.040	.183
$F_{\text{mag}\to\text{attr}}$	.109	.297	.268	.305	.228	.173	.159	.319
$F_{\text{attn}\to\text{attr}}^{\text{rollout}}$	.002	.021	.016	.011	.017	.004	.007	.014
$F_{\text{attn}\to\text{attr}}^{\text{mean}}$	.003	.013	.013	.022	.018	.003	.099	.036
$F_{\text{attn}\to\text{attr}}^{\text{max}}$	.002	.013	.016	.010	.022	.002	.069	.033

dimensions. These results may be confounded by longer prompts and the larger spread of pairwise log-odds.

In the Appendix, Section B contains expanded visualizations and calibration results, Section C contains full per-layer logit lens results, Section D demonstrates confidence-dependence in fidelity measurements, Section E proposes normalized root mean squared error (NRMSE) as an asymmetric variant of r^2, Section F elaborates on the mechanistic causes of the prediction–attribution gap, Section G studies the effect of additionally modifying models' system prompts, Section H extends our analysis to multiple-choice benchmarks, and Section I makes the connection to CKA explicit.

6. Discussion

The access-validity inversion. While F_{pred} measures prediction-level agreement, attribution fidelity is causal: leave-one-out ablations amount to a coarse-grained gradient of the prediction with respect to context, so mechanistic agreement requires this stricter condition beyond basic prediction agreement. For all model pairs not including Qwen 2.5–0.5B, $F_{\text{pred}} > F_{\text{attr}}$, and F_{attr} is generally very low.

Furthermore, we find most representation-level properties to have limited ability to predict attributions. API-compatible attribution metrics are more causally relevant than some white-box signals. In fact, median F_{repr} and F_{cross} are very poorly aligned (see Figure 9), implying representation-level transfer is largely uncorrelated with causal prediction. These negative results on attention add to the growing literature reexamining its explanatory value (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Bibal et al., 2022).

A plausible explanation for the access-validity inversion is that training forces models to align representations *in high-probability subspaces*, but the null space of the top-K readouts is less constrained. Consequently, models may agree on their overall predictions and even gross structural response (F_{mag}) while pointing in very different directions. Interestingly, this divergence is detectable with only API-level access: you can't see *how* these representations differ, but you can see *that* they do. (Since we observe that log-probs diverge between true and false prompts only in instruction fine-tuned models (see Section B.2), it's even possible that the access-validity inversion is posttraining-specific.)

Surrogate fidelity as local linear geometry. The Platonic Representation Hypothesis (Huh et al., 2024) holds that models converge to a shared representation, citing evidence such as model stitching, which unifies two models' representations via a learned linear transformation (similar to the A discussed in Equation (16)). Under this framing, our results are ironic: while *appearances* do indeed converge (high F_{pred}), the *Forms* themselves actually fan out across different models (low F_{align}).

The polytope lens relaxation (Black et al., 2022) affords models freedom to shuffle operations within closed consistency regions, so low surrogate fidelity need not imply deep representational divergence. This more closely matches how Huh et al. (2024) actually demonstrate their claims, eschewing the strict orthogonal transformation assumptions underpinning CKA in favor of nearest-neighbor overlaps.

Practical guidance. Our results imply that surrogate validity must be measured rather than assumed, and suggest a concrete way for practitioners to make surrogate validity claims. Prediction fidelity is a useful screen, but it consistently overstates agreement at the level needed for explanation: on BoolQ with word-level ablations, $F_{\text{pred}} = 0.708$ whereas $F_{\text{attr}} = 0.175$. Representational similarities can

also be misleading. For instance, F_{mag} and F_{attn} are high across model pairs, suggesting high similarity more generally, but directional alignment breaks the pattern: F_{align} is consistently low, driving down F_{attr} with it.

The appropriate fidelity test depends on the level of access and the claim being made. Both F_{pred} and F_{attr} can be measured using log-probabilities alone, and should be reported whenever comparing models to make claims about surrogate explanations. For directly testing open-to-closed mechanism transfer, F_{cross} requires open-weight access to the surrogate. In our evaluation, no surrogate-side representation signal clears this bar. Validating at the level of the explanation being made and the behavior being tested will help ground claims about generalization to frontier models. Note that the intervention can and should be chosen to match the behavior of interest: for instance, leave-one-out ablations are useful for claims about input evidence or saliency, but may not be appropriate for circuit-level claims.

Limitations and future work. Our framework is a first step towards standardized surrogate fidelity evaluation; we expect practices to evolve with input from the MI community.

We focus on binary classification tasks where log-odds give a reliable readout, with exploratory extensions into last-token and multiple-choice settings. An extension to open-ended text generation, which represents the vast majority of LLM use cases and does not admit an obvious scalar summary, is a natural next step requiring new scoring functions.

Our focus on leave-one-out attribution trades mechanistic depth for API compatibility; moreover, ablations shifting inputs off-distribution may have unintended side effects. It is therefore natural to extend our work to more commonly used *representation-level* interventions like activation patching, richer mechanistic signals (steering vectors, SAE features), and meaning-preserving input perturbations (Romanou et al., 2026). Low cross-seed consistency between SAE features (Chanin et al., 2024; Leask et al., 2025) might imply similarly low representation fidelity.

7. Related Work

Feature attribution for LLMs. Leave-one-out attribution (Koh & Liang, 2017), ContextCite (Cohen-Wang et al., 2024), and local surrogate methods such as MExGen (Paes et al., 2025) and gSmILE (Dehghani et al., 2025) estimate input-level importance for generative models. AttriBoT (Liu et al., 2025) scales attribution to long contexts via efficient approximation. These methods only explain *what* a model's prediction depends on; we connect attributions to internal mechanism by exploiting the linearity of log-odds. In this regard, Carlini et al. (2024) is also influential to our work in its use of log-probabilities to probe model mechanisms.

Proxy and surrogate models. Prior work asks whether proxy models can cheaply approximate a target's predictions (Hinton et al., 2015) or attributions (Ribeiro et al., 2016; Koh et al., 2026), often by training or fine-tuning a surrogate. On the other hand, we ask when a surrogate-derived explanation is valid evidence about a target model's *mechanism*.

Mechanistic interpretability. Circuit analysis (Olah et al., 2020; Elhage et al., 2021) has been scaled (Wang et al., 2022; Conmy et al., 2023) and automated (Goldowsky-Dill et al., 2023); sparse autoencoders offer a complementary feature-level decomposition (Bricken et al., 2023; Cunningham et al., 2023; Templeton et al., 2024). Causally relevant directions—steering vectors (Turner et al., 2023; Rimsky et al., 2024; Li et al., 2023), linear probes (Belinkov, 2022; Hewitt & Manning, 2019), crosscoder latents (Dunefsky et al., 2024; Lindsey et al., 2024; Ameisen et al., 2025)—identify features that influence behavior but require full model access. We relax this requirement and test whether surrogate-derived features transfer to closed targets.

Attention as explanation. Attention weights are widely used as explanations but their validity is contested (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019; Serrano & Smith, 2019; Bastings & Filippova, 2020; Bibal et al., 2022). Our F_{cross} metric provides direct evidence: attention patterns are highly consistent across scales yet uncorrelated with causal attributions, so attention stability is not evidence of mechanistic agreement.

Representation similarity. CKA (Kornblith et al., 2019), SVCCA (Raghu et al., 2017), and RSA (Kriegeskorte et al., 2008) compare hidden representations across models. The linear representation hypothesis (Park et al., 2024) and polytope lens (Black et al., 2022) characterize how models encode concepts geometrically. Our results partially agree: attention converges across scales, but its dissociation from attribution fidelity shows representational convergence does not imply functional equivalence.

8. Conclusion

We introduce a hierarchy of surrogate fidelity metrics (prediction, attribution, and representational) grounded in the distinction between observational, causal, and structural agreement. We find that neither prediction– nor representation-level fidelity implies attribution fidelity. When interpreting open surrogates as proxies for closed systems, we recommend that practitioners validate surrogates at the level of analysis they intend to perform, and release our evaluation framework to support this practice. More broadly, we offer an opinionated first proposal and invite the MI community to refine it: cross-model fidelity deserves to be discussed explicitly, and the right framework to measure it may still need to evolve.

References

Abnar, S. and Zuidema, W. Quantifying attention flow in transformers. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4190–4197, 2020.

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Agrawal, L. A., Tan, S., Soylu, D., Ziems, N., Khare, R., Opsahl-Ong, K., Singhvi, A., Shandilya, H., Ryan, M. J., Jiang, M., Potts, C., Sen, K., Dimakis, A. G., Stoica, I., Klein, D., Zaharia, M., and Khattab, O. Gepa: Reflective prompt evolution can outperform reinforcement learning, 2026. URL https://arxiv.org/abs/2507.19457.

Ameisen, E., Lindsey, J., Pearce, A., Gurnee, W., Turner, N. L., Chen, B., Citro, C., Abrahams, D., Carter, S., Hosmer, B., Marcus, J., Sklar, M., Templeton, A., Bricken, T., McDougall, C., Cunningham, H., Henighan, T., Jermyn, A., Jones, A., Persic, A., Qi, Z., Ben Thompson, T., Zimmerman, S., Rivoire, K., Conerly, T., Olah, C., and Batson, J. Circuit tracing: Revealing computational graphs in language models. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/methods.html.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bastings, J. and Filippova, K. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 149–155, 2020.

Belinkov, Y. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022.

Belrose, N., Ostrovsky, I., McKinney, L., Furman, Z., Smith, L., Halawi, D., Biderman, S., and Steinhardt, J. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023.

Bibal, A., Cardon, R., Alfter, D., Wilkens, R., Wang, X., François, T., and Watrin, P. Is attention explanation? an introduction to the debate. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, pp. 3889–3900, 2022.

Black, S., Sharkey, L., Grinsztajn, L., Winsor, E., Braun, D., Merizian, J., Parker, K., Guevara, C. R., Millidge, B., Alfour, G., et al. Interpreting neural networks through the polytope lens. *arXiv preprint arXiv:2211.12312*, 2022.

Bricken, T., Templeton, A., Batson, J., Chen, B., Jermyn, A., Conerly, T., Turner, N., Anil, C., Denison, C., Askell, A., et al. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2(5):6, 2023.

Carlini, N., Paleka, D., Dvijotham, K. D., Steinke, T., Hayase, J., Cooper, A. F., Lee, K., Jagielski, M., Nasr, M., Conmy, A., et al. Stealing part of a production language model. *arXiv preprint arXiv:2403.06634*, 2024.

Chanin, D., Wilken-Smith, J., Dulka, T., Bhatnagar, H., Golechha, S., and Bloom, J. A is for absorption: Studying feature splitting and absorption in sparse autoencoders. *arXiv preprint arXiv:2409.14507*, 2024.

Clark, C., Lee, K., Chang, M.-W., Kwiatkowski, T., Collins, M., and Toutanova, K. Boolq: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)*, pp. 2924–2936, 2019.

Cohen-Wang, B., Shah, H., Georgiev, K., and Madry, A. Contextcite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37: 95764–95807, 2024.

Comanici, G., Bieber, E., Schaekermann, M., Pasupat, I., Sachdeva, N., Dhillon, I., Blistein, M., Ram, O., Zhang, D., Rosen, E., et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Conmy, A., Mavor-Parker, A., Lynch, A., Heimersheim, S., and Garriga-Alonso, A. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352, 2023.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

Dehghani, Z., Akram, M. N., Aslansefat, K., Khan, A., and Papadopoulos, Y. Explaining large language models with gsmile, 2025. URL https://arxiv.org/abs/2505.21657.

Dunefsky, J., Chlenski, P., and Nanda, N. Transcoders find interpretable llm feature circuits. *Advances in Neural Information Processing Systems*, 37:24375–24410, 2024.

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Goldowsky-Dill, N., MacLeod, C., Sato, L., and Arora, A. Localizing model behavior with path patching. *arXiv preprint arXiv:2304.05969*, 2023.

Hewitt, J. and Manning, C. D. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, 2019.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

Jain, S. and Wallace, B. C. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3543–3556, 2019.

Johnson, W. B. and Lindenstrauss, J. Extensions of lipschitz mappings into hilbert space. *Contemporary mathematics*, 26:189–206, 1984. URL https://api.semanticscholar.org/CorpusID:117819162.

Koh, P. W. and Liang, P. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pp. 1885–1894. PMLR, 2017.

