Do Language Models Dream of Answers before They Speak?

Anonymous authors
Paper under double-blind review

Abstract

At the core of language model intelligence lies the formation of task-relevant representations, internal encodings of the information underlying the model's inferential outcomes. However, the representational dynamics governing the transformation of latent states into model outputs remain insufficiently characterized. We study the evolution dynamics of task-relevant representations during inference and observe that *a task-relevant direction supporting highly accurate readout that generalizes across datasets emerges abruptly at an intermediate model layer*. To characterize the geometry underlying this readout, we develop and empirically validate a theoretical framework grounded in real-world language models, accounting for interactions among components and deriving empirically verifiable criteria for the direction's spectral dominance. Furthermore, we substantiate the readout's effectiveness and semantic alignment with the model's native answer preferences through visual evidence reranking experiments for retrieval-augmented generation. By connecting the internal organization of task-relevant representations to model outputs, our work provides a principled basis for interpreting representational dynamics and applying internal model judgments in real-world systems.

1 Introduction

Decision formation is a fundamental component of intelligent reasoning, underpinning a broad range of studies of reasoning dynamics (Wu et al., 2026a), uncertainty calibration (Xiao et al., 2026), and inference-time intervention (Rimsky et al., 2024). Across these settings, a shared premise is that intermediate computation gives rise to an internal state that reflects the model's evolving decision, typically modeled as a latent variable and inferred indirectly from behavior or intermediate representations (Belrose et al.; Jiang et al., 2024; Feng et al., 2025). Yet this presumed internal state remains largely unsubstantiated, with limited understanding of its existence and representational form (Makelov et al., 2024; Friedman et al., 2024). The resulting ambiguity creates a critical vulnerability for safety-oriented monitoring and steering, where repeated intervention on an incorrect internal target can compound into persistent or unintended behavioral shifts (Braun et al., 2025; Korznikov et al., 2025; Xiong et al., 2026). Resolving this ambiguity requires an explicit representational account of decision formation (Marks & Tegmark, 2023; Lugoloobi et al., 2026), which motivates a fundamental question: *Does an emerging decision acquire a stable internal organization prior to its expression in model behavior?*

Existing interpretability approaches characterize internal decision-related computation from several complementary perspectives, spanning accessibility, functional effects, and information dynamics (Feng et al., 2025; Rimsky et al., 2024; Marks et al., 2025). Probing methods instantiate the presumed latent decision state as a decodable signal in intermediate activations (Orgad et al., 2025; Kossen et al., 2024). Yet linear decodability alone does not establish geometric dominance (Park et al., 2024; Friedman et al., 2024), as the recovered signal may occupy only a weak or peripheral direction in the underlying space. Task-vector methods further compress task-relevant information into explicit latent directions (Han et al., 2025), but the existence of a compact vector does not imply that the model's evolving decision is intrinsically organized around that direction (Yang et al., 2025; Dong et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2026). Information-flow and attribution methods trace the propagation and aggregation of decision-relevant signals, illuminating their transmission (Marks et al., 2025; Kharlapenko et al., 2025) while leaving the final geometric organization of the emerging decision unresolved. Despite these complementary advances, the intrinsic geometric organization of the emerging decision remains unclear, particularly its consolidation around a dominant internal axis (Park et al., 2024; Yang et al., 2025).

To address this question, we develop a spectral framework for characterizing the geometric organization of decision-related variation in intermediate representations. The framework is grounded in a fundamental geometric principle: coherent organization of an emerging decision around an internal axis should manifest as concentrated variation along a dominant direction, with reduced dispersion across competing directions. Building on this principle, we construct a task-conditioned decomposition of intermediate computation that separates decision-aligned organization from competing representational variation and formalizes the conditions required for the former to become globally dominant. This formulation yields an operational criterion for identifying a decision axis from the intrinsic geometry of internal states and enables its formation to be tracked across layers and computational stages.

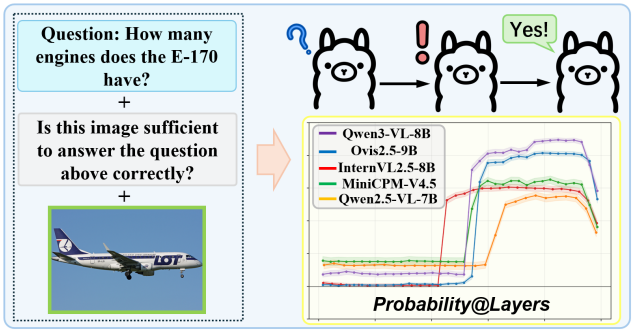

Figure 1: **Abrupt geometric consolidation during decision formation.** Given a question and candidate visual evidence, task-relevant information remains weak and distributed across early computation, followed by a sharp transition at a localized stage. Across multiple vision-language models, this transition marks the rapid consolidation of decision-related variation into a dominant internal direction, motivating our formulation of the *decision axis* as the geometric form of an emerging decision prior to behavioral expression.

Across models, we uncover a distinct geometric transition during inference, in which initially distributed decision-related variation rapidly consolidates around a dominant internal direction. We term this emergent direction the *decision axis*, as shown in Figure 1. This abrupt transition suggests that decision formation is not merely reflected by progressively stronger decodability, but instead involves a localized reorganization of the model's internal geometry. Orthogonal decomposition further shows that task-relevant differences are disproportionately concentrated along the decision axis relative to its share of overall representational variation, distinguishing it from both generic high-variance directions and weak task-informative signals. The decision axis also closely tracks the model's eventual behavioral decision, indicating that it captures a privileged internal state of decision formation rather than an incidental mode of variation. These findings suggest that *decision formation acquires a concrete geometric form through an abrupt and localized reorganization of internal computation, with task-relevant variation preferentially consolidating into a dominant* decision axis *aligned with subsequent behavior.* Our contributions are summarized as follows:

- **Problem Formulation.** We identify the representational ambiguity surrounding the latent decision state and formulate decision formation as a problem of internal geometric organization in representation space. This formulation enables direct characterization of the internal process underlying decision formation, moving the analysis beyond proxy-based inference.

- **Geometric Consolidation of Decision Formation.** We introduce a theoretically grounded spectral framework for characterizing the emergence of decision geometry, distinguishing decision-aligned organization from competing internal variation and tracking its evolution across computation. We reveal that decision formation undergoes an abrupt and localized geometric transition, with distributed decision-related variation consolidating into a dominant *decision axis* aligned with subsequent behavior.

- **Practical Transferability of Decision Axis.** We demonstrate that the decision axis transfers directly to downstream decision settings as an internal task-relevance signal without additional task-specific modeling. This direct transfer yields consistent performance gains across settings, validating the decision axis as an effective internal representation of task relevance.

2 Related Work

Representation-Level Interpretability and Task Encoding. Representation-level interpretability examines how task information is encoded and accessed in hidden states (Zhao et al., 2024; Luo et al., 2024; Marro, 2026; Wehner et al., 2025; Bartoszcze et al., 2025). Probing and intermediate decoding recover latent attributes or predictions from frozen activations (Maiya et al., 2025; Sun et al., 2025; Ni et al., 2026; Waldis et al., 2026; Luo et al., 2024; Marro, 2026), with Patchscopes decoding transplanted states in an inspection context (Ghandeharioun et al., 2024). Task-vector methods support transfer or steering through compact task encodings (Todd et al., 2024; Liu et al., 2024; Tikhonov et al., 2026; Zheng et al., 2026; Wehner et al., 2025; Bartoszcze et al., 2025), such as causally optimized combinations of attention-head representations learned from demonstrations (Saglam et al., 2025). Beyond decodability and compact task encoding, we characterize the abrupt emergence of a spectrally dominant decision axis.

Mechanistic Interpretability and Causal Computation. Explaining model behavior through component interactions is central to mechanistic interpretability (Rai et al., 2024; Ranaldi, 2025; Lin et al., 2025; Mueller et al., 2026; Kowalska & Kwaśnicka, 2026; Cai et al., 2026). Predictive pathways are reconstructed through information-flow and circuit analysis (Hsu et al., 2025; Mueller et al., 2025; Rai et al., 2024; Lin et al., 2025), exemplified by Information Flow Routes' single-pass token-level attribution graphs (Ferrando & Voita, 2024), while intervention methods quantify component effects (Zhang & Nanda, 2024; Teplica et al., 2025; Zhang & Wang, 2026; Mueller et al., 2026; Cai et al., 2026), as in joint causal tracing of attention heads and MLP neurons (Yan et al., 2026). Our framework links these component computations to global decision geometry, deriving sufficient conditions for spectral dominance and alignment from exact value, routing, and interaction contributions.

3 Theoretical Foundations of Decision Geometry

We study when the dominant direction of attention covariance aligns with a specified reference decision axis, and how head–token contributions provide a sufficient certificate. Three quantities govern this guarantee: variance along the axis, the strongest orthogonal variance, and their covariance coupling. Heads need not align individually or act independently.

3.1 Directional variance and the spectral guarantee

Setup and Notations. We consider a decoder-only Transformer with fixed parameters and analyze a specified output representation at a fixed layer and token position. Let $\mathbf{A}_i \in \mathbb{R}^d$ denote the attention output after output projection and before residual addition. The task conditions are encoded by a balanced task label $\tau_i \in \{-0.5, 0.5\}$ and $\mathbb{E}[\tau_i] = 0$.All expectations are taken under a common analysis distribution. For empirical analyses, these expectations are evaluated using a common set of nonnegative weights summing to one. Fix a unit reference decision axis $\mathbf{u}_T$, held constant across samples, and let $\mathbf{P}_T^{\perp} = \mathbf{I} - \mathbf{u}_T\mathbf{u}_T^{\top}$. For $\widetilde{\mathbf{A}}_i = \mathbf{A}_i - \mathbb{E}[\mathbf{A}_i]$, define Let $\mathbf{\Sigma}_A = \mathbb{E}[\widetilde{\mathbf{A}}_i\widetilde{\mathbf{A}}_i^{\top}]$ denote the attention covariance. Define the regression coefficient and the corresponding remainder as:

$$\mathbf{d}_A = \frac{\mathbb{E}[\tau_i\widetilde{\mathbf{A}}_i]}{\sigma_\tau^2}, \qquad \boldsymbol{\varepsilon}_{A,i} = \widetilde{\mathbf{A}}_i - \tau_i\mathbf{d}_A. \tag{3.1}$$

The coefficient decomposes along and orthogonal to the reference decision axis as $\mathbf{d}_A = \Gamma_A\mathbf{u}_T + \mathbf{e}_A^{\perp}$, where $\Gamma_A = \mathbf{u}_T^{\top}\mathbf{d}_A$ and $\mathbf{e}_A^{\perp} = \mathbf{P}_T^{\perp}\mathbf{d}_A$. Since $\mathbb{E}[\tau_i\boldsymbol{\varepsilon}_{A,i}] = \mathbf{0}$ (Lemma A.1), yielding the exact covariance decomposition:

$$\mathbf{\Sigma}_A = \sigma_\tau^2\mathbf{d}_A\mathbf{d}_A^{\top} + \mathbf{\Sigma}_{\varepsilon_A}, \tag{3.2}$$

where $\mathbf{\Sigma}_{\varepsilon_A} = \mathbb{E}[\boldsymbol{\varepsilon}_{A,i}\boldsymbol{\varepsilon}_{A,i}^\top]$.

Definition 3.1 (Directional variances and coupling)**.** *Define*

$$\begin{aligned} q_\varepsilon &= \mathbf{u}_T^\top \mathbf{\Sigma}_{\varepsilon_A} \mathbf{u}_T, \quad \nu_\varepsilon = \lambda_{\max}(\mathbf{P}_T^\perp \mathbf{\Sigma}_{\varepsilon_A} \mathbf{P}_T^\perp), \quad V_\parallel = \sigma_\tau^2 \Gamma_A^2 + q_\varepsilon, \\ V_\perp &= \lambda_{\max}(\mathbf{P}_T^\perp \mathbf{\Sigma}_A \mathbf{P}_T^\perp), \quad D_A = V_\parallel - V_\perp, \quad \mathrm{Mix} = \|\mathbf{P}_T^\perp \mathbf{\Sigma}_A \mathbf{u}_T\|_2. \end{aligned} \tag{3.3}$$

Here, q_ε and $V_\parallel$ are the remainder and total variances along the reference decision axis, while ν_ε and $V_\perp$ are their respective maximal orthogonal variances. The margin D_A measures the reference decision axis's variance advantage, and Mix *measures its covariance coupling with the orthogonal complement.*

By definition, a positive D_A guarantees a positive eigengap, while a smaller ratio Mix/D_A provides a stronger alignment guarantee. Let $\lambda_1 \geq \lambda_2 \geq \cdots$ denote the eigenvalues of $\mathbf{\Sigma}_A$, and let $\mathbf{v}_1$ be a unit eigenvector associated with λ_1.

Theorem 3.2 (Spectral dominance and alignment)**.** *Under the setup above, define the residual-aware margin*

$$M_{\mathrm{RA}} = \sigma_\tau^2\big(\Gamma_A^2 - \|\mathbf{e}_A^\perp\|_2^2\big) + q_\varepsilon - \nu_\varepsilon. \tag{3.4}$$

Then $D_A \geq M_{\mathrm{RA}}$. If $D_A > 0$, $\mathbf{v}_1$ is unique up to sign, and

$$\lambda_1 - \lambda_2 \geq D_A, \qquad \mathrm{Align}(\mathbf{v}_1, \mathbf{u}_T) = (\mathbf{v}_1^\top \mathbf{u}_T)^2 \geq \frac{1}{2}\left(1 + \frac{D_A}{\sqrt{D_A^2 + 4\,\mathrm{Mix}^2}}\right). \tag{3.5}$$

In particular, $M_{\mathrm{RA}} > 0$ is sufficient for positive dominance and a positive eigengap.

The theorem guarantees a unique leading direction under reference-decision axis variance dominance, with close alignment when coupling is small relative to this advantage. Appendix A.2.1 derives the eigengap and alignment bounds from Rayleigh quotients and projected eigenvector equations, respectively.

Corollary 3.3 (Pairwise readout stability)**.** *Let $\mathbf{u}_T, \mathbf{v}$ be unit vectors with $(\mathbf{v}^\top \mathbf{u}_T)^2 \geq a$, where $0 < a \leq 1$. Orient $\mathbf{v}$ so that $\mathbf{v}^\top \mathbf{u}_T \geq 0$, and set $\delta_a = \sqrt{2 - 2\sqrt{a}}$. For every $\mathbf{z}, \mathbf{z}' \in \mathbb{R}^d$,*

$$\big|\mathbf{v}^\top(\mathbf{z} - \mathbf{z}') - \mathbf{u}_T^\top(\mathbf{z} - \mathbf{z}')\big| \leq \delta_a \|\mathbf{z} - \mathbf{z}'\|_2. \tag{3.6}$$

Hence a reference score difference $\mathbf{u}_T^\top(\mathbf{z} - \mathbf{z}') > \delta_a \|\mathbf{z} - \mathbf{z}'\|_2$ guarantees $\mathbf{v}^\top \mathbf{z} > \mathbf{v}^\top \mathbf{z}'$.

Corollary 3.3 bounds the discrepancy between projections onto the decision-relevant direction and the reference axis, justifying the subsequent readout as an approximation to reference-axis scores and providing a conditional theoretical basis for downstream ranking. Appendix A.2.2 gives the full proof.

3.2 Tracing the certificate to attention contributions

We express the attention output as $\mathbf{A}_i = \sum_{a \in \mathcal{A}} \alpha_{i,a} \mathbf{y}_{i,a}$, where $\mathcal{A}$ is a fixed, finite index set comprising all query-head–token pairs $a = (h, t)$. $\alpha_{i,a} \geq 0$ is the inference weight and $\mathbf{y}_{i,a} \in \mathbb{R}^d$ is the value after the fixed head output projection. For $\bar{\alpha}_a = \mathbb{E}[\alpha_{i,a}]$, $\bar{\mathbf{y}}_a = \mathbb{E}[\mathbf{y}_{i,a}]$, $\delta\alpha_{i,a} = \alpha_{i,a} - \bar{\alpha}_a$, and $\delta\mathbf{y}_{i,a} = \mathbf{y}_{i,a} - \bar{\mathbf{y}}_a$, define:

$$\mathbf{d}_a^V = \bar{\alpha}_a \frac{\mathbb{E}[\tau_i \delta\mathbf{y}_{i,a}]}{\sigma_\tau^2}, \quad \mathbf{d}_a^R = \frac{\mathbb{E}[\tau_i \delta\alpha_{i,a}]}{\sigma_\tau^2} \bar{\mathbf{y}}_a, \quad \mathbf{d}_a^X = \frac{\mathbb{E}[\tau_i \delta\alpha_{i,a} \delta\mathbf{y}_{i,a}]}{\sigma_\tau^2}. \tag{3.7}$$

Proposition 3.4 (Exact attention attribution)**.** *Under the centered task encoding, the regression coefficient admits the exact decomposition*

$$\mathbf{d}_A = \sum_{a \in \mathcal{A}} (\mathbf{d}_a^V + \mathbf{d}_a^R + \mathbf{d}_a^X). \tag{3.8}$$

Proposition 3.4 decomposes $\mathbf{d}_A$ into value, routing, and interaction contributions, providing a basis for subsequent bounds on its reference-decision axis and orthogonal components. Appendix A.3.1 derives this decomposition by product expansion and interprets the interaction contribution in terms of conditional covariances.