Koh, W., Suk, J., Han, S., Yun, S.-Y., and Shin, J. Predicting llm reasoning performance with small proxy model, 2026. URL https://arxiv.org/abs/2509.21013.

Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al. Captum: A unified and generic model interpretability library for pytorch. *arXiv preprint arXiv:2009.07896*, 2020.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMLR, 2019.

Kriegeskorte, N., Mur, M., and Bandettini, P. A. Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, 2:249, 2008.

Lai, G., Xie, Q., Liu, H., Yang, Y., and Hovy, E. RACE: Large-scale ReAding comprehension dataset from examinations. In Palmer, M., Hwa, R., and Riedel, S. (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 785–794, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1082. URL https://aclanthology.org/D17-1082/.

Leask, P., Bussmann, B., Pearce, M., Bloom, J., Tigges, C., Moubayed, N. A., Sharkey, L., and Nanda, N. Sparse autoencoders do not find canonical units of analysis. *arXiv preprint arXiv:2502.04878*, 2025.

Li, K., Patel, O., Viégas, F., Pfister, H., and Wattenberg, M. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in Neural Information Processing Systems*, 36:41451–41530, 2023.

Lindsey, J., Templeton, A., Marcus, J., Conerly, T., Batson, J., and Olah, C. Sparse crosscoders for cross-layer features and model diffing. *Transformer Circuits Thread*, pp. 3982–3992, 2024.

Liu, F., Kandpal, N., and Raffel, C. Attribot: A bag of tricks for efficiently approximating leave-one-out context attribution, 2025. URL https://arxiv.org/abs/2411.15102.

Nauta, M., Trienes, J., Pathak, S., Nguyen, E., Peters, M., Schmitt, Y., Schlötterer, J., van Keulen, M., and Seifert, C. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, July 2023. ISSN 1557-7341. doi: 10.1145/3583558. URL http://dx.doi.org/10.1145/3583558.

Nie, Y., Williams, A., Dinan, E., Bansal, M., Weston, J., and Kiela, D. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 4885–4901, 2020.

nostalgebraist. Interpreting GPT: The Logit Lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens. Accessed: 2026-04-23.

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., and Carter, S. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001, 2020.

Paes, L. M., Wei, D., Do, H. J., Strobelt, H., Luss, R., Dhurandhar, A., Nagireddy, M., Ramamurthy, K. N.,

Sattigeri, P., Geyer, W., and Ghosh, S. Multi-level explanations for generative language models, 2025. URL https://arxiv.org/abs/2403.14459.

Paperno, D., Kruszewski, G., Lazaridou, A., Pham, Q. N., Bernardi, R., Pezzelle, S., Baroni, M., Boleda, G., and Fernández, R. The lambada dataset: Word prediction requiring a broad discourse context, 2016. URL https://arxiv.org/abs/1606.06031.

Park, K., Choe, Y. J., and Veitch, V. The linear representation hypothesis and the geometry of large language models, 2024. URL https://arxiv.org/abs/2311.03658.

Pearl, J. *Causality*. Cambridge university press, 2009.

Raghu, M., Gilmer, J., Yosinski, J., and Sohl-Dickstein, J. Svcca: Singular vector canonical correlation analysis for deep learning dynamics and interpretability. *Advances in neural information processing systems*, 30, 2017.

Ribeiro, M. T., Singh, S., and Guestrin, C. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Rimsky, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., and Turner, A. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

Robert, P. and Escoufier, Y. A unifying tool for linear multivariate statistical methods: the rv-coefficient. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 25(3):257–265, 1976.

Romanou, A., Ibrahim, M., Ross, C., Shaib, C., Oktar, K., Bell, S. J., Ovalle, A., Dodge, J., Bosselut, A., Sinha, K., and Williams, A. Brittlebench: Quantifying llm robustness via prompt sensitivity, 2026. URL https://arxiv.org/abs/2603.13285.

Sakaguchi, K., Bras, R. L., Bhagavatula, C., and Choi, Y. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Serrano, S. and Smith, N. A. Is attention interpretable? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 2931–2951, 2019.

Templeton, A., Conerly, T., Marcus, J., Lindsey, J., Bricken, T., Chen, B., Pearce, A., Citro, C., Ameisen, E., Jones, A., Cunningham, H., Turner, N. L., McDougall, C., MacDiarmid, M., Freeman, C. D., Sumers, T. R., Rees, E., Batson, J., Jermyn, A., Carter, S., Olah, C., and Henighan, T. Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet. *Transformer Circuits Thread*, 2024. URL https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., and MacDiarmid, M. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., and Steinhardt, J. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

Wiegreffe, S. and Pinter, Y. Attention is not not explanation. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 11–20, 2019.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pp. 38–45, 2020.

A. Experimental Details

Models. We evaluate the Qwen2.5 model family (Bai et al., 2023) at four scales: 0.5B, 3B, 7B, and 14B parameters, using the instruction-tuned variants (Qwen2.5-{0.5B,3B,7B,14B}-Instruct). For the scaling analysis in Section B.2, we additionally evaluate the base (non-instruction-tuned) variants at the same scales plus 32B. All models are loaded in `bfloat16` precision on a single NVIDIA A100 80 GB GPU using HuggingFace Transformers (Wolf et al., 2020) with `device_map="auto"`. Models are loaded and unloaded sequentially to fit within GPU memory.

Inference. We extract logits from a single forward pass with no sampling—we do not generate text. For each prompt, we compute the log-softmax over the *full vocabulary* at the last token position (i.e., the position where the model would begin generating). This differs from API-based approaches that return only the top-k log-probabilities; our local inference provides exact log-probabilities for any token. Temperature is not applicable as we operate on raw logits without sampling.

Log-odds scoring. For a prompt x and label classes $\mathcal{P}$ (positive) and $\mathcal{N}$ (negative), we define the log-odds as:

$$\ell(x) = \log \sum_{t \in \mathcal{P}} P(t \mid x) - \log \sum_{t \in \mathcal{N}} P(t \mid x) \tag{17}$$

where $P(t \mid x)$ is the next-token probability for token t. For BoolQ, $\mathcal{P}$ and $\mathcal{N}$ correspond to the tokens "true" and "false" respectively. To handle tokenizer variation, we expand each label to include case variants (`true`, `True`, `TRUE`) and prefix variants (no prefix, space prefix, underscore prefix), yielding 9 token variants per label. Log-probabilities across variants are aggregated via log-sum-exp.

For ANLI, $\mathcal{P}$ and $\mathcal{N}$ correspond to token prefixes "ent" and "contr" (with the same case/prefix expansion); for WinoGrande, "1" and "2" (no expansion needed for single digits).

Prompt templates. Each benchmark uses a task-specific system prompt and a structured user prompt. For BoolQ:

> **System:** "You are a general-purpose binary question-answering machine. You will be shown a passage (marked 'Context'), followed by a true/false question (marked 'Question') about that passage. Answer the question on the basis of the context, and respond only with 'true' or 'false'."
>
> **User:** "Context: {passage}. Question: {question}."

ANLI uses "Premise" / "Hypothesis" labels with an entailment/neutral/contradiction instruction; WinoGrande uses "Sentence" / "Option 1" / "Option 2" with a blank-filling instruction. Prompts are formatted using each model's native chat template via `tokenizer.apply_chat_template()`.

Segmentation. We segment prompts using Captum's InterpretableInput framework (Kokhlikyan et al., 2020). For sentence-level experiments, we use the `"sentence"` TextSegmentInput; for word-level experiments, the `"word"` TextSegmentInput. Segmentation operates on the user message text only (not the system prompt).

Ablation scoring. For each prompt, we perform leave-one-out ablation: for each of N segments, we remove that segment and re-score the modified prompt. The ablation response for segment i is:

$$a_i(x) = \ell(x) - \ell(x_{\setminus i}) \tag{18}$$

where $x_{\setminus i}$ is the prompt with segment i removed. This requires $N + 1$ forward passes per prompt (one for the original, one per ablation).

Attention scoring. We extract attention weights from a single forward pass with `output_attentions=True`. Each transformer layer l produces an attention matrix $A^{(l)} \in \mathbb{R}^{H \times T \times T}$ where H is the number of heads and T is the sequence length. We consider three aggregation strategies:

- **Mean pooling:** Average over all heads and layers, then extract the last row (attention from the prediction position to all input tokens).
- **Max pooling:** Maximum over all heads and layers, then extract the last row.

- **Attention rollout** (Abnar & Zuidema, 2020): At each layer, average over heads, mix 50/50 with the identity matrix (accounting for residual connections), renormalize rows, and multiply into a running product. The result captures attention flow through the full network.

Token-level attention is aggregated to segment-level scores by summing attention weights over the tokens belonging to each segment (mapped via tokenizer offset positions).

Agreement metrics. Cross-model agreement is measured via Pearson correlation: for each benchmark, we concatenate all scalar measurements into a single vector and compute r^2. All reported correlations include bootstrapped 95% confidence intervals ($B = 1{,}000$ resamples of the per-prompt correlation distribution).

Benchmarks. We evaluate on five benchmarks: BoolQ (Clark et al., 2019) (3,270 validation examples), ANLI (Nie et al., 2020) R1/R2/R3 (1,000/1,000/1,200 test examples), and WinoGrande (Sakaguchi et al., 2021) (1,267 validation examples).

Compute. All experiments are run on a single NVIDIA A100 80 GB GPU. Sentence-level scoring (attention + ablation) for one benchmark $\times$ four models takes approximately 2–4 hours depending on dataset size. Word-level ablation is substantially more expensive ($\sim$100$\times$ more forward passes) and runs for approximately 24 hours per benchmark.

B. Additional results

B.1. Detailed cross-representation agreement heatmap

A detailed heatmap of attention scoring methods and their agreement with F_{attr} as cross-predictors is given in Figure 6.

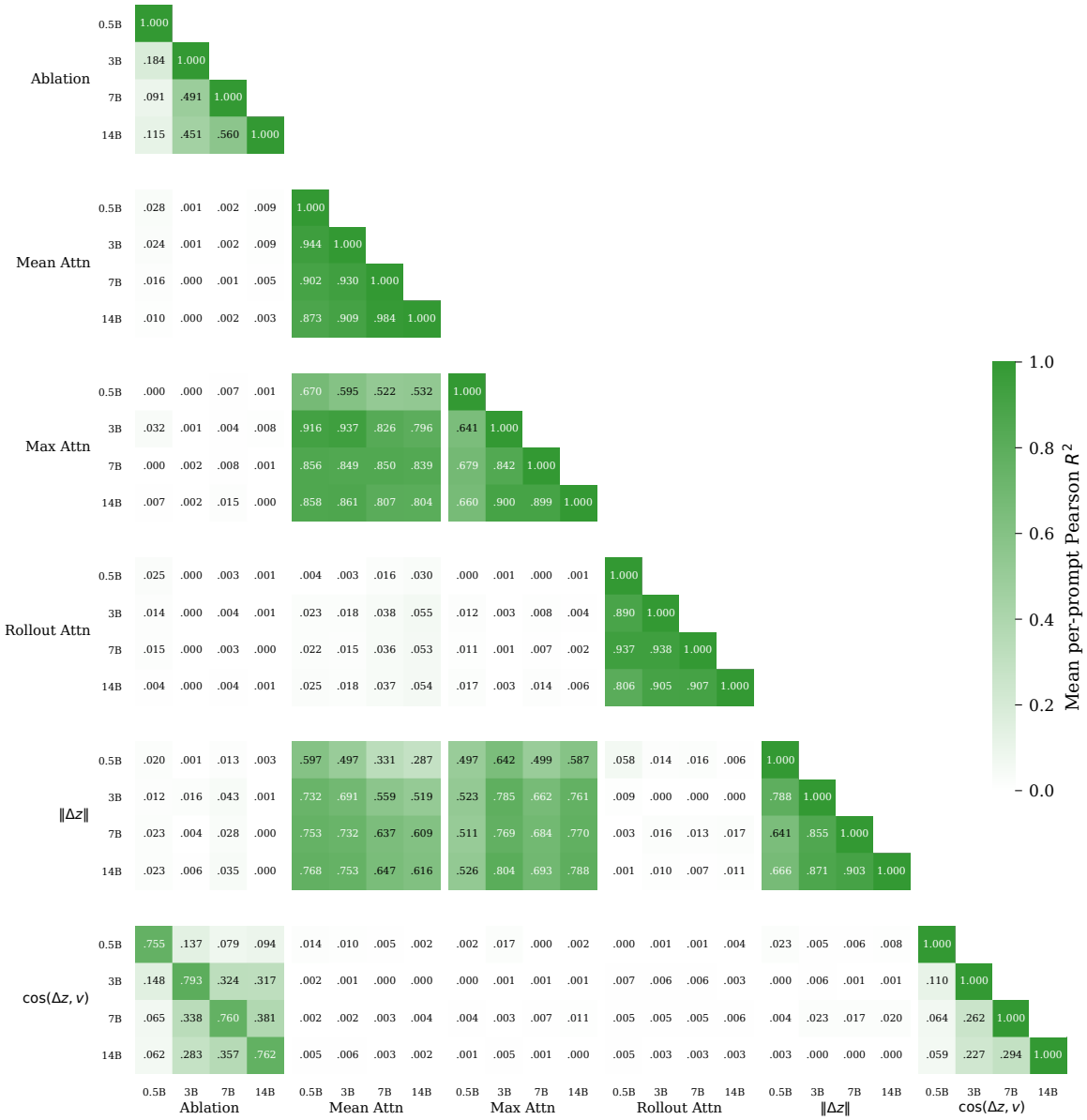


Figure 6. Fidelity metrics across a variety of measurements for Qwen 2.5 models with 0.5B, 3B, 7B, and 14B parameters. We measure attributions under ablation; mean-, max-, and rollout-pooled attention scores; perturbation norms; and perturbation alignment—all on BoolQ at the sentence level. We find that ablation scores are generally hard to predict, even using surrogate models of different sizes; predicting ablation scores using any attention-based metric is unreliable.