In terms of directional quantities of the full attention output to support the spectral certificate, the preceding decomposition in Proposition 3.4 requires further reformulation. The regression coefficients and remainders are therefore aggregated over a complete block partition, with all cross-block covariance terms retained detailed in Appendix A.4–A.5. The resulting expressions connect the individual attention contributions to the aggregate quantities bounded in the following corollary as:

Corollary 3.5 (Certificate from component bounds)**.** *Suppose* $G_0 > L_0 \geq 0$ *and*

$$\Gamma_A \geq G_0, \qquad \|\mathbf{e}_A^{\perp}\|_2 \leq L_0, \qquad D_\varepsilon := q_\varepsilon - \nu_\varepsilon \geq R_0 > 0. \tag{3.9}$$

Then the component margin satisfies

$$D_0 = \sigma_\tau^2(G_0^2 - L_0^2) + R_0 > 0. \tag{3.10}$$

Consequently, $D_A \geq M_{\mathrm{RA}} \geq D_0 > 0$ *and* $\lambda_1 - \lambda_2 \geq D_0 > 0$*, so* $\mathbf{v}_1$ *is unique up to sign. If upper bound* $\mathrm{Mix} \leq B_0$*,* $B_0 \geq 0$*, is available, then*

$$\mathrm{Align}(\mathbf{v}_1, \mathbf{u}_T) \geq \frac{1}{2}\left(1 + \frac{D_0}{\sqrt{D_0^2 + 4B_0^2}}\right). \tag{3.11}$$

Corollary 3.5 guarantees a unique decision-relevant direction $\mathbf{v}_1$ of the attention covariance with its alignment to the reference decision axis $\mathbf{u}_T$ under the stated component assumptions and coupling bound and Appendix A.2.3 provides the proof. Together with the exact attribution, this result connects the spectral guarantee to attention computations, establishing sufficient conditions under which the combined value, routing, and interaction contributions support decision-axis variance dominance and principal-direction alignment.

4 Decision-Axis Identification and Mechanistic Tracing

4.1 Specification of Fundamental Theoretical Quantities

We partition the benchmark at the question level into fit, development, and test splits in a ratio of $10 : 1 : 9$. The three splits are allocated to direction estimation, head selection, and assumption validation respectively, thereby mitigating post-selection bias and enabling conditional inference with independently specified directions and head subsets. For each VLM layer, let $\mathbf{A}_i$ denote the projected attention output at the final prompt position of fit-split sample i, before residual addition. We estimate the reference decision axis as the unit direction maximizing the projected mean contrast between task conditions on the fit split: $\widehat{\mathbf{u}}_T = \frac{\widehat{\mathbb{E}}_{\mathrm{fit}}[\tau_i \mathbf{A}_i]}{\left\|\widehat{\mathbb{E}}_{\mathrm{fit}}[\tau_i \mathbf{A}_i]\right\|_2}$, where $\widehat{\mathbb{E}}_{\mathrm{fit}}$ denotes the empirical expectation on the fit split, with equal total weight assigned to each task condition. The decision-relevant direction $\widehat{\mathbf{v}}_1$ is defined as the unit leading eigenvector of the fit-split attention covariance as

$$\widehat{\mathbf{v}}_1 \in \underset{\|\mathbf{v}\|_2=1}{\arg\max}\, \widehat{\mathbb{E}}_{\mathrm{fit}}\left[\left(\mathbf{v}^\top\left(\mathbf{A}_i - \widehat{\mathbb{E}}_{\mathrm{fit}}[\mathbf{A}_i]\right)\right)^2\right], \tag{4.1}$$

with its sign chosen such that $\widehat{\mathbf{v}}_1^\top \widehat{\mathbf{u}}_T \geq 0$. Implementation details for preprocessing and direction estimation are provided in Appendix C.1.

We further select a head subset $\mathcal{H}_\star$ on the development split, comprising heads with the strongest influence on the decision readout. The selection is based on the magnitude of aggregate readout effects induced by matched cross-condition head replacement, with the full procedure detailed in Appendix C.1. Conditional on these fixed specifications, the remaining statistical quantities are estimated on the test split in accordance with Section 3, enabling empirical evaluation of the theoretical assumptions and associated spectral certificates.

4.2 Decision-Relevant Representation Readout

To bridge the gap between theoretical quantities and downstream tasks, we introduce a decision-relevant representation g via projection onto $\widehat{\mathbf{v}}_1$, grounded in the theoretically guaranteed alignment between $\mathbf{v}_1$ and $\mathbf{u}_T$ under the stated assumptions. Restoring the layer

index ℓ, we define the readout for sample i as the projection of the attention output, centered by its fit-split mean, onto the corresponding decision-relevant direction:

$$g_\ell(i) = \widehat{\mathbf{v}}_{1,\ell}^\top \left(\mathbf{A}_{\ell,i} - \widehat{\mathbb{E}}_{\text{fit}}[\mathbf{A}_{\ell,i}] \right). \tag{4.2}$$

which approximates the reference-axis score with the approximation error bounded by the product of the directional discrepancy and the norm of the centered attention output. To elucidate the formation and propagation of decision-relevant information, we extend this readout formulation to the outputs of individual components within each layer. Component-specific activation extraction and readout procedures are detailed in Appendix C.2.

4.3 Operationalizing Representation Readouts for Downstream Tasks

To substantiate both decision fidelity and practical utility, we evaluate representation-level readouts and the output-level probability estimates derived from them as relevance scores for visual evidence reranking in retrieval-augmented generation (RAG), which quantifies the relevance of each image–query pair to rerank candidate images as supporting evidence for accurate question answering. For each query q, let $\mathcal{I}_q = \{I_{q,j}\}_{j=1}^{n_q}$ denote its retrieved candidate images. The relevance score of each image–query pair determines a ranking permutation π_q satisfying

$$s_{q,j} = s(q, I_{q,j}), \qquad s_{q,\pi_q(1)} \geq \cdots \geq s_{q,\pi_q(n_q)}. \tag{4.3}$$

The relevance score s is derived from the VLM's native assessment of evidential sufficiency through standard multimodal inference, with the evaluation protocol detailed in Appendix C.3. For representation-level reranking, the relevance score $s = g_\ell(i)$ is obtained by projecting the centered attention output at the layer exhibiting the abrupt transition onto the corresponding decision-relevant direction. The layer-localization procedure is detailed in Appendix C.4. For output-level reranking, logistic regression maps $g_\ell(i)$ to the estimated probability of an affirmative relevance judgment by the VLM:

$$\widehat{p}_\ell(i) = \sigma(a_\ell + b_\ell g_\ell(i)), \tag{4.4}$$

where σ denotes the logistic sigmoid. Probability fitting and calibration-based layer selection are detailed in Appendix C.5. The estimated probability $\widehat{p}_\ell(i)$ serves as the image–query relevance score s. This formulation grounds intermediate representations in real-world tasks, enabling systematic assessment of their correspondence with native model decisions and their utility for downstream task performance.

5 Experiments

5.1 Experimental Setup

Benchmarks. We conduct experiments on four benchmarks: CUB-200-2011 (Wah), FGVC-Aircraft (Maji et al., 2013), Stanford Cars (Krause et al., 2013), and Visual-RAG (Wu et al., 2026b) and construct CAC from the first three with three levels of distractor difficulty and partition its question instances into fit, development, and test splits in an ratio of $10:1:9$. For Visual-RAG, we construct reranking pools for individual queries and partition the queries into test and calibration splits in a $4:1$ ratio. Detailed preparation procedures are provided in Appendix B.

Models. Our experiments cover five families of state-of-the-art open-source MLLMs, including Qwen3-VL-8B (Bai et al., 2025a), Ovis2.5-9B (Lu et al., 2025), MiniCPM-V-4.5 (Yu et al., 2025), Qwen2.5-VL-7B (Bai et al., 2025b) and InternVL2.5-8B (Chen et al., 2025), spanning multiple architectures and parameter scales.

Evaluation Metrics. We evaluate reranking performance using standard retrieval metrics, including mAP, nDCG, MRR, Recall, and AUROC assessing relevance discrimination. Detailed metric definitions and computation procedures are provided in the Appendix D.1.

5.2 Functional Validation of Decision-Relevant Directions

Table 1: **Visual evidence reranking on Visual-RAG.** Readouts g_k estimated on CAC are compared with baselines, including logit lens at the same transition layers k. mAP, nDCG@10, and recall are percentages; brackets give 95% bootstrap confidence intervals for mAP. Bold indicates the best result per column.

Method	mAP [95% CI]	nDCG@10	MRR	R@5	R@10	AUROC
CLIP ViT-L/14-336	21.9 [19.8, 24.2]	23.0	0.367	12.4	20.0	0.613
GME-Qwen2-VL-2B	24.0 [21.5, 26.6]	24.4	0.344	13.5	23.1	0.634
MM-Embed	24.1 [21.6, 26.4]	26.4	0.407	14.5	23.9	0.639
GME-Qwen2-VL-7B	25.1 [22.6, 27.8]	26.3	0.381	13.2	23.2	0.631
VLM2Vec-Full	27.5 [24.8, 30.4]	29.5	0.439	15.3	24.1	0.662
BGE-VL-large	29.9 [26.8, 32.9]	31.4	0.440	18.0	27.2	0.665
UniME-V2	30.5 [27.5, 33.7]	33.2	0.453	18.5	29.3	0.680
SigLIP 2	30.8 [28.0, 33.8]	33.5	0.452	19.4	30.5	0.695
LamRA-Rank	36.1 [33.0, 39.3]	41.8	0.544	25.0	37.6	0.742
LogitLens_{17}, Qwen2.5-VL-7B	13.2 [11.7, 14.8]	10.3	0.175	5.4	10.4	0.446
LogitLens_{19}, MiniCPM-V-4.5	21.8 [19.3, 24.5]	22.4	0.333	11.4	19.6	0.605
LogitLens_{14}, InternVL2.5-8B	31.2 [28.3, 34.3]	33.8	0.436	18.9	32.6	0.717
LogitLens_{19}, Qwen3-VL-8B	31.8 [28.5, 35.1]	34.1	0.465	18.8	30.1	0.673
LogitLens_{20}, Ovis2.5-9B	45.8 [42.2, 49.4]	49.7	0.618	30.4	44.2	0.796
g_{14}, InternVL2.5-8B	39.4 [36.0, 42.9]	43.0	0.553	26.1	38.7	0.749
g_{17}, Qwen2.5-VL-7B	40.5 [37.2, 43.7]	45.0	0.569	28.7	42.1	0.778
g_{19}, MiniCPM-V-4.5	44.1 [40.6, 47.5]	48.0	0.593	29.7	42.5	0.783
g_{19}, Qwen3-VL-8B	48.6 [45.0, 52.1]	53.8	0.643	33.1	49.2	0.806
g_{20}, Ovis2.5-9B	**49.9 [46.5, 53.2]**	**55.6**	**0.680**	**35.1**	**50.9**	**0.823**

To assess whether the observed signal encodes task-relevant information, we conduct functional validation as described in Section 4.3. Table 1 compares signal-based reranking using five VLMs with nine state-of-the-art methods on Visual-RAG benchmarks across multiple metrics. We also evaluate logit lens at the corresponding transition layers, contrasting direct intermediate decoding through the native output mapping with readout along a decision-relevant axis identified from standardized representation covariance. Across evaluations, signal-based reranking achieves competitive or superior performance, with Ovis2.5-9B attaining 49.9% mAP versus 36.1% for LamRA-Rank, the strongest baseline. Although estimated on our CAC benchmark, the empirical decision-relevant directions remain effective on Visual-RAG, providing evidence that they capture task-relevant rather than incidental statistical structure. Our readout also substantially outperforms logit lens, supporting the downstream utility of a decision axis characterized by intrinsic geometric dominance beyond decodability alone. Effective visual evidence selection across benchmarks demonstrates the direction's practical utility and generalizability, functionally validating its interpretation and generality as a decision-relevant readout direction.

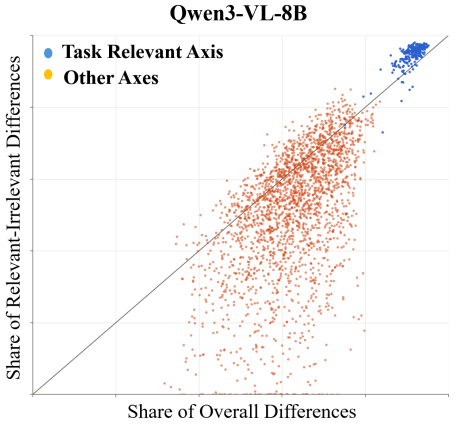

Figure 2: **Concentration of task-relevant variation along the decision axis.** CAC question–axis pairs compare variance share with relevance-separation share for v_1 (blue) and others (orange).

We further assess the direction's task relevance and uniqueness through orthogonal decomposition of representations, comparing its contribution to relevance separation with those of other principal directions. In Figure 2, each point represents one axis evaluated over all candidate images paired with a CAC question. Larger x- and y-values indicate greater contributions to overall candidate variation and separation between relevant and irrelevant candidates, respectively. Points above the diagonal indicate stronger concentration of

decision-related differences than overall candidate variation. Points corresponding to the decision-relevant direction cluster in the upper-right region, predominantly above the diagonal, indicating that it captures substantial overall variation while concentrating relevance-related differences, supporting its observed readout effectiveness. Additional results appear in Appendix E, with the full procedure and results in Appendix D.2. Together, these findings establish the readout axis as a privileged, transferable representation of task relevance, providing direct and reliable access to task information encoded within the model.

5.3 Architectural Localization of Decision Readout Emergence

We further examine individual components within the abrupt transition layer to localize the emergence of the decision-relevant readout at the architectural level and delineate the structural scope of our theoretical analysis. Specifically, we quantify component-wise variation in task performance and agreement with native VLM decisions using the representation-level readouts and their derived output-level probability estimates respectively in Visual-RAG benchmark. Further details are provided in Appendices C.2 and D.3. As shown in Table 2, the decision-relevant readout jumps from near-chance performance to approximately the level attained at the block output at A_k, reaching mean AUROC and output-token agreement of 0.79 and 0.67, respectively. Meanwhile, we present Figure 3 to provide a visual illustration. More results are detailed in Appendix E. These results show that the attention update itself supports decision-relevant readout, localizing the layer-level emergence within attention and providing an empirical basis for our theoretical analysis of its spectral structure. These findings localize both the onset of the readout transition and the predominant gain in readout performance to the attention update within the identified jump layer, delineating the structural scope of our theoretical framework and demonstrating that the attention output and subsequent representations already support decision-relevant readout. The profile construction is detailed in Appendix D.4.

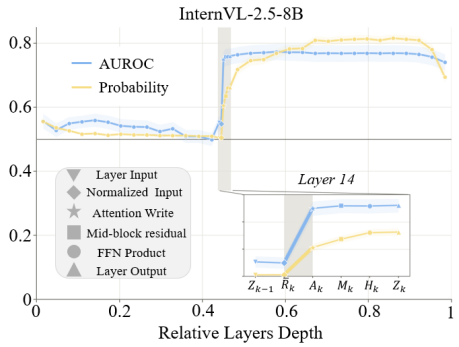

Figure 3: **Readout emergence within attention.** AUROC (blue) and output-token agreement (yellow) rise at A_{14} in InternVL2.5-8B; shading denotes pointwise 95% confidence intervals.

5.4 Theoretical Validation

To empirically ground our framework, we systematically assess spectral dominance conditions and certificates alongside quantitative alignment guarantees between the relevant task direction and reference decision axis. We first examine empirical assumptions and spectral certificates specifying sufficient conditions for variance dominance and the resulting hierarchy of spectral margins. We estimate these quantities on CAC, testing positivity with simultaneous one-sided 95% bootstrap lower confidence bounds (LCBs) from resampling question clusters within each dataset. Table 3 reports strictly positive simultaneous LCBs for both checks and certificates across all evaluated model families, with certificate point estimates preserving the theoretical ordering. These results demonstrate that real-world representations jointly satisfy the sufficient conditions and admit non-vacuous spectral guarantees, extending empirical support from premises to predicted consequences. Appendix D.5 details the bootstrap procedure.

We further assess the central geometric conclusion through the quantitative alignment guarantee between the decision-relevant direction v_1 and reference decision axis u_T. Following this guarantee, we derive its empirical counterpart from spectral estimates rather than direct similarity measurements and evaluate it against the stringent threshold $a^\star = 0.80$, using LCBs to assess statistical robustness. The theoretical bounds indicate consistently high alignment across models, providing conclusion-level validation of the assumptions and certificates and a principled explanation for the strong observed directional correspondence.

Table 2: **Architectural localization of decision readout emergence.** Stagewise readouts at transition layer k are evaluated on Visual-RAG using relevance AUROC and output-token agreement. Agreement is the mean readout-derived probability assigned to the model's native decision. Brackets give 95% bootstrap confidence intervals.