B.2. Output log-odds scaling in Qwen

We plot the log-odds of 100 randomly-selected prompts over different Qwen model sizes, as well as the true- and false-class medians, in Figure 7. We repeat this process for base and instruct models, and find that instruct fine-tuning appears to drive log-odds differentiation in models.

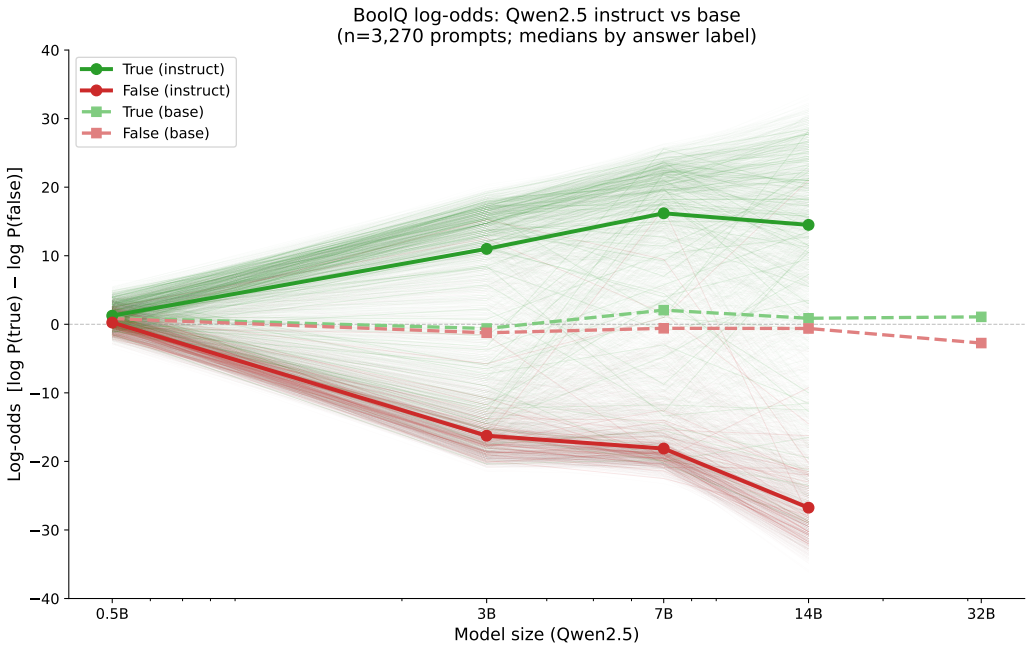


Figure 7. Median log-odds by ground-truth label for instruct (solid circles) and base (dashed squares) models, with individual prompt trajectories shown as translucent lines. Instruction-tuned models exhibit increasing class separation with scale, while base models remain near zero regardless of size. This suggests that the prediction signal measured by F_{pred} is largely a product of instruction tuning rather than pre-training alone.

B.3. Base log-odds distribution

In Figure 8, we plot histograms of base log-odds distributions for different models on the BoolQ dataset. Histogram bars are subdivided by the proportion of true and false responses (ground truth) in the corresponding bin. Note that Qwen-0.5B has a highly irregular distribution compared to the rest of the models; this is a likely explanation for why it has extremely low F_{pred} with all other models.

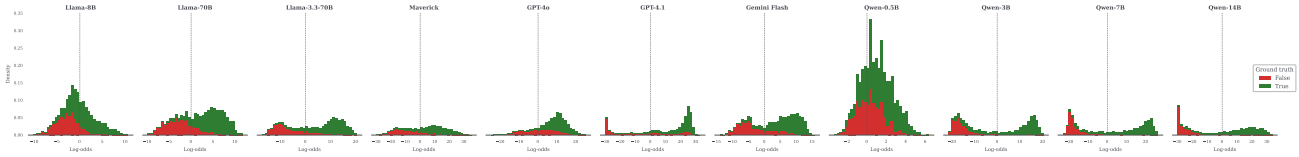


Figure 8. **Distribution of model log-odds on BoolQ by ground-truth label.** Each panel shows stacked histograms of log-odds for a single model, with green = TRUE and red = FALSE. Models above 3B parameters produce bimodal distributions with clear class separation. Qwen-0.5B exhibits substantial overlap between classes, consistent with its low prediction fidelity. The scale of log-odds varies widely across models (e.g., $|\ell| > 30$ for GPT-4.1 vs. $|\ell| < 10$ for Llama-8B), motivating the use of rank correlation.

B.4. Metric calibration

In Figure 9, we compare Spearman ρ and r^2, as well as F_{repr} and F_{cross} scores, in open models. Each point represents one pair of distinct open models.

Figure 10 shows that the log-sum-exp over variants of the "True" token in BoolQ is highly correlated with the individual "true" vs "false" direction. In practice, only a single difference of unembeddings is linear; the log-add-exp breaks this nonlinearity. Because these quantities are so similar to one another, we do not expect significant issues from this approximation.

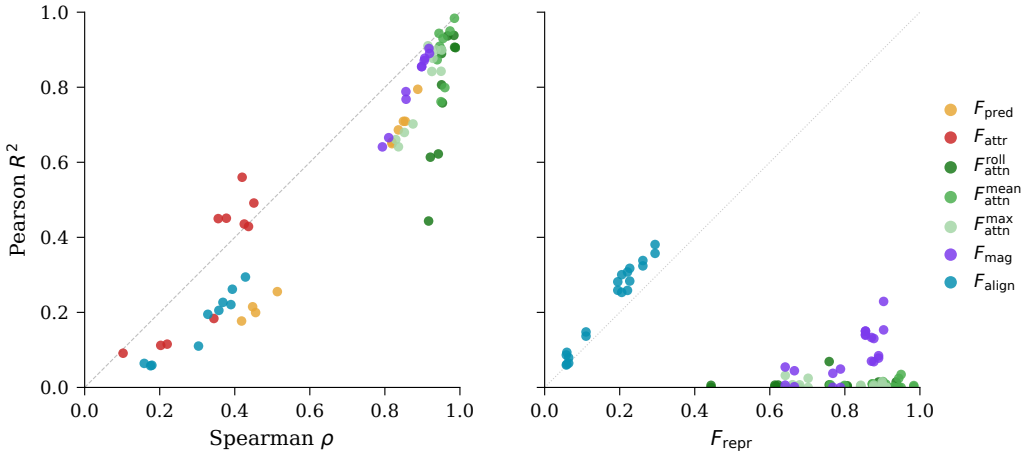


Figure 9. **Left:** a scatterplot of Spearman ρ against Pearson r^2. Note that only F_{attr} has $r^2 > \rho$. **Right:** a scatterplot of F_{repr} versus F_{cross} for different representation-level quantities. Note that F_{pred} and F_{attr} are excluded here, and that only F_{align} has better cross- than self-prediction.

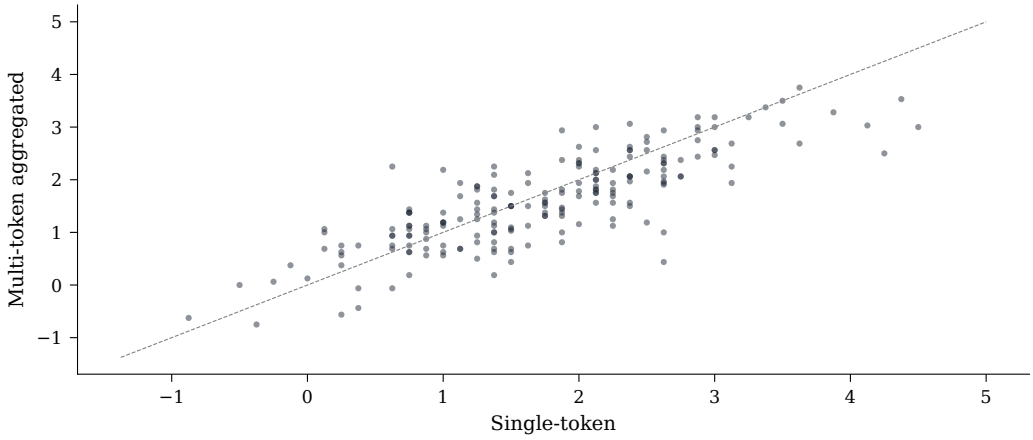
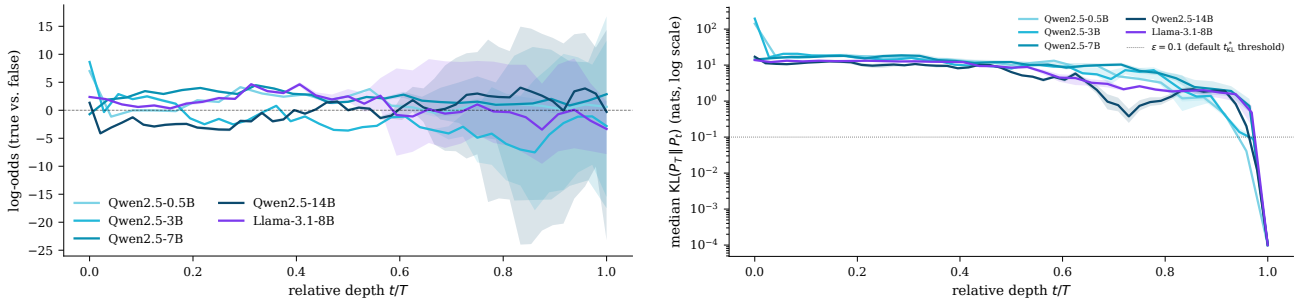


Figure 10. The correlation between the log-sum-exp of a set of tokens and a single-token variant is high, likely due to tight clustering between semantically similar tokens in unembedding space. Across the open models used in this paper, the average unembedding vector cosine similarity is 0.648 for true tokens, 0.647 for false tokens, and 0.515 between true and false tokens. The high cosine similarity between true and false tokens implies that the true-vs-false direction captured in the log-odds is only a small fraction of what is represented in the LLM's last layer.

C. Per-layer fidelity trajectories

We extend the surrogate-fidelity framework of Section 3 from the final layer to every layer of the residual stream, asking at *which depth* the cross-model agreement on each fidelity level emerges. Throughout, we use the *logit lens* (nostalgebraist, 2020) convention introduced in Section 2: for every $l \in \{0, 1, \ldots, L\}$ we read $\ell^{(l)}(x) := \mathcal{N}(z^{(l)}) \cdot v$ where $\mathcal{N}$ is the model's *final* RMSNorm applied to the layer-l residual; index 0 is the embedding-layer projection $\mathcal{N}(z^{(0)}) \cdot v$ and indices $1, \ldots, L$ are the post-block residuals.[3] Per-layer arrays therefore have length $L + 1$. We use the open-weight subset of the model pool (Qwen-2.5 {0.5B, 3B, 7B, 14B}-Instruct and Llama-3.1-8B-Instruct), the BoolQ validation split ($n = 500$ prompts), and the same sentence-level segmenter as Section 4, yielding 2,105 (PROMPT, SEGMENT) pairs that all five models share.