Model	k	Z_{k-1}	$\bar{R}_k$	A_k	M_k	Z_k
Task Level (AUROC)						
Qwen3-VL-8B	19	0.59 [0.57, 0.61]	0.59 [0.57, 0.61]	0.81 [0.78, 0.83]	0.81 [0.79, 0.84]	0.81 [0.78, 0.83]
Qwen2.5-VL-7B	17	0.52 [0.49, 0.54]	0.52 [0.49, 0.54]	0.78 [0.75, 0.80]	0.56 [0.53, 0.58]	0.63 [0.61, 0.65]
InternVL2.5-8B	14	0.55 [0.53, 0.57]	0.55 [0.53, 0.57]	0.75 [0.72, 0.77]	0.76 [0.73, 0.78]	0.76 [0.73, 0.78]
MiniCPM-V-4.5	19	0.51 [0.49, 0.53]	0.50 [0.48, 0.52]	0.78 [0.76, 0.81]	0.73 [0.70, 0.75]	0.79 [0.77, 0.81]
Ovis2.5-9B	20	0.71 [0.69, 0.73]	0.70 [0.68, 0.72]	0.82 [0.80, 0.84]	0.82 [0.80, 0.84]	0.83 [0.81, 0.85]
Mean	–	**0.57**	**0.57**	**0.79**	**0.73**	**0.76**
Output level (Output-Token Agreement)						
Qwen3-VL-8B	19	0.49 [0.49, 0.49]	0.49 [0.49, 0.49]	0.76 [0.73, 0.78]	0.72 [0.70, 0.74]	0.75 [0.73, 0.78]
Qwen2.5-VL-7B	17	0.58 [0.58, 0.59]	0.58 [0.58, 0.59]	0.64 [0.63, 0.66]	0.59 [0.58, 0.59]	0.61 [0.60, 0.61]
InternVL2.5-8B	14	0.50 [0.50, 0.50]	0.51 [0.50, 0.51]	0.61 [0.59, 0.62]	0.64 [0.62, 0.65]	0.66 [0.65, 0.68]
MiniCPM-V-4.5	19	0.56 [0.56, 0.56]	0.56 [0.56, 0.56]	0.60 [0.59, 0.61]	0.58 [0.57, 0.58]	0.63 [0.62, 0.64]
Ovis2.5-9B	20	0.50 [0.49, 0.50]	0.49 [0.49, 0.50]	0.74 [0.72, 0.75]	0.57 [0.57, 0.58]	0.74 [0.73, 0.76]
Mean	–	**0.53**	**0.53**	**0.67**	**0.62**	**0.68**

Table 3: **Validation of spectral dominance and alignment guarantees.** CAC results report assumption checks, spectral certificates, and the theory-derived lower bound Align_{RA} on squared alignment. Brackets give simultaneous and unadjusted nominal one-sided 95% bootstrap lower confidence bounds (LCBs), with simultaneous correction within each model's specified endpoint families.

	Qwen3-VL-8B	Qwen2.5-VL-7B	InternVL2.5-8B	MiniCPM-V-4.5	Ovis2.5-9B
Empirical Assumptions					
A-S: G_0	2.33[2.21, 2.22]✓	1.19[1.13, 1.13]✓	0.69[0.65, 0.65]✓	3.01[2.90, 2.91]✓	1.94[1.85, 1.86]✓
A-RD: D_ε	5.92[5.21, 5.40]✓	3.07[2.62, 2.75]✓	4.32[3.80, 3.98]✓	13.95[12.86, 13.18]✓	7.62[6.86, 7.09]✓
A-M: $G_0 - L_0$	2.10[1.99, 2.05]✓	0.99[0.93, 0.98]✓	0.45[0.36, 0.44]✓	2.64[2.52, 2.61]✓	1.63[1.52, 1.59]✓
Spectral Certificates					
D_A	11.08[10.43, 10.43]✓	5.37[4.97, 4.97]✓	6.98[6.54, 6.54]✓	24.57[23.53, 23.53]✓	13.33[12.67, 12.67]✓
M_{RA}	11.06[10.22, 10.43]✓	5.36[4.83, 4.97]✓	6.97[6.34, 6.54]✓	24.54[23.15, 23.51]✓	13.31[12.41, 12.66]✓
D_0	7.26[6.53, 6.72]✓	3.41[2.96, 3.09]✓	4.42[3.91, 4.09]✓	16.18[15.09, 15.41]✓	8.54[7.79, 8.02]✓
Alignment ($a_\star$=0.80)					
Align_{RA}	0.963[0.958, 0.959]✓	0.933[0.922, 0.924]✓	0.884[0.871, 0.874]✓	0.959[0.955, 0.956]✓	0.903[0.894, 0.895]✓

Together with task-level results, this agreement grounds the observed decision-relevant axis in the framework's spectral mechanism, completing validation from downstream utility to underlying organization and supporting its explanatory completeness and empirical validity.

6 Conclusions

We introduce a theoretically grounded spectral framework for quantitatively analyzing decision formation through internal representation geometry, moving beyond proxy-based inference to directly characterize the underlying decision process. Our study reveals abrupt, localized geometric consolidation during decision formation, as distributed decision-related variation converges onto a dominant decision axis. The resulting axis closely tracks subsequent model behavior and admits mechanistic characterization through stage-wise localization and attention-level decomposition. We further demonstrate its direct transfer to downstream decision settings as an effective task-relevance signal without additional task-specific modeling. Extensive experiments across models and settings establish a geometric account of decision formation and a principled foundation for mechanistic analysis and internal monitoring of emerging model decisions.

AI USE STATEMENT

In this work, we used generative AI tools for providing feedback on research experiments and coding assistant method implementation.

We have **not** used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses, cleaning and reformat dataset, support qualitative and thematic data analysis.

Interpreting results and assisting with translation are not applicable to this work.

Specifically, coding agents assisted in implementing our theoretical formulas and estimation procedures, including decision-axis extraction, computation of spectral certificates and alignment bounds, and projection-based readouts for visual evidence reranking. They also supported baseline implementation and selection of the image–question relevance prompt.

REFERENCES

Technical report.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025a. URL https://arxiv.org/abs/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025b. URL https://arxiv.org/abs/2502.13923.

Lukasz Bartoszcze, Sarthak Munshi, Bryan Sukidi, Jennifer Yen, Zejia Yang, David Williams-King, Linh Le, Kosi Asuzu, and Carsten Maple. Representation engineering for large-language models: Survey and research challenges. *arXiv preprint arXiv:2502.17601*, 2025.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2025. *URL https://arxiv. org/abs/2303.08112*.

Joschka Braun, Carsten Eickhoff, David Krueger, Seyed Ali Bahrainian, and Dmitrii Krasheninnikov. Understanding (un) reliability of steering vectors in language models. *arXiv preprint arXiv:2505.22637*, 2025.

Jie Cai, Mafizur Rahman, James Enouen, Lijun Qian, and Yan Liu. A survey on actionable interpretability in large language models. In Diego Calvanese (ed.), *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence, IJCAI-26*, pp. 7787–7797. International Joint Conferences on Artificial Intelligence Organization, 8 2026. doi: 10.24963/ijcai.2026/865. URL https://doi.org/10.24963/ijcai.2026/865. Survey Track.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming

Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Yuxin Dong, Jiachen Jiang, Zhihui Zhu, and Xia Ning. Understanding task vectors in in-context learning: Emergence, functionality, and limitations. In *International Conference on Learning Representations*, volume 2026, pp. 19393–19428, 2026.

Jiahai Feng, Stuart Russell, and Jacob Steinhardt. Monitoring latent world states in language models with propositional probes. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 19337–19359, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3132d0405fabe24b2a7b6cd7ba9de6b5-Paper-Conference.pdf.

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17432–17445, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.965. URL https://aclanthology.org/2024.emnlp-main.965/.

Dan Friedman, Andrew Kyle Lampinen, Lucas Dixon, Danqi Chen, and Asma Ghandeharioun. Interpretability illusions in the generalization of simplified models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 14035–14059. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/friedman24a.html.

Asma Ghandeharioun, Avi Caciularu, Adam Pearce, Lucas Dixon, and Mor Geva. Patchscopes: A unifying framework for inspecting hidden representations of language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15466–15490. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghandeharioun24a.html.

Seungwook Han, Jinyeop Song, Jeff Gore, and Pulkit Agrawal. Emergence and effectiveness of task vectors in in-context learning: An encoder decoder perspective. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21871–21897. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/han25h.html.

Aliyah Hsu, Georgia Zhou, Yeshwanth Cherapanamjeri, Yaxuan Huang, Anobel Odisho, Peter Carroll, and Bin Yu. Efficient automated circuit discovery in transformers using contextual decomposition. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 57880–57898, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/916ee60e315531d6b3954af8a8dc3437-Paper-Conference.pdf.

Yibo Jiang, Goutham Rajendran, Pradeep Kumar Ravikumar, Bryon Aragam, and Victor Veitch. On the origins of linear representations in large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 21879–21911. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jiang24d.html.

Dmitrii Kharlapenko, Stepan Shabalin, Arthur Conmy, and Neel Nanda. Scaling sparse feature circuits for studying in-context learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 29819–29851. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/kharlapenko25a.html.

Anton Korznikov, Andrey Galichin, Alexey Dontsov, Oleg Y Rogov, Ivan Oseledets, and Elena Tutubalina. The rogue scalpel: Activation steering compromises llm safety. *arXiv preprint arXiv:2509.22067*, 2025.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in llms. *arXiv preprint arXiv:2406.15927*, 2024.

Bianka Kowalska and Halina Kwaśnicka. Unboxing the black box: A survey on mechanistic interpretability for algorithmic understanding of neural networks. *Machine Learning*, 115 (8):190, 2026.

Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, June 2013.

Zihao Lin, Samyadeep Basu, Mohammad Beigi, Varun Manjunatha, Ryan A Rossi, Zichao Wang, Yufan Zhou, Sriram Balasubramanian, Arman Zarei, Keivan Rezaei, et al. A survey on mechanistic interpretability for multi-modal foundation models. *arXiv preprint arXiv:2502.17516*, 2025.

Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32287–32307. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/liu24bx.html.

Shiyin Lu, Yang Li, Yu Xia, Yuwei Hu, Shanshan Zhao, Yanqing Ma, Zhichao Wei, Yinglun Li, Lunhao Duan, Jianshan Zhao, Yuxuan Han, Haijun Li, Wanying Chen, Junke Tang, Chengkun Hou, Zhixing Du, Tianli Zhou, Wenjie Zhang, Huping Ding, Jiahe Li, Wen Li, Gui Hu, Yiliang Gu, Siran Yang, Jiamang Wang, Hailong Sun, Yibo Wang, Hui Sun, Jinlong Huang, Yuping He, Shengze Shi, Weihong Zhang, Guodong Zheng, Junpeng Jiang, Sensen Gao, Yi-Feng Wu, Sijia Chen, Yuhui Chen, Qing-Guo Chen, Zhao Xu, Weihua Luo, and Kaifu Zhang. Ovis2.5 technical report, 2025. URL https://arxiv.org/abs/2508.11737.

William Lugoloobi, Thomas Foster, William Bankes, and Chris Russell. Llms encode their failures: Predicting success from pre-generation activations. *arXiv preprint arXiv:2602.09924*, 2026.

Siwen Luo, Hamish Ivison, Soyeon Caren Han, and Josiah Poon. Local interpretations for explainable natural language processing: A survey. *ACM Comput. Surv.*, 56(9), April 2024. ISSN 0360-0300. doi: 10.1145/3649450. URL https://doi.org/10.1145/3649450.

Sharan Maiya, Yinhong Liu, Ramit Debnath, and Anna Korhonen. Improving preference extraction in LLMs by identifying latent knowledge through classifying probes. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9061–9081, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.444. URL https://aclanthology.org/2025.acl-long.444/.

S. Maji, J. Kannala, E. Rahtu, M. Blaschko, and A. Vedaldi. Fine-grained visual classification of aircraft. Technical report, 2013.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 26486–26515, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/70b8505ac79e3e131756f793cd80eb8d-Paper-Conference.pdf.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. *arXiv preprint arXiv:2310.06824*, 2023.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 23888–23923, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/3ba4d47a83e498c2b1a0868cba20f6de-Paper-Conference.pdf.

Rémy Marro. Compositional meaning representations in LLMs: a critical review of probing studies. In Saif M. Mohammad and Nedjma Ousidhoum (eds.), *Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pp. 198–210, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-413-2. doi: 10.18653/v1/2026.starsem-conference.13. URL https://aclanthology.org/2026.starsem-conference.13/.

Aaron Mueller, Atticus Geiger, Sarah Wiegreffe, Dana Arad, Iván Arcuschin, Adam Belfki, Yik Siu Chan, Jaden Fried Fiotto-Kaufman, Tal Haklay, Michael Hanna, Jing Huang, Rohan Gupta, Yaniv Nikankin, Hadas Orgad, Nikhil Prakash, Anja Reusch, Aruna Sankaranarayanan, Shun Shao, Alessandro Stolfo, Martin Tutek, Amir Zur, David Bau, and Yonatan Belinkov. MIB: A mechanistic interpretability benchmark. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 45069–45108. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/mueller25a.html.

Aaron Mueller, Jannik Brinkmann, Millicent Li, Samuel Marks, Koyena Pal, Nikhil Prakash, Can Rager, Aruna Sankaranarayanan, Arnab Sen Sharma, Jiuding Sun, Eric Todd, David Bau, and Yonatan Belinkov. The quest for the right mediator: Surveying mechanistic interpretability for NLP through the lens of causal mediation analysis. *Computational Linguistics*, 52(1):331–378, March 2026. doi: 10.1162/coli.a.572. URL https://aclanthology.org/2026.cl-1.10/.

Jingwei Ni, Ekaterina Fadeeva, Tianyi Wu, Mubashara Akhtar, Jiaheng Zhang, Elliott Ash, Markus Leippold, Timothy Baldwin, See-Kiong Ng, Artem Shelmanov, and Mrinmaya Sachan. ReProbe: Efficient test-time scaling of multi-step reasoning by probing internal states of large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11667–11689, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.536. URL https://aclanthology.org/2026.acl-long.536/.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.),

Proceedings of the 41st International Conference on Machine Learning, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024.

Leonardo Ranaldi. Survey on the role of mechanistic interpretability in generative ai. *Big Data and Cognitive Computing*, 9(8), 2025. ISSN 2504-2289. doi: 10.3390/bdcc9080193. URL https://www.mdpi.com/2504-2289/9/8/193.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

Baturay Saglam, Xinyang Hu, Zhuoran Yang, Dionysis Kalogerias, and Amin Karbasi. Learning task representations from in-context learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 6634–6663, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.345. URL https://aclanthology.org/2025.findings-acl.345/.

Yucheng Sun, Alessandro Stolfo, and Mrinmaya Sachan. Probing for arithmetic errors in language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8111–8128, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.411. URL https://aclanthology.org/2025.emnlp-main.411/.

Carter Teplica, Yixin Liu, Arman Cohan, and Tim G. J. Rudner. SCIURus: Shared circuits for interpretable uncertainty representations in language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 12451–12469, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.618. URL https://aclanthology.org/2025.naacl-long.618/.

Pavel Tikhonov, Ivan Oseledets, and Elena Tutubalina. One task vector is not enough: A large-scale study for in-context learning. In Santosh T.Y.S.S., Juan Diego Rodriguez, and Ona de Gibert (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pp. 1451–1463, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-393-7. doi: 10.18653/v1/2026.acl-srw.127. URL https://aclanthology.org/2026.acl-srw.127/.

Eric Todd, Millicent Li, Arnab Sen Sharma, Aaron Mueller, Byron Wallace, and David Bau. Function vectors in large language models. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 17282–17333, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/4ae163cb8788970e53b4fd9578141139-Paper-Conference.pdf.

Andreas Waldis, Vagrant Gautam, Anne Lauscher, Dietrich Klakow, and Iryna Gurevych. Aligned probing: Relating toxic behavior and model internals. *Transactions of the Association for Computational Linguistics*, 14:271–291, 2026. doi: 10.1162/tacl.a.613. URL https://aclanthology.org/2026.tacl-1.14/.

Jan Wehner, Sahar Abdelnabi, Daniel Tan, David Krueger, and Mario Fritz. Taxonomy, opportunities, and challenges of representation engineering for large language models. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=2U1KIfmaU9. Survey Certification.

Junda Wu, Yuxin Xiong, Xintong Li, Sheldon Yu, Zhengmian Hu, Tong Yu, Rui Wang, Xiang Chen, Jingbo Shang, and Julian McAuley. Ctrls: Chain-of-thought reasoning via latent state transition. In Emtiyaz Khan, Yingzhen Li, Arno Solin, and Aaditya Ramdas (eds.), *Proceedings of The 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 2728–2736. PMLR, 02–05 May 2026a. URL https://proceedings.mlr.press/v300/wu26a.html.

Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. Visual-rag: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '26)*, 2026b. doi: 10.1145/3805712.3808615.

Zeguan Xiao, Diyang Dou, Boya Xiong, Yun Chen, and Guanhua Chen. Enhancing uncertainty estimation in llms with expectation of aggregated internal belief. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 34043–34051, 2026.

Chen Xiong, Zhiyuan He, Pin-Yu Chen, Ching-Yun Ko, and Tsung-Yi Ho. Steering externalities: Benign activation steering unintentionally increases jailbreak risk for large language models. *arXiv preprint arXiv:2602.04896*, 2026.

Zirui Yan, Dennis Wei, Dmitriy A Katz, Prasanna Sattigeri, and Ali Tajer. Multi-component causal tracing in large language models. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3398–3418, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.154. URL https://aclanthology.org/2026.acl-long.154/.