C.1. Trajectories of $\ell^{(l)}$ and the lens distribution



(a) Median $\ell^{(l)}$ (with IQR ribbons) per relative depth l/L across $n = 500$ BoolQ prompts. Trajectories converge through relative depths 0.1–0.5 and re-diverge in the back half.

(b) Median $\mathrm{KL}\big(P^{(L)} \,\|\, P^{(l)}\big)$ per relative depth, where $P^{(l)} = \mathrm{softmax}(W_U^\top \mathcal{N}(z^{(l)}))$ is the lens distribution at layer l. The final-layer distribution is reached only in the last 1–3 layers under any reasonable threshold (the dotted line marks $\varepsilon = 0.1$ nats).

Figure 11. Per-layer trajectories and KL.

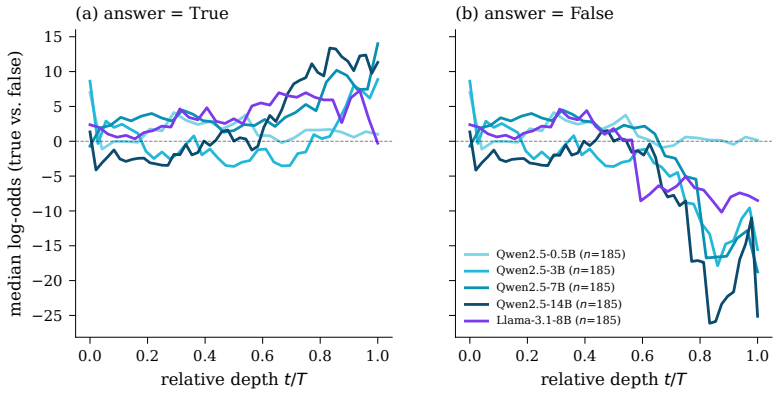


Figure 12. Same as Figure 11a but split by ground-truth answer. The expected sign separation appears almost entirely in the second half of depth and is cleanest for Llama-3.1-8B and Qwen-14B.

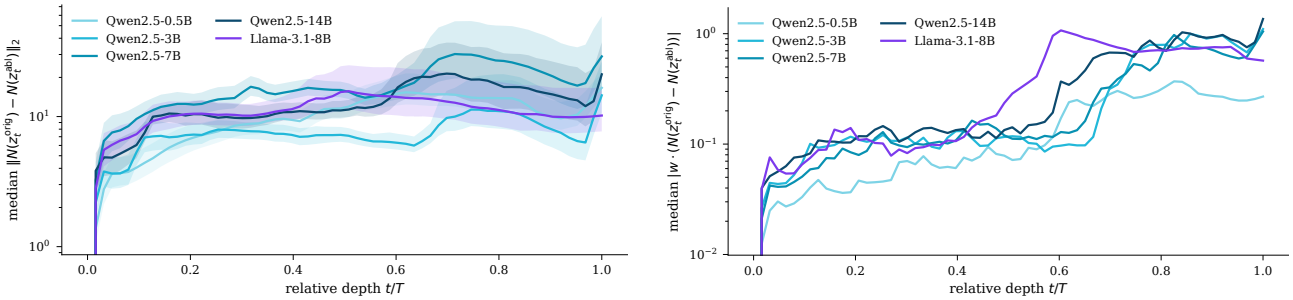
Figure 11a plots the per-layer trajectories of $\ell^{(l)}$. All five models start with a substantial embedding-layer bias whose sign and magnitude depend on the model's static embedding of the assistant-marker token at the prediction position; readers should treat the layer-0 slot as a per-tokenizer baseline rather than a content-dependent prediction. Figure 12 splits the same trajectories by the ground-truth answer: the expected sign separation between the True- and False-prompt sub-populations

[3] We restrict the per-layer extension to single-token positive/negative labels so that the linearization $\ell = z \cdot v$ from Equation (4) holds exactly modulo bit-precision noise, allowing $\ell^{(l)}$ to be read directly from $\mathcal{N}(z^{(l)})$ without instantiating W_U at every layer. We verified this identity on Qwen-0.5B for all $L = 24$ decoder layers: the maximum discrepancy between $\mathcal{N}(z^{(l)}) \cdot v$ and the explicit $W_U[\text{POS}] - W_U[\text{NEG}]$ projection of $\mathcal{N}(z^{(l)})$ is 9.6×10^{-2} versus a five-bf16-ULP threshold of 2.4×10^{-1}. The analogous DLA-decomposition identity from Equation (6) holds to 8.6×10^{-2}.

emerges almost entirely in the back half of depth. Figure 11b summarizes the same data as the per-layer $\mathrm{KL}\big(P^{(L)} \,\|\, P^{(l)}\big)$ — the lens distribution falls within 0.1 nats of the final-layer distribution only in the last 1–3 layers for every model.

C.2. Per-layer attribution decomposition

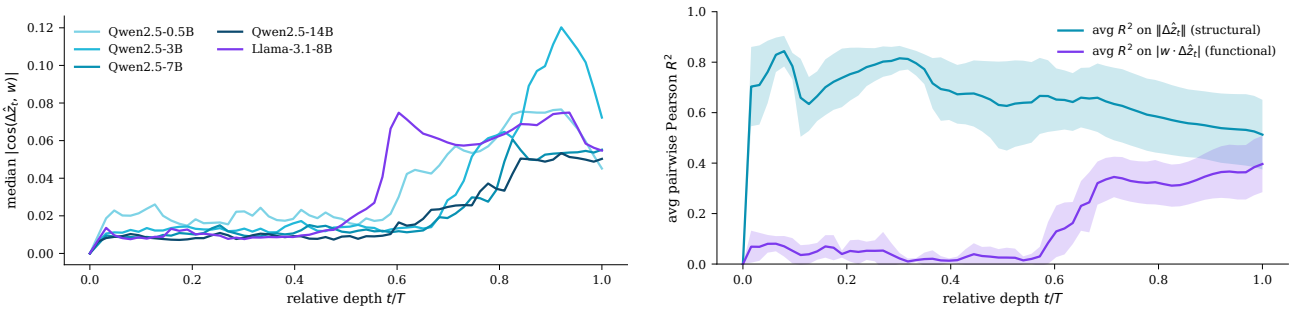
Following the factorization in Equation (8) we record three per-$(\textsc{prompt}, \textsc{segment}, \textsc{layer})$ scalars under the same leave-one-out segment ablation as Section 3: $\|\Delta z^{(l)}\|$, $\Delta\ell^{(l)} = \Delta z^{(l)} \cdot v$, and $\cos(\Delta z^{(l)}, v)$, where $\Delta z^{(l)} := \mathcal{N}(z^{(l)}(x)) - \mathcal{N}(z^{(l)}(\tilde{x}))$. Figures 13a, 13b and 14a show the per-model medians.



(a) Median $\|\Delta z^{(l)}\|$ per relative depth. The structural footprint of an ablation appears almost immediately (depth $\lesssim 0.1$) and stays roughly flat thereafter.

(b) Median $|\Delta\ell^{(l)}|$ per relative depth. The readout-aligned component is small ($\sim$0.05–0.1) for the first half of depth and grows by 1–2 orders of magnitude in the last 30% of layers.

Figure 13. Per-layer structural perturbation norm and readout-aligned attribution magnitude.



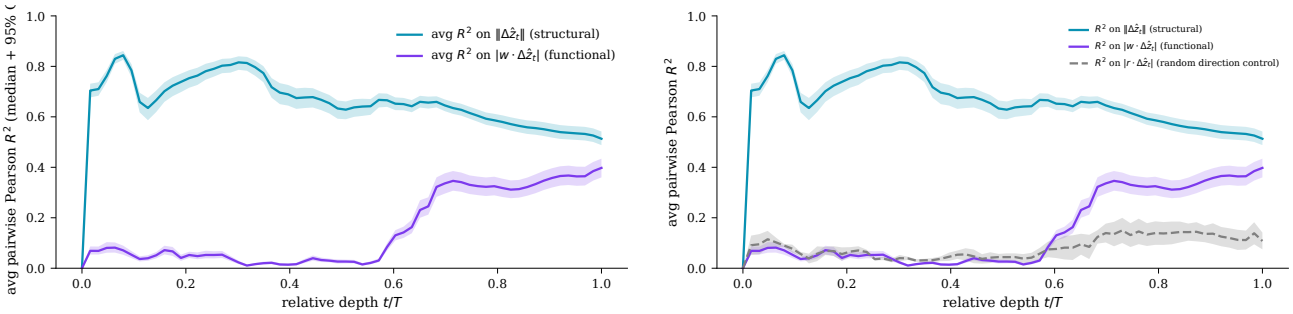
(a) Median $|\cos(\Delta z^{(l)}, v)|$ per relative depth. Comparing with Figures 13a and 13b shows that the late rise in $|\Delta\ell^{(l)}|$ is driven by the rising $|\cos|$ factor of Equation (8), not by larger perturbations.

(b) Pair-averaged $F^{(l)}_{\mathrm{mag}}$ (cyan, structural) and $F^{(l)}_{\mathrm{attr}}$ (purple, functional) versus relative depth, with $\pm 1\sigma$ ribbons across the 10 model pairs.

Figure 14. Per-layer directional alignment and pair-averaged attribution magnitudes.

C.3. Cross-model fidelity at depth

We compute the per-layer extensions $F^{(l)}_{\mathrm{mag}}$ and $F^{(l)}_{\mathrm{attr}}$ of the metrics defined in Equations (11) and (12) on the 2,105 $(\textsc{prompt}, \textsc{segment})$ pairs shared by all five models. Per-model trajectories are resampled to a common length-64 grid in relative depth $l/L \in [0, 1]$ via linear interpolation in l/L space; pairwise Pearson R^2 (the coefficient of determination) is computed at each of the 64 grid points, i.e. the squared Pearson correlation between the two models' per-segment attribution vectors at that depth.



(a) Bootstrap-CI version of Figure 14b ($B = 500$ resamples of (PROMPT, SEGMENT) pairs with replacement; the 95% CI is the empirical 2.5/97.5 percentile of the pair-averaged R^2 at each depth bin). CIs are $\leq \pm 0.05$ in R^2 at every depth.

(b) The structural and functional pair-averaged curves of Figure 15a with the random-direction control overlay (gray dashed). The control rises late too, plateauing at $R^2 \approx 0.13$–0.15 before ending at $R^2 \approx 0.11$ at the readout. The v-specific premium over the random baseline at the readout is therefore ≈ 0.29 in R^2.

Figure 15. Per-layer correlation analysis with bootstrap confidence intervals and random-direction control.

Figures 14b and 15a aggregate over the 10 model pairs. The headline pattern at this scale is a clean *depth dissociation*:

- Structural agreement $F^{(l)}_{\text{mag}}$ is high ($\sim$0.65–0.85) across most depths past the embedding slot, then tapers to $\sim$0.51 at the readout.
- Functional agreement $F^{(l)}_{\text{attr}}$ is low ($\lesssim$0.08) for the first $\sim$55% of depth, rises sharply between depths 0.55 and 0.7, and reaches $\sim$0.40 at the readout.
- The final-layer $F^{(L)}_{\text{attr}} \approx 0.40$ matches the main-paper F_{attr} baseline (Table 1, BoolQ Open→Open) to within sampling noise.

Random-direction control. The dissociation in Figure 15a could in principle be a generic property of late-layer perturbations: any fixed direction in residual-stream space might exhibit a low-then-high cross-model agreement curve. To test this, we replace v in $F^{(l)}_{\text{attr}}$'s definition with a per-model random unit direction r matched in norm to $\|v\|$ (averaging over $K = 5$ random draws per model on a 200-prompt subsample). Figure 15b overlays the random-direction curve on the structural and functional curves of Figure 15a.