Liu Yang, Ziqian Lin, Kangwook Lee, Dimitris Papailiopoulos, and Robert Nowak. Task vectors in in-context learning: Emergence, formation, and benefit. *arXiv preprint arXiv:2501.09240*, 2025.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, Bokai Xu, Junbo Cui, Yingjing Xu, Liqing Ruan, Luoyuan Zhang, Hanyu Liu, Jingkun Tang, Hongyuan Liu, Qining Guo, Wenhao Hu, Bingxiang He, Jie Zhou, Jie Cai, Ji Qi, Zonghao Guo, Chi Chen, Guoyang Zeng, Yuxuan Li, Ganqu Cui, Ning Ding, Xu Han, Yuan Yao, Zhiyuan Liu, and Maosong Sun. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe, 2025. URL https://arxiv.org/abs/2509.18154.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/06a52a54c8ee03cd86771136bc91eb1f-Paper-Conference.pdf.

Luyang Zhang and Jialu Wang. When attribution patching lies: Diagnosis and a second-order correction. *arXiv preprint arXiv:2606.09899*, 2026.

Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2), February 2024. ISSN 2157-6904. doi: 10.1145/3639372. URL https://doi.org/10.1145/3639372.

Bowen Zheng, Ming Ma, Zhongqiao Lin, and Tianming Yang. Label words as local task vectors in in-context learning. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL*

2026, pp. 6319–6332, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.314. URL https://aclanthology.org/2026.findings-acl.314/.

A Theory and proofs for the attention spectral framework

This appendix proves every formal result in Section 3 and supplies the complete attention decomposition, moment calculations, and scope qualifications. The proof dependencies are explicit: Lemma A.1 establishes the regression covariance identity; the directional decomposition then yields Theorem 3.2 and Corollary 3.5; Proposition 3.4 and Lemma A.2 connect their quantities to head blocks. Corollary 3.3 translates an alignment bound into conditional pairwise score stability.

Main-paper statement	Complete proof
Theorem 3.2: spectral guarantee	Section A.2.1
Corollary 3.3: ranking stability	Section A.2.2
Proposition 3.4: exact attribution	Section A.3.1
Corollary 3.5: bounded certificate	Section A.2.3

A.1 Setup and task regression

Fix the model parameters, layer, and attention readout position. Let $\mathbf{A}_i \in \mathbb{R}^d$, $d \geq 2$, be the attention write after the linear output projection, with $\mathbb{E}\|\mathbf{A}_i\|_2^2 < \infty$. The index i denotes a sample. Throughout, $\mathbb{E}$ uses one fixed analysis distribution with balanced binary labels:

$$\tau_i \in \{-\tfrac{1}{2}, \tfrac{1}{2}\}, \qquad \Pr(\tau_i = \tfrac{1}{2}) = \Pr(\tau_i = -\tfrac{1}{2}) = \tfrac{1}{2}, \qquad \sigma_\tau^2 = \mathbb{E}[\tau_i^2] = \tfrac{1}{4}. \tag{A.1}$$

Alternatively, $\mathbb{E}$ may be an empirical average with fixed nonnegative weights that sum to one and assign total weight $1/2$ to each condition. Within either condition, these weights are renormalized by its mass $1/2$ when computing conditional moments. Every mean, coefficient, and covariance must use these same weights; in particular, no separate sample-covariance correction is applied. Independence of samples is not needed for the algebra below.

Fix a unit task axis $\mathbf{u}_T$, and write $\mathbf{P}_T^\perp = \mathbf{I} - \mathbf{u}_T\mathbf{u}_T^\top$. The axis is held constant in every expectation below. For an axis learned from independent training data, the population statements may be read conditional on that fitted design, with the balanced evaluation distribution in equation A.1. The axis, its orientation, the representation, and all selection rules must be specified for the analysis. Defining this axis to be the PC1 whose alignment is being explained would make the conclusion circular. Vectors are columns, $^\top$ denotes transpose, and $\|\cdot\|_2$ denotes the Euclidean norm. All covariance matrices and orthogonal projections in the spectral analysis act in the same output space and use the same Euclidean metric.

Define the centered write, its covariance, and its task coefficient by

$$\begin{aligned} \widetilde{\mathbf{A}}_i &= \mathbf{A}_i - \mathbb{E}[\mathbf{A}_i], \quad \mathbf{\Sigma}_A = \mathbb{E}[\widetilde{\mathbf{A}}_i\widetilde{\mathbf{A}}_i^\top], \\ \mathbf{d}_A &= \frac{\mathbb{E}[\tau_i\widetilde{\mathbf{A}}_i]}{\sigma_\tau^2}, \qquad \boldsymbol{\varepsilon}_{A,i} = \widetilde{\mathbf{A}}_i - \tau_i\mathbf{d}_A. \end{aligned} \tag{A.2}$$

The coefficient decomposes as

$$\mathbf{d}_A = \Gamma_A\mathbf{u}_T + \mathbf{e}_A^\perp, \qquad \Gamma_A = \mathbf{u}_T^\top\mathbf{d}_A, \qquad \mathbf{e}_A^\perp = \mathbf{P}_T^\perp\mathbf{d}_A. \tag{A.3}$$

Here Γ_A is signed. The regression remainder $\boldsymbol{\varepsilon}_{A,i}$ is distinct from the Transformer residual stream.

Lemma A.1 (Regression and within-condition covariance)**.** *Let* $\boldsymbol{\mu}_\pm = \mathbb{E}[\mathbf{A}_i \mid \tau_i = \pm\frac{1}{2}]$. *Then*

$$\mathbf{d}_A = \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-, \qquad \boldsymbol{\varepsilon}_{A,i} = \mathbf{A}_i - \mathbb{E}[\mathbf{A}_i \mid \tau_i], \qquad \mathbb{E}[\boldsymbol{\varepsilon}_{A,i} \mid \tau_i] = \mathbf{0}. \tag{A.4}$$

Consequently, with $\mathbf{\Sigma}_{\varepsilon_A} = \mathbb{E}[\boldsymbol{\varepsilon}_{A,i}\boldsymbol{\varepsilon}_{A,i}^\top]$,

$$\begin{aligned} \mathbf{\Sigma}_A &= \sigma_\tau^2\mathbf{d}_A\mathbf{d}_A^\top + \mathbf{\Sigma}_{\varepsilon_A}, \\ \mathbf{\Sigma}_{\varepsilon_A} &= \tfrac{1}{2}\operatorname{Cov}(\mathbf{A}_i \mid \tau_i = \tfrac{1}{2}) + \tfrac{1}{2}\operatorname{Cov}(\mathbf{A}_i \mid \tau_i = -\tfrac{1}{2}). \end{aligned} \tag{A.5}$$

Proof. Balance gives $\mathbb{E}[\mathbf{A}_i] = (\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-)/2$ and $\mathbb{E}[\tau_i \mathbf{A}_i] = (\boldsymbol{\mu}_+ - \boldsymbol{\mu}_-)/4$. Substitution into equation A.2 proves $\mathbf{d}_A = \boldsymbol{\mu}_+ - \boldsymbol{\mu}_-$, and

$$\mathbb{E}[\mathbf{A}_i \mid \tau_i] = \tfrac{1}{2}(\boldsymbol{\mu}_+ + \boldsymbol{\mu}_-) + \tau_i \mathbf{d}_A.$$

This proves equation A.4. Thus $\mathbb{E}[\boldsymbol{\varepsilon}_{A,i}] = \mathbf{0}$ and $\mathbb{E}[\tau_i \boldsymbol{\varepsilon}_{A,i}] = \mathbf{0}$. Expanding the second moment of $\tau_i \mathbf{d}_A + \boldsymbol{\varepsilon}_{A,i}$ eliminates both cross terms, and conditioning the remainder second moment on τ_i gives the second identity in equation A.5. Only zero cross moments are used; the task and remainder need not be independent. □

A.2 Directional covariance and the main spectral guarantee

Resolve the remainder along and orthogonal to the task axis:

$$\begin{aligned} s_i &= \mathbf{u}_T^\top \boldsymbol{\varepsilon}_{A,i}, \quad & \boldsymbol{\zeta}_i &= \mathbf{P}_T^\perp \boldsymbol{\varepsilon}_{A,i}, \\ q_\varepsilon &= \mathbb{E}[s_i^2], & \mathbf{C}_\varepsilon &= \mathbb{E}[\boldsymbol{\zeta}_i \boldsymbol{\zeta}_i^\top], \\ \mathbf{m}_\varepsilon &= \mathbb{E}[\boldsymbol{\zeta}_i s_i]. \end{aligned} \tag{A.6}$$

Since $\boldsymbol{\varepsilon}_{A,i} = s_i \mathbf{u}_T + \boldsymbol{\zeta}_i$ and $\boldsymbol{\zeta}_i \perp \mathbf{u}_T$, expanding its second moment gives

$$\boldsymbol{\Sigma}_{\varepsilon_A} = q_\varepsilon \mathbf{u}_T \mathbf{u}_T^\top + \mathbf{u}_T \mathbf{m}_\varepsilon^\top + \mathbf{m}_\varepsilon \mathbf{u}_T^\top + \mathbf{C}_\varepsilon. \tag{A.7}$$