The random-direction control rises late as well, plateauing at $R^2 \approx 0.13$–0.15 before ending at $R^2 \approx 0.11$ at the readout — a non-negligible but minority fraction of the headline $F^{(L)}_{\text{attr}} \approx 0.40$. The v-specific premium over the random baseline at the readout layer is therefore ≈ 0.29 in R^2. We read this as a partial qualification of the framework's main-text claim that low F_{attr} reflects models "disagreeing on which inputs drive their predictions": of the cross-model agreement that F_{attr} measures at the readout, roughly three quarters is v-specific (i.e., genuinely about the readout direction) and roughly one quarter is direction-agnostic late-stage residual-stream consistency.

C.4. Caveats and scope

The per-layer extension inherits the assumptions of Sections 2 and 3 but adds three of its own. (i) The logit-lens convention applies $\mathcal{N}$ at every depth even though intermediate layers are not trained to be readable through $(\mathcal{N}, W_U)$; the cross-model curves in this section should therefore be read as comparing models under a shared reading convention rather than as a claim about each model's "internal predictions" at intermediate layers. A tuned-lens robustness check (Belrose et al., 2023) is straightforward to add but was not prioritized — the random-direction control (Section C.3) is more diagnostic for the headline because it directly tests whether the late-rise pattern is v-specific. (ii) We use single-token POS/NEG labels so the linearization $\ell^{(l)} = \mathcal{N}(z^{(l)}) \cdot v$ holds exactly modulo bf16 noise; under the multi-token aggregation used for the main-paper $F_{\text{pred}}/F_{\text{attr}}$ the projection-based identity becomes a softmax-weighted approximation, which changes both calibration and absolute scales. (iii) Our regex sentence segmenter produces fewer segments per BoolQ passage than the production sentence-tokenizer used in Section 4; the $F^{(L)}_{\text{attr}}$ baseline we measure here (≈ 0.40) is consistent with the main-text Open→Open F_{attr} baseline within sampling noise, but absolute counts of (PROMPT, SEGMENT) pairs are not directly comparable. None of these caveats affects the qualitative shape of the depth dissociation, which is robust under all controls in Figures 15a and 15b.

D. Confidence stratification

The fidelity metrics reported in Table 1 aggregate over all prompts in a benchmark. However, prompts vary widely in difficulty: on BoolQ, model log-odds range from $|\ell| < 0.5$ (near-chance) to $|\ell| > 30$ (near-certain). We ask whether surrogate fidelity varies systematically with model confidence, operationalized as $|\ell|$.

D.1. Fidelity versus confidence

We bin BoolQ prompts into five quintiles by mean $|\ell|$ across the four Qwen instruct models and compute F_{pred}, F_{attr}, and F_{attn} within each bin. Figure 16 shows the result.

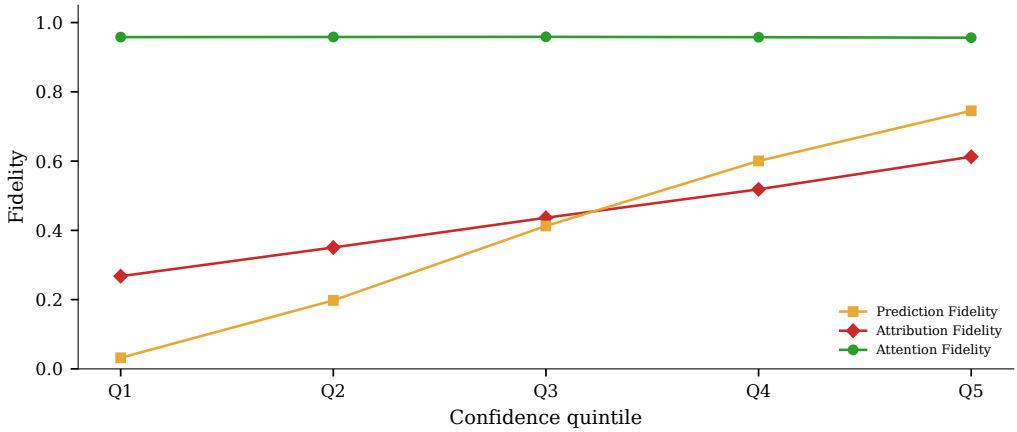


Figure 16. F_{pred}, F_{attr}, and F_{attn} versus model confidence quintile on BoolQ (mean pairwise Pearson R^2 across all six Qwen 2.5 instruct pairs). Prediction fidelity rises sharply from Q1 to Q5 ($0.03 \rightarrow 0.75$); attribution fidelity more than doubles ($0.27 \rightarrow 0.61$); attention fidelity is flat (0.958 ± 0.001).

Prediction fidelity rises from $F_{\text{pred}} = 0.03$ in the lowest-confidence quintile to 0.75 in the highest—a $> 20\times$ increase. Attribution fidelity follows a similar trajectory, climbing from $F_{\text{attr}} = 0.27$ to 0.61 ($+129\%$). By contrast, attention fidelity is invariant to confidence: $F_{\text{attn}} = 0.958 \pm 0.001$ across all five bins.

The strong confidence dependence of F_{pred} carries a practical implication: when models are confident about a prompt, the surrogate's prediction is a reliable proxy for the target's ($F_{\text{pred}} = 0.75$ in the top quintile). For uncertain prompts, however, surrogate predictions are substantially less trustworthy ($F_{\text{pred}} = 0.03$).

The confidence dependence of F_{attr} has a natural mechanistic explanation. Low-confidence prompts produce small log-odds, and ablating a sentence from such a prompt yields a correspondingly small change in log-odds. When most attribution values in a prompt are near zero, the rank ordering among them is dominated by noise rather than signal, deflating R^2. This is the same near-zero concentration effect noted in Section 5.2: high-confidence prompts have larger, more discriminative attribution vectors on which models can meaningfully agree.

The flatness of F_{attn} reinforces the dissociation identified in Section 5.3: attention agreement reflects architectural similarity rather than decision-level consensus, and is therefore insensitive to whether the model finds a prompt easy or hard.

D.2. Whose confidence?

The previous analysis stratifies by the mean confidence across all four Qwen models. A practitioner may instead want to condition on a single model's confidence—for instance, the target's (to assess when a surrogate's explanations are trustworthy for a given target) or the surrogate's (to decide whether a particular surrogate prediction merits further analysis). Figure 17 compares these choices.

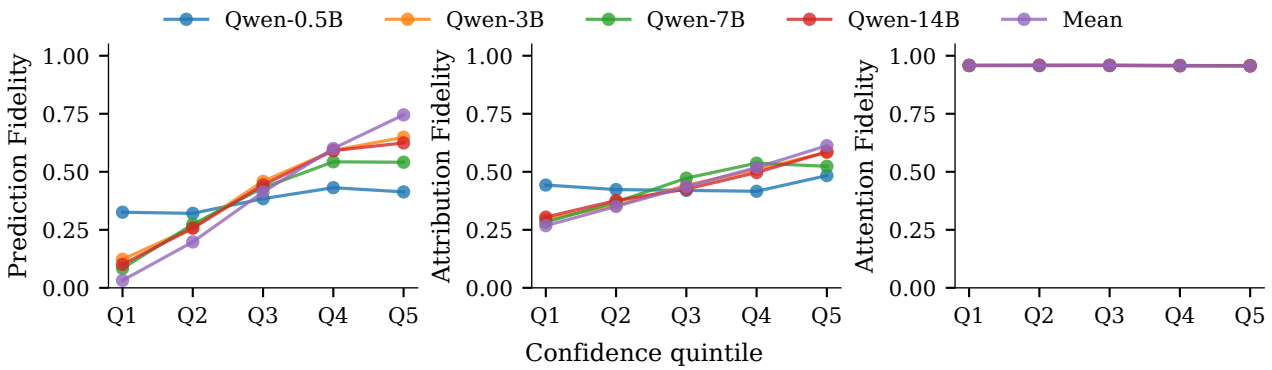


Figure 17. Fidelity versus confidence quintile, stratified by each individual Qwen model's $|\ell|$ as well as their mean. The slope is similar regardless of which model's confidence defines the bins: model confidence is a robust predictor of fidelity independent of which model supplies it.

For both F_{pred} and F_{attr}, the fidelity-versus-confidence slope is qualitatively similar regardless of whether we stratify by the smallest model (Qwen-0.5B), the largest (Qwen-14B), or the mean. This is unsurprising in retrospect: because F_{pred} is itself moderate to high within the Qwen family, a prompt that one model finds easy tends to be easy for the others as well. For F_{attn}, no stratification produces any gradient, consistent with attention's insensitivity to prompt difficulty.

Taken together, these results indicate that a practitioner can cheaply screen prompts for surrogate reliability using any available model's $|\ell|$ as a proxy for confidence: prompts in the top quintile ($|\ell| \gtrsim 10$ for Qwen-14B) yield $F_{\text{attr}} \approx 0.61$, while prompts in the bottom quintile ($|\ell| \lesssim 2$) yield $F_{\text{attr}} \approx 0.27$. The output–behavior gap narrows substantially when conditioned on model confidence: at the high-confidence end the gap between prediction and attribution fidelity is $(0.75, 0.61)$, compared to $(0.03, 0.27)$ at the low end.

E. Directional asymmetry via NRMSE

The agreement metrics used in the main text—Spearman ρ and Pearson R^2—are symmetric: swapping which model is called "surrogate" and which is called "target" leaves the statistic unchanged. This symmetry is desirable for summarizing overall agreement, but it cannot detect *directional* asymmetry: the possibility that using a small model as a surrogate for a large one is systematically worse (or better) than the reverse.

To test for directional effects we introduce normalized root-mean-square error (NRMSE), defined for a surrogate s predicting a target t as

$$\mathrm{NRMSE}(s \to t) \;=\; \frac{\mathrm{RMSE}(s,t)}{\sigma(y_t)}, \tag{19}$$

where y_t is the vector of ground-truth values from model t and $\sigma(y_t)$ is its standard deviation. The denominator normalizes the error by the spread of the truth: $\mathrm{NRMSE} < 1$ means the surrogate is more accurate than the naïve baseline of always predicting $\bar{y}_t$ (the mean of the target), while $\mathrm{NRMSE} > 1$ means it is worse.

Because $\sigma(y_t)$ depends on which model is treated as truth, $\mathrm{NRMSE}(s \to t) \neq \mathrm{NRMSE}(t \to s)$ in general—even though $\mathrm{RMSE}(s,t) = \mathrm{RMSE}(t,s)$ is symmetric. This asymmetry reveals which direction of surrogacy is more reliable: if both directions yield low NRMSE, surrogacy is bidirectional; if one direction is high while the reverse is low, the asymmetry identifies the problematic direction. In our data, $\mathrm{NRMSE}(0.5\mathrm{B} \to 14\mathrm{B}) = 15.2$ while $\mathrm{NRMSE}(14\mathrm{B} \to 0.5\mathrm{B}) \approx 1.0$—the small model's predictions deviate from the 14B model's outputs by $15\times$ its own standard deviation, indicating systematic failure rather than mere noise. Thus, *a high* $\mathrm{NRMSE}(s \to t)$ *flags that model s is an unreliable surrogate for model t, even when the symmetric Spearman ρ between them appears moderate.*

We restrict this analysis to F_{pred} and F_{attr}; raw attention magnitudes differ by up to $10{,}000\times$ across model sizes, causing NRMSE to explode to $\sim$100,000 for some pairs—a scale artifact rather than a meaningful directional difference.

Figure 18 shows 4×4 NRMSE heatmaps for F_{pred} and F_{attr} across the Qwen 2.5 instruct family (0.5B, 3B, 7B, 14B), with rows indexing the surrogate and columns indexing the target.

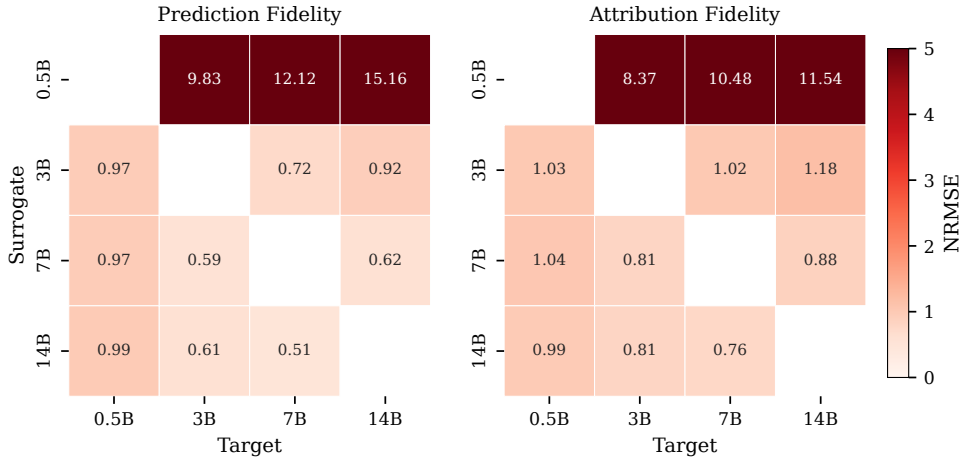


Figure 18. NRMSE heatmaps for F_{pred} (left) and F_{attr} (right) across the four Qwen 2.5 instruct models on BoolQ. Row = surrogate, column = target. Values below 1.0 indicate the surrogate outperforms the mean baseline; values above 1.0 indicate it underperforms. Asymmetry is concentrated in the 0.5B row and column: using 0.5B as a surrogate for 14B yields catastrophic NRMSE (15.2 for F_{pred}, 11.5 for F_{attr}), while the reverse direction gives ≈ 1.0. Among 3B+ models, NRMSE is below 1.0 in most directions.

The heatmaps reveal a sharp qualitative boundary at 3B parameters. For F_{pred}, $\mathrm{NRMSE}(0.5\mathrm{B} \to 14\mathrm{B}) = 15.2$—catastrophically worse than the mean baseline—while the reverse gives $\mathrm{NRMSE} \approx 1.0$. F_{attr} follows the same pattern: 11.5 versus ≈ 1.0. Among the three larger models (3B, 7B, 14B), NRMSE is below 1.0 in most directions for both metrics, indicating that surrogacy is approximately symmetric and consistently better than the mean baseline. The practical implication is that directional asymmetry is a concern only when the surrogate is substantially smaller than the target—specifically, below the $\sim$3B threshold identified in the main text—and that above this threshold a practitioner can treat surrogacy as approximately bidirectional.

F. Representation fidelity decomposition

As shown in Equation (8), the attribution $\Delta\ell = \|\Delta z\| \, \|v\| \, \cos(\Delta z, v)$ decomposes into a perturbation magnitude and an alignment term. The representation-level metrics F_{mag} and F_{align} (Equations (12) and (13)) measure cross-model agreement on each factor separately. Here we examine how the gap between these two sub-metrics explains the low F_{attr} observed in the main text.

Figure 19 shows a 5×5 split heatmap of F_{mag} and F_{align} across the five open models on BoolQ.

Figure 19. Representation fidelity decomposition on BoolQ. Upper triangle: F_{mag} (Pearson R^2 of perturbation magnitudes $\|\Delta z\|$). Lower triangle: F_{align} (Pearson R^2 of readout alignments γ). Models strongly agree on *how much* representations change when a segment is ablated ($F_{\text{mag}} = 0.64$–0.90), but weakly agree on *how much* that change projects onto the readout direction ($F_{\text{align}} = 0.06$–0.29).

F_{mag} is uniformly high: even the weakest pair (Qwen-0.5B $\leftrightarrow$ Qwen-7B, $R^2 = 0.641$) exceeds F_{attr} by a wide margin. Cross-family agreement (Llama-8B vs. any Qwen model) is comparably strong ($F_{\text{mag}} \geq 0.768$), indicating that perturbation magnitude is largely architecture-independent.

F_{align} is substantially lower (mean 0.17, range 0.058–0.294). The gap between F_{mag} and F_{align} identifies the bottleneck for attribution fidelity: models agree that ablating a given segment causes a large representational shift, but disagree on *how much* that shift projects onto the readout direction v. Because $\gamma = \cos(\Delta z, v)$ depends on each model's readout direction—which encodes how the model distinguishes positive from negative classes in representation space—low F_{align} indicates that different models have learned fundamentally different internal geometries for the same binary classification task.

This decomposition clarifies the hierarchy of fidelity observed in the main text:

- $F_{\text{attn}} \approx 0.96$: models attend to the same tokens.
- $F_{\text{mag}} \approx 0.81$: ablating a segment perturbs representations by a similar amount across models.
- $F_{\text{align}} \approx 0.17$: models encode the task along different readout directions, so the same perturbation projects differently onto each model's decision axis.
- $F_{\text{attr}} \approx 0.33$: the causal effect on the output disagrees (a joint consequence of F_{mag} and F_{align}).

The practical implication is that the "output–behavior gap" ($F_{\text{pred}} \gg F_{\text{attr}}$) is not because models disagree about which segments are structurally important—they largely do agree, as F_{mag} shows—but because each model's decision axis differs. A surrogate's attribution ranking is unreliable not because it fails to detect the right perturbations, but because the direction along which perturbations translate into output changes is model-specific.

G. System-prompt perturbation as a secondary behavioral fidelity signal

The behavioral fidelity framework introduced in Section 3 measures attribution via input ablation: removing a segment of the prompt and observing the change in log-odds. We now consider an orthogonal causal intervention: mutate the system prompt, and ask whether models that respond similarly to context perturbation also exhibit greater surrogate fidelity. Where ablation measures sensitivity to the *absence* of information, system-prompt perturbation measures sensitivity to its *framing*.

Concretely, we use GEPA (Agrawal et al., 2026), a state-of-the-art prompt optimization algorithm, both as a method for generating a diverse set of system-prompt candidates and as a tool for per-model optimization. This enables us to assess two strict tests of context-level fidelity: (i) *perturbation-response similarity*: do models whose predictions shift in correlated ways under the same system-prompt variation constitute better surrogates for one another? (ii) *prompt transferability*: does a system prompt optimized with respect to donor model M_D improve performance when applied to recipient model M_R, and does the degree of transfer serve as a directional surrogacy measure in contrast to the similarity metrics of Section 3?

Experimental protocol. We evaluate the same eleven models, spanning four families (Qwen, Llama, GPT, Gemini), described in Section 4.3 on the ANLI R3 test split ($n = 400$), chosen for its difficulty to ensure headroom for prompt-driven gains and reason labels required by GEPA's feedback mechanism. We construct three prompt conditions: (i) *candidates* – ten system prompts selected for maximal tree-path diversity from a single GEPA run targeting GPT-4.1 (GPT-5.2 as reflector), applied uniformly to all models as shared context-level perturbations; (ii) *per-model best* – the top-performing prompt from an independent GEPA run targeting each model individually; and (iii) the *baseline* system prompt used as the control. All GEPA runs use 20 optimization steps with minibatches of 20 from the ANLI R3 development split.

G.1. Perturbation-response fidelity

We ask whether replacing the system prompt with a GEPA-generated alternative drives two models' predictions into alignment, and whether this alignment extends to *how* those predictions shift. For each prompt condition and model, we compute F_{pred} (Pearson R^2 of log-odds) and F_{attr} (Pearson R^2 of $\Delta\ell$ relative to the baseline). Table 3 reports the results.

The prediction-level finding is clear: replacing the baseline system prompt with a more structured alternative increases cross-model prediction agreement. All system prompt condition sets raise F_{pred} by approximately 25% over the baseline. The gain holds across all family groupings and is insensitive to prompt source; even per-model best prompts, where each model is under its own independently optimized prompt, act as a non-uniform perturbation, suggesting that prompt optimization drives models toward a shared output regime independent of perturbation direction.

Attribution fidelity tells a sharply different story. Mean pairwise F_{attr} under diverse candidates is $R^2 = .092$, comparable to the $R^2 = .241$ reported for sentence-level input ablation on ANLI R3 (Table 2), and drops further to .044 under cross-model best prompts. The gap between F_{pred} and F_{attr} (.480 vs. .092) mirrors the output–behavior dissociation observed under input ablation: models that converge on predictions do not converge on how those predictions respond to context perturbation. The dissociation is most pronounced for GPT-4o and GPT-4.1, which achieve the highest F_{pred} in our evaluation ($R^2 = .840$)

Table 3. **Cross-model fidelity under system-prompt perturbation on ANLI R3.** $n = 400$ prompts, 11 models across 4 families. Each cell reports the mean pairwise Pearson R^2 over the relevant model-pair set. F_{pred} measures agreement on log-odds under a shared prompt condition; F_{attr} measures agreement on $\Delta\ell$ induced by replacing the baseline prompt. *Top*: Prediction fidelity increases under optimized prompts (baseline $\bar{R}^2 = .385 \rightarrow$ candidates $\bar{R}^2 = .480$), indicating that structured system prompts drive cross-model output convergence. *Bottom*: Attribution fidelity, defined by the shift in log-odds from perturbing the baseline system prompt, is uniformly low and does not improve with prompt specificity (candidates $\bar{R}^2 = .092 \rightarrow$ cross-model best $\bar{R}^2 = .044$).

	Condition	All	W. Qwen	W. Qwen⁻	W. Llama	W. Llama⁻	W. GPT	Cross
F_{pred}	Baseline	**.385**	.135	.278	.600	.716	.838	.362
	Candidates	**.480**	.305	.430	.602	.730	.840	.469
	Cross-model best	**.478**	.297	.541	.610	.816	.839	.467
	Per-model best[§]	**.479**	.303	.596	.620	.808	.863	.468
F_{attr}	Candidates	**.092**	.030	.077	.204	.213	.095	.080
	Cross-model best	**.044**	.023	.062	.113	.163	.000	.038
	Per-model best[§]	**.069**	.058	.058	.089	.035	.033	.071

W. Qwen⁻ and W. Llama⁻ exclude Qwen-0.5B and Llama-8B respectively.
[§] Non-uniform perturbation: each model evaluated under its own optimized prompt.

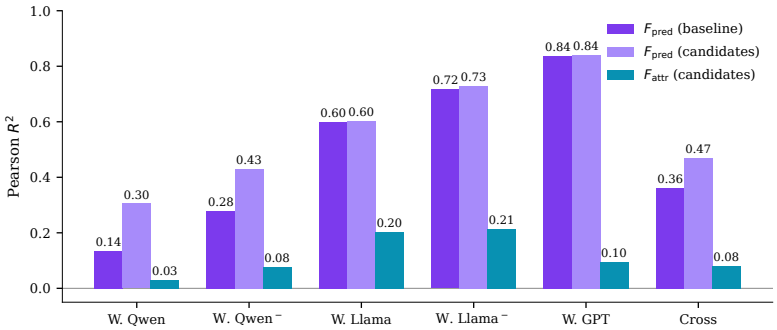


Figure 20. **Prediction and attribution fidelity under system-prompt perturbation on ANLI R3.** Pearson R^2 grouped by model-pair family. **Dark purple:** F_{pred} under the baseline system prompt. **Light purple:** F_{pred} under diverse GEPA-generated candidates. **Cyan:** F_{attr} under the same candidates, computed as the Pearson R^2 of $\Delta\ell$ induced by replacing the baseline prompt. Structured system prompts drive output convergence while attribution fidelity remains uniformly low, the output–behavior gap persists under prompt optimization.

but yet near-zero F_{attr} under diverse candidates (.095) and effectively zero under cross-model best prompts (.000). Despite near-perfect agreement on log-odds, prompts that shift one model's per item predictions do not reliably shift the other's, exemplifying the critical failure mode identified in Section 5.2: high output fidelity masking divergent causal structure.

G.2. Cross-model prompt transfer

The bilateral fidelity metrics in Section G.1 treat prompt perturbation as a shared intervention. We now ask a directed question: can a prompt optimized for one model transfer its behavioral gains to another, and does the degree of transfer serve as a directional surrogacy measure that the symmetric metrics of Section 3 cannot provide?

For each donor–recipient pair (M_D, M_R), we apply the system prompt optimized for M_D to M_R and measure the accuracy change relative to M_R's baseline. Figure 21 (left panel) reports the full 11×11 transfer matrix. Cross-model transfer is predominantly neutral or negative (mean off-diagonal Δaccuracy $= -1.4\%$): prompts that adapt to one model's failure patterns do not generally remedy another's. Crucially, the matrix is asymmetric, transferring M_D's prompt to M_R yields a different outcome than the reverse. In line with the small accuracy deltas, sample-level agreement between M_R's predictions under M_D's optimized prompt and M_R's own baseline is high (right panel; off-diagonal mean 90.4%). Optimized prompts

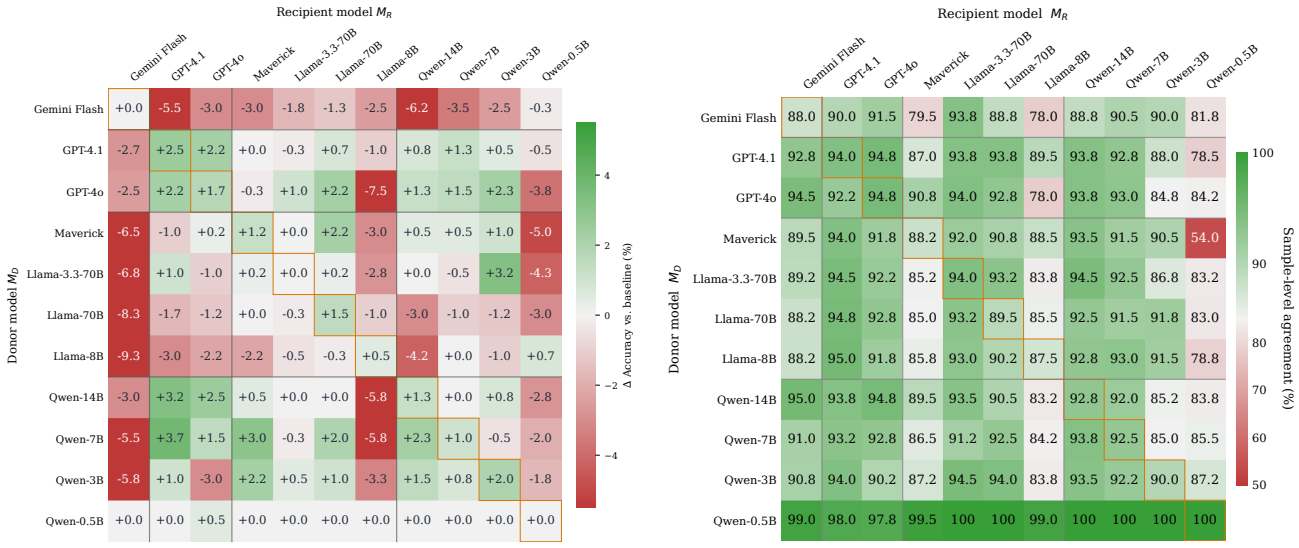


Figure 21. **Cross-model system-prompt transfer on ANLI R3** ($n = 400$). *Left*: Δaccuracy when recipient model M_R is evaluated under a system prompt optimized for donor model M_D, relative to M_R's baseline performance. Cross-model transfer is predominantly neutral or negative. *Right*: sample-level prediction agreement between M_R under M_D's optimized prompt and M_R's own baseline (off-diagonal mean 90.4%). Optimized prompts largely preserve per-sample decisions despite shifting aggregate accuracy.

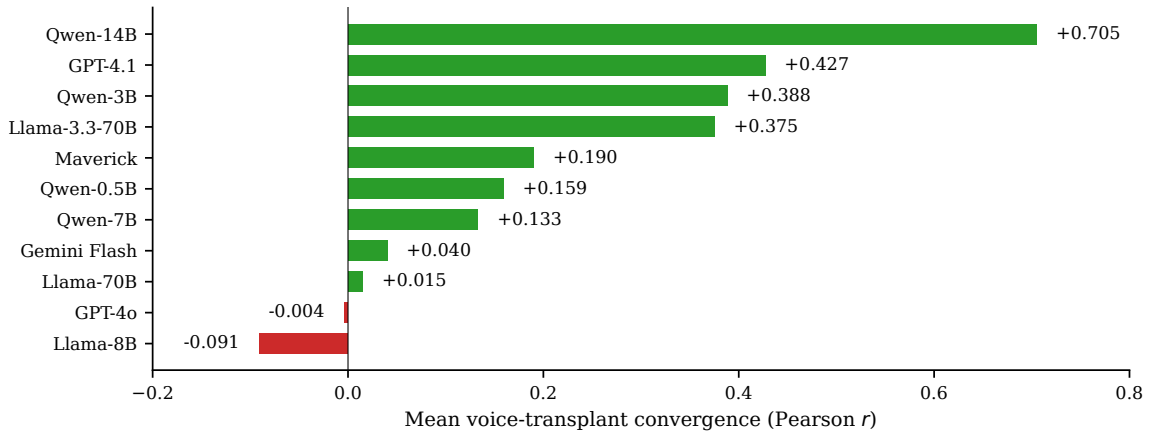


Figure 22. **Recipient susceptibility to voice transplant on ANLI R3.** For each ordered pair (M_D, M_R), we define $\mathrm{VT}(M_D \to M_R) =$ Pearson r on the gap in donor and recipient baseline log-odds $(\ell_{M_D} - \ell_{M_R})$ and the recipient's log-odds shift under the donor's optimized prompt. Bars show the mean VT across all non-self donors; positive values indicate the recipient shifts toward the donor, negative values indicate a shift away. Qwen-14B ($r = +0.71$) consistently converges toward its donors, while Llama-8B ($r = -0.09$) resists.

largely preserve which samples each model answers correctly, perturbing a minority of per-item decisions. The principal exception is Maverick → Qwen-0.5B (54.0%), consistent with Qwen-0.5B's position as a general outlier (Section 5.1).

G.3. Voice transplant: prompt as directional bias vector

F_{pred} and F_{attr} measure bilateral similarity: are two models alike? They do not measure whether one model's behavior can be steered toward another's. We introduce the *voice-transplant correlation* to capture this asymmetric notion of surrogacy. For each ordered pair (M_D, M_R), we define

$$\mathrm{VT}(M_D \to M_R) = r\Big(\ell^{\text{baseline}}_{M_D} - \ell^{\text{baseline}}_{M_R},\ \Delta\ell^{\text{best}_{M_D}}_{M_R}\Big) \tag{20}$$

where $\ell^{\text{baseline}}_{M_D} - \ell^{\text{baseline}}_{M_R}$ is the natural difference in log-odds output between the donor and recipient under the baseline prompt, and $\Delta\ell^{\text{best}_{M_D}}_{M_R}$ is the shift in the recipient's log-odds induced by applying the donor's optimized prompt. Positive VT indicates the recipient shifts toward the donor; negative VT indicates the recipient shifts away.

Figure 22 reports mean VT received across all non-self donors, ranked by susceptibility. Qwen-14B achieves the highest mean VT ($r = +0.71$): applying any other model's optimized prompt consistently shifts its log-odds toward the donor's baseline, indicating strong behavioral malleability. Most models demonstrate some degree of positive convergence, GPT-4.1 ($r = +0.43$), Qwen-3B ($r = +0.39$), Llama-3.3-70B ($r = +0.38$), spanning model families and architectures, suggesting that prompt-based alignment is broadly viable. At the other end, Llama-8B ($r = -0.09$) and GPT-4o ($r = -0.004$) show negligible or negative convergence: applying another model's optimized prompt does not reliably close the log-odds gap.

G.4. Summary

System-prompt perturbation replicates the central finding under a complementary form of causal intervention. Perturbing the instructional context drives cross-model output convergence while attribution fidelity remains uniformly low, independent of if the perturbation is a diverse set of candidates, or optimized for each model itself. The dissociation is starkest for GPT-4o and GPT-4.1, which achieve the highest prediction fidelity in our evaluation (0.840) yet near-zero attribution fidelity under every prompt condition tested. Cross-model prompt transfer shows limited success in our evaluation, though the degree and direction of transfer vary across pairs, indicating that prompt-level surrogacy is inherently directional. Finally, we propose voice-transplant (VT) susceptibility as a lightweight, API-accessible measure of a model's impressionability, its capacity to be behaviorally steered toward a target. This provides MI practitioners a preliminary diagnostic for identifying which surrogates are amenable to prompt-based alignment before committing to mechanistic analysis.

H. Extension to multi-class classification

The framework of Section 3 assumes binary classification, where the log-odds $\ell(x) = \log P(y^+) - \log P(y^-)$ reduce to a single scalar per prompt. We now extend prediction and attribution fidelity to K-class settings (e.g., $K = 4$ for multiple-choice benchmarks such as RACE), where a single scalar cannot capture the full structure of a model's prediction.

Pairwise log-odds representation. For a K-class task with label set $\mathcal{Y} = y_1, \ldots, y_K$, we represent each model's prediction as the vector of all $\binom{K}{2}$ pairwise log-odds:

$$\boldsymbol{\ell}(x) = \big(\log P(y_a \mid x) - \log P(y_b \mid x)\big)_{(a,b)\in\mathcal{C}}, \tag{21}$$

where $\mathcal{C} = (a, b) : 1 \leq a < b \leq K$. For a 4-class task, this yields a 6-dimensional vector per prompt.

This representation has two desirable properties. First, it is *label-free*: no knowledge of the correct answer is required, since all class pairs are included symmetrically. Second, pairwise log-odds are invariant to the softmax normalization constant C (Equation (3)), so they isolate relative preferences among the valid options and are unaffected by how much probability mass a model places on non-option tokens. The binary case is recovered exactly: when $K = 2$, $\binom{2}{2} = 1$ and $\boldsymbol{\ell}(x)$ reduces to the scalar $\ell(x)$ of Equation (4).

Ablation in the multi-class setting. Under leave-one-out ablation, each segment removal changes the prediction vector from $\boldsymbol{\ell}(x)$ to $\boldsymbol{\ell}(x_{\setminus i})$. The vector-valued ablation response is:

$$\boldsymbol{\Delta\ell} i(x) = \boldsymbol{\ell}(x) - \boldsymbol{\ell}(x\setminus i) \in \mathbb{R}^{\binom{K}{2}}, \tag{22}$$

which captures how removing segment i shifts the relative preference between every class pair simultaneously.

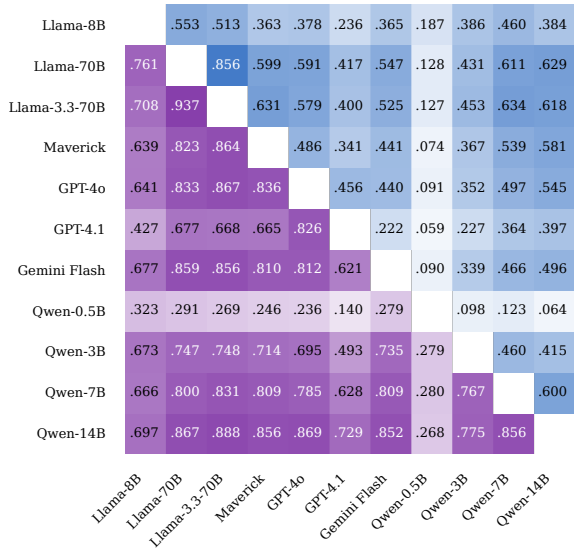


Figure 23. **RACE cross-model Multivariate RV fidelity scores.** Pairwise F_{pred} (**purple**) and F_{attr} (**blue**) heatmaps for eleven models across four families.

RV coefficient. To measure agreement between two models' vector-valued predictions or attributions, we require a multivariate extension of correlation. We use the *RV coefficient* (Robert & Escoufier, 1976), which measures the closeness of two data matrices in the Hilbert–Schmidt inner product space.

For two centered data matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n\times p}$ (where n is the number of observations and $p = \binom{K}{2}$), the RV coefficient is:

$$\mathrm{RV}(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{tr}\big(\mathbf{X}^\top\mathbf{Y}, \mathbf{Y}^\top\mathbf{X}\big)}{\sqrt{\mathrm{tr}(\mathbf{X}^\top\mathbf{X}, \mathbf{X}^\top\mathbf{X}) \cdot \mathrm{tr}(\mathbf{Y}^\top\mathbf{Y}, \mathbf{Y}^\top\mathbf{Y})}}. \tag{23}$$

The RV coefficient ranges from 0 (no agreement) to 1 (proportional matrices). When $p = 1$, it reduces to the squared Pearson correlation R^2.

Multi-class fidelity metrics. Prediction and attribution fidelity extend using the pairwise log-odds vectors and RV:

$$F_{\text{pred}}(M_S, M_T; \mathcal{D}) = \text{RV}(\boldsymbol{L}_{M_S}, \boldsymbol{L}_{M_T}), \tag{24}$$

$$F_{\text{attr}}(M_S, M_T; \mathcal{D}) = \text{RV}(\boldsymbol{\Delta L}_{M_S}, \boldsymbol{\Delta L}_{M_T}), \tag{25}$$

where $\boldsymbol{L}_M \in \mathbb{R}^{|\mathcal{D}| \times \binom{K}{2}}$ stacks the per-prompt prediction vectors, and $\boldsymbol{\Delta L}_M \in \mathbb{R}^{N \times \binom{K}{2}}$ stacks all segment-level ablation response vectors globally across all prompts ($N = \sum x N_x$, where N_x is the number of segments in prompt x).

Both metrics compute the RV coefficient on the full data matrices globally. This parallels the binary case, where F_{pred} and F_{attr} are global correlations across all prompts or (prompt, segment) pairs respectively.

H.1. Demonstrative experiment: RACE

We evaluate multi-class fidelity on RACE (Lai et al., 2017), a 4-way multiple-choice reading comprehension benchmark with 4,934 test prompts. Each prompt consists of a multi-paragraph article followed by a question with four options (A–D), making it well-suited for sentence-level ablation: passages contain 8–10 sentences on average.

Results. We score 11 models spanning four families: Qwen2.5, and Llama-3/4, GPT and Gemini. Figure 23 shows 11×11 pairwise heatmaps of F_{pred} and F_{attr} on RACE. Comparing this to the heatmap in Figure 3, which is evaluated on the binary dataset of BoolQ, the relative model fidelities of both prediction and attribution are well aligned between these two datasets.

I. Connection to representational similarity

The progression from binary log-odds to multi-class pairwise log-odds suggests a deeper perspective: our fidelity metrics are *task-relevant projections* of a full representational similarity measure, while accessible through log-probabilities alone.

From scalars to vectors to representations. In the binary case, prediction fidelity compares the scalar log-odds $\ell(x) = z \cdot v$ across prompts, where $v = u_+ - u_-$ is the log-odds direction (Equation (4)). This is a projection of the residual stream $z \in \mathbb{R}^d$ onto a single direction in representation space. In the multi-class case (Section H), the pairwise log-odds vector $\boldsymbol{\ell}(x) \in \mathbb{R}^{\binom{K}{2}}$ corresponds to projecting z onto $\binom{K}{2}$ directions simultaneously:

$$\ell_{ab}(x) = z \cdot (u_a - u_b), \qquad (a, b) \in \mathcal{C}, \tag{26}$$

where u_a, u_b are columns of the unembedding matrix W_U. Stacking these projections, the full pairwise log-odds vector is $\boldsymbol{\ell}(x) = Pz$, where $P \in \mathbb{R}^{\binom{K}{2} \times d}$.

The same applies to attributions: the vector-valued ablation response $\boldsymbol{\Delta\ell}_i(x) = P\Delta z_i$ (Equation (22)) is the projection of the representation perturbation $\Delta z_i = z - z_i'$ onto the same task-relevant subspace.

At the limit, one could skip the projection entirely and compare the full residual-stream matrices $\mathbf{Z}_M \in \mathbb{R}^{n \times d}$ directly. This requires white-box access to model internals, but yields a complete picture of representational similarity.

Centered Kernel Alignment. The natural measure for comparing full representation matrices is *Centered Kernel Alignment* (CKA) (Kornblith et al., 2019), which has become the standard for comparing neural network representations across architectures. For centered data matrices $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$, linear CKA is defined as:

$$\mathrm{CKA}(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{tr}\big(\mathbf{X}\mathbf{X}^\top \mathbf{Y}\mathbf{Y}^\top\big)}{\sqrt{\mathrm{tr}((\mathbf{X}\mathbf{X}^\top)^2) \cdot \mathrm{tr}((\mathbf{Y}\mathbf{Y}^\top)^2)}}. \tag{27}$$

This is the cosine similarity between the two models' Gram matrices $\mathbf{X}\mathbf{X}^\top$ and $\mathbf{Y}\mathbf{Y}^\top$ in the Hilbert–Schmidt inner product space. Each Gram matrix entry $(\mathbf{X}\mathbf{X}^\top)_{ij} = x_i \cdot x_j$ captures the dot-product similarity between two prompts' representations within a model; CKA asks whether these pairwise similarity structures agree across models.

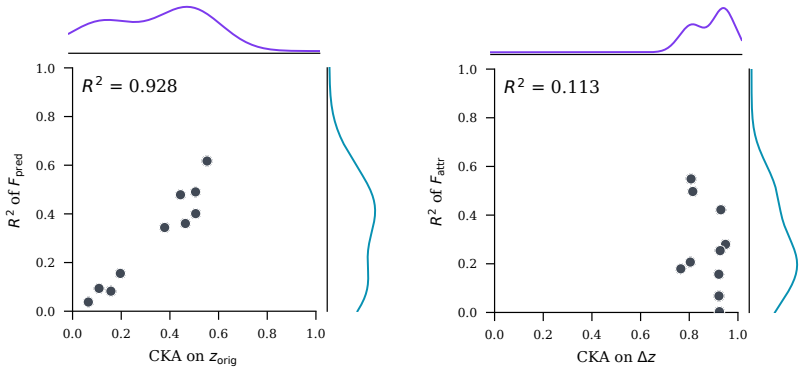


Figure 24. **Correlation between representation CKA and R^2 Fidelity on BoolQ.** The prediction fidelity and CKA are perfectly aligned, but surprisingly the attribution fidelity says the opposite.

Linear CKA is mathematically equivalent to the RV coefficient. When applied to scalar data ($d = 1$), it reduces to the squared Pearson correlation r^2 — which is the binary prediction fidelity metric F_{pred} of Equation (10). The RV coefficient (Equation (23)) applied to multi-class pairwise log-odds is therefore CKA on the projected representations $P\mathbf{Z}$. Our fidelity framework forms a unified hierarchy:

$$\underbrace{r^2(\ell_{M_S}, \ell_{M_T})}_{\text{binary } F\text{pred}} \subset \underbrace{\mathrm{RV}(\boldsymbol{L}_{M_S}, \boldsymbol{L}_{M_T})}_{\text{multi-class } F\text{pred}} \subset \underbrace{\mathrm{CKA}(\mathbf{Z}_{M_S}, \mathbf{Z}_{M_T})}_{\text{representational similarity}} \tag{28}$$

Each level projects the residual stream onto progressively fewer task-relevant directions: d (full representation), $\binom{K}{2}$ (pairwise log-odds), or 1 (binary log-odds).

Task-relevant projections vs. random projections. Our projections are not arbitrary — they use the unembedding directions $(u_a - u_b)$ that the model was *trained* to separate. The training objective (next-token prediction) optimizes the residual stream z so that its projection onto unembedding vectors produces the correct probability distribution. These directions therefore capture maximal task-relevant variance in z by construction.

By contrast, the Johnson–Lindenstrauss lemma (Johnson & Lindenstrauss, 1984) guarantees that random projections from $\mathbb{R}^d$ to $\mathbb{R}^k$ preserve pairwise dot products up to $(1 \pm \varepsilon)$ distortion when $k \geq O(\log n/\varepsilon^2)$. For $n = 5{,}000$ prompts and $\varepsilon = 0.1$, this requires $k \approx 850$ random dimensions to approximate the full Gram matrix. Our task-relevant projections use only $k = \binom{K}{2}$ directions (6 for a 4-class task) yet capture the subspace that determines the classification output, i.e., the operationally relevant subspace for surrogate fidelity.

This perspective also clarifies the practical value of our fidelity framework: it provides a *window into representational similarity through black-box API access*. Full CKA requires white-box access to hidden states; our projected variant requires only log-probabilities for the related tokens.

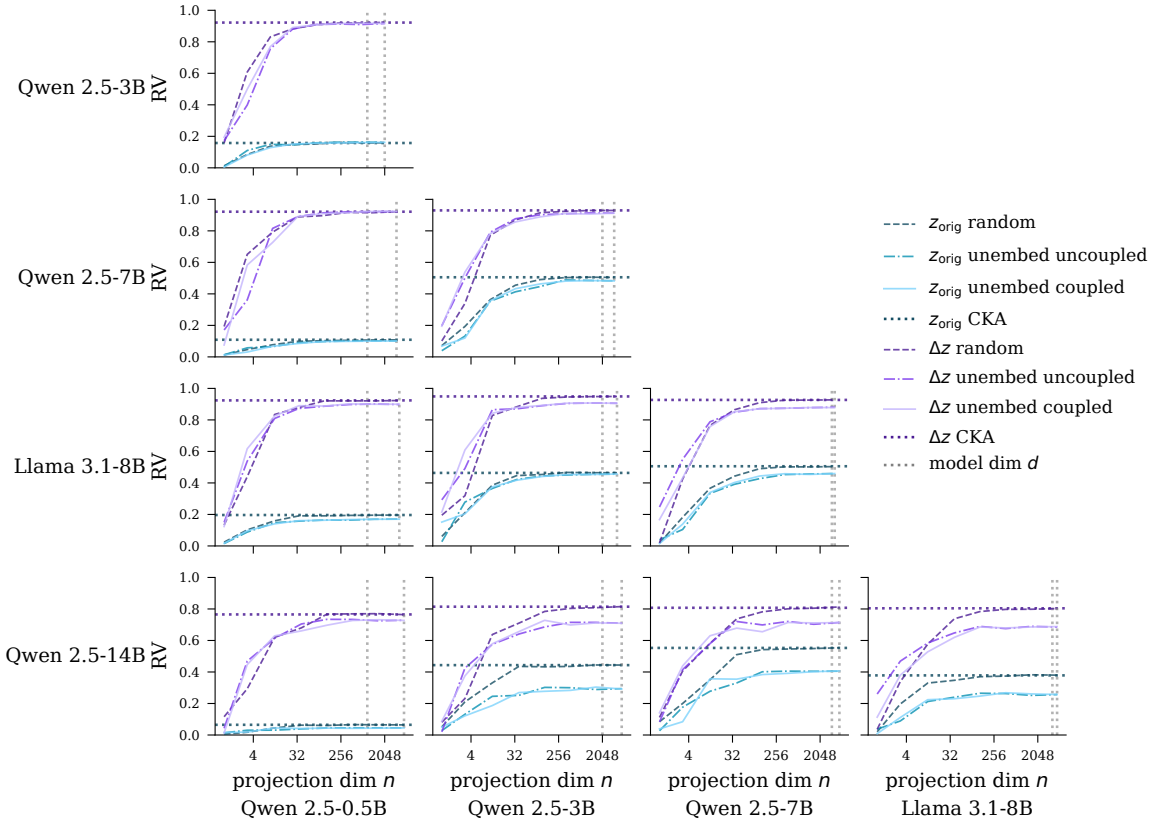


Figure 25. **RV convergence to CKA.** X-axis: projection dimension n (log scale, 1 to d). Y-axis: RV on projected data. Three curves (random Gaussian, unembeddings coupled, unembeddings uncoupled) for prediction and attribution respectively converging to horizontal dashed line (representation CKA).

Projected vs. full representational CKA. To validate that our log-odds fidelity metrics are faithful proxies for representational similarity, we compute both full CKA on the residual stream z and r^2 on binary log-odds on BoolQ for all pairs of open-source models (Qwen2.5-0.5B, 3B, 7B, 14B-Instruct and Llama-3.1-8B, yielding 10 model pairs). This is done for both prediction and attribution fidelity. Figure 24 shows the correlation between the two measures across model pairs.

Convergence of random projections. To quantify the efficiency of task-relevant projections, we compare three projection strategies as a function of projection dimension n from $n = 1$ over $n = d$ (the model's latent dimension):

1. **Random Gaussian:** project z onto n i.i.d. random unit vectors.
2. **Unembeddings coupled:** project z onto n randomly sampled tokens unembeddings (same token for M_S and M_T).

3. **Unembeddings uncoupled:** project z onto n randomly sampled tokens unembeddings (independent tokens for M_S and M_T).

For each n and each strategy, we compute RV on the n-dimensional projected data and track convergence toward full CKA. Figure 25 shows convergence curves of different model pairs. From the plots, we can see coupled unembeddings usually converge a little faster, but to a lower plateau, while only random projection can reach CKA-like levels. This aligns with our expectation that the unembeddings capture a much more relevant subspace, which is effective but limited in covering the whole representation space.

Comparing with the R^2 fidelity values shown in Figure 3, we can see the heatmap's numbers are much higher than the corresponding model pairs' curve at the same dimension $d = 1$. This shows our task-relevant log-odds is an effective approach to estimate models' representational agreement.