Equivalently,

$$\begin{aligned} q_\varepsilon &= \mathbf{u}_T^\top \boldsymbol{\Sigma}_{\varepsilon_A} \mathbf{u}_T, \quad & \mathbf{C}_\varepsilon &= \mathbf{P}_T^\perp \boldsymbol{\Sigma}_{\varepsilon_A} \mathbf{P}_T^\perp, \\ \mathbf{m}_\varepsilon &= \mathbf{P}_T^\perp \boldsymbol{\Sigma}_{\varepsilon_A} \mathbf{u}_T. \end{aligned} \tag{A.8}$$

Thus $q_\varepsilon \geq 0$, $\mathbf{C}_\varepsilon \succeq 0$, $\mathbf{C}_\varepsilon \mathbf{u}_T = \mathbf{0}$, and $\mathbf{m}_\varepsilon \perp \mathbf{u}_T$. Define the orthogonal block and the coupling vector of the full covariance:

$$\begin{aligned} \mathbf{C}_A &= \sigma_\tau^2 \mathbf{e}_A^\perp (\mathbf{e}_A^\perp)^\top + \mathbf{C}_\varepsilon, \\ \mathbf{m}_A &= \sigma_\tau^2 \Gamma_A \mathbf{e}_A^\perp + \mathbf{m}_\varepsilon. \end{aligned} \tag{A.9}$$

Lemma A.1 and equation A.7 give $\mathbf{C}_A = \mathbf{P}_T^\perp \boldsymbol{\Sigma}_A \mathbf{P}_T^\perp$ and $\mathbf{m}_A = \mathbf{P}_T^\perp \boldsymbol{\Sigma}_A \mathbf{u}_T$. To make the block dimensions explicit, choose $\mathbf{U}_\perp \in \mathbb{R}^{d \times (d-1)}$ with orthonormal columns spanning $\mathbf{u}_T^\perp$. Then $\mathbf{U}_\perp \mathbf{U}_\perp^\top = \mathbf{P}_T^\perp$ and $\mathbf{U} = [\, \mathbf{u}_T \;\; \mathbf{U}_\perp \,]$ is orthogonal. We have

$$\mathbf{U}^\top \boldsymbol{\Sigma}_A \mathbf{U} = \begin{bmatrix} V_\parallel & \mathbf{m}_A^\top \mathbf{U}_\perp \\ \mathbf{U}_\perp^\top \mathbf{m}_A & \mathbf{U}_\perp^\top \mathbf{C}_A \mathbf{U}_\perp \end{bmatrix}, \qquad \begin{aligned} V_\parallel &= \mathbf{u}_T^\top \boldsymbol{\Sigma}_A \mathbf{u}_T \\ &= \sigma_\tau^2 \Gamma_A^2 + q_\varepsilon, \\ V_\perp &= \lambda_{\max}(\mathbf{C}_A), \\ \mathrm{Mix} &= \|\mathbf{m}_A\|_2. \end{aligned} \tag{A.10}$$

The matrices $\mathbf{C}_A, \mathbf{C}_\varepsilon$ act on $\mathbb{R}^d$ and annihilate $\mathbf{u}_T$; their restrictions act on $\mathbb{R}^{d-1}$. Positivity and $d \geq 2$ ensure that adding the zero eigenvalue along $\mathbf{u}_T$ does not change their largest eigenvalues. Therefore

$$V_\perp = \max_{\substack{\|\mathbf{w}\|_2 = 1 \\ \mathbf{w} \perp \mathbf{u}_T}} \mathbf{w}^\top \boldsymbol{\Sigma}_A \mathbf{w}.$$

Define the directional dominance and the remainder contrast by

$$D_A = V_\parallel - V_\perp, \qquad \nu_\varepsilon = \lambda_{\max}(\mathbf{C}_\varepsilon), \qquad D_\varepsilon = q_\varepsilon - \nu_\varepsilon. \tag{A.11}$$

In particular, positive dominance is equivalent to the exact condition

$$\sigma_\tau^2 \Gamma_A^2 + q_\varepsilon > \lambda_{\max}\left(\sigma_\tau^2 \mathbf{e}_A^\perp (\mathbf{e}_A^\perp)^\top + \mathbf{C}_\varepsilon\right). \tag{A.12}$$

Define the residual-aware margin, retaining the direction of the regression remainder variance, by

$$M_{\mathrm{RA}} = \sigma_\tau^2 \left(\Gamma_A^2 - \|\mathbf{e}_A^\perp\|_2^2\right) + D_\varepsilon. \tag{A.13}$$

Let $\lambda_1 \geq \lambda_2 \geq \cdots$ be the eigenvalues of $\boldsymbol{\Sigma}_A$. For a unit leading eigenvector $\mathbf{v}_1$, write $\mathrm{Align}(\mathbf{v}_1, \mathbf{u}_T) = (\mathbf{v}_1^\top \mathbf{u}_T)^2$; this quantity is invariant to either vector's sign.

A.2.1 Proof of Theorem 3.2

We first prove the following exact refinement of the margin bound:

$$\begin{aligned} D_A &= M_{\mathrm{RA}} + \Omega_\perp \ \geq M_{\mathrm{RA}}, \\ \Omega_\perp &= \sigma_\tau^2 \|\mathbf{e}_A^\perp\|_2^2 + \nu_\varepsilon - \lambda_{\max}(\mathbf{C}_A) \ \geq 0. \end{aligned} \tag{A.14}$$

The nonnegative slack identifies the only relaxation used to obtain $D_A \geq M_{\mathrm{RA}}$. We then prove the eigengap and alignment claims in equation 3.5.

Proof. For every unit $\mathbf{w} \perp \mathbf{u}_T$,

$$\mathbf{w}^\top \mathbf{C}_A \mathbf{w} = \sigma_\tau^2 (\mathbf{w}^\top \mathbf{e}_A^\perp)^2 + \mathbf{w}^\top \mathbf{C}_\varepsilon \mathbf{w} \leq \sigma_\tau^2 \|\mathbf{e}_A^\perp\|_2^2 + \nu_\varepsilon.$$

Maximizing over $\mathbf{w}$ proves $\Omega_\perp \geq 0$. Substituting the definition of $\Omega_\perp$ into $V_\parallel - V_\perp$ gives equation A.14.

The Rayleigh quotient gives $\lambda_1 \geq V_\parallel$. Let $\mathbf{v}_1, \mathbf{v}_2$ be orthonormal eigenvectors for λ_1, λ_2. Their two-dimensional span intersects the codimension-one space $\mathbf{u}_T^\perp$ nontrivially. A unit vector $\mathbf{z} = a\mathbf{v}_1 + b\mathbf{v}_2$ in this intersection has $a^2 + b^2 = 1$ and $\mathbf{z}^\top \mathbf{\Sigma}_A \mathbf{z} = a^2\lambda_1 + b^2\lambda_2 \geq \lambda_2$. It also has Rayleigh quotient at most $V_\perp$. Hence $\lambda_2 \leq V_\perp$, proving $\lambda_1 - \lambda_2 \geq D_A$ and uniqueness when $D_A > 0$.

For the alignment bound, write $\mathbf{v}_1 = x\mathbf{u}_T + \mathbf{w}$, where $\mathbf{w} \perp \mathbf{u}_T$ and $x^2 + \|\mathbf{w}\|_2^2 = 1$. Projecting $\mathbf{\Sigma}_A \mathbf{v}_1 = \lambda_1 \mathbf{v}_1$ onto the axis and its orthogonal complement gives the ambient-space equations

$$\lambda_1 x = V_\parallel x + \mathbf{m}_A^\top \mathbf{w}, \qquad \lambda_1 \mathbf{w} = x\mathbf{m}_A + \mathbf{C}_A \mathbf{w}. \tag{A.15}$$

Since $\lambda_1 \geq V_\parallel > V_\perp$, the second equation rules out $x = 0$; otherwise $\mathbf{w}$ would be a unit eigenvector of $\mathbf{C}_A$ with eigenvalue exceeding its maximum. Choose the sign so that $x > 0$. If $\mathrm{Mix} = 0$, the same equation forces $\mathbf{w} = \mathbf{0}$, so alignment is one and the bound holds with equality. If $\mathrm{Mix} > 0$, it also rules out $\mathbf{w} = \mathbf{0}$. In this case, set

$$r = \frac{\|\mathbf{w}\|_2}{x} > 0, \qquad \mathbf{n} = \frac{\mathbf{w}}{\|\mathbf{w}\|_2}, \qquad \omega = \mathbf{m}_A^\top \mathbf{n}, \qquad \mu = \mathbf{n}^\top \mathbf{C}_A \mathbf{n}.$$

Dividing the first equation of equation A.15 by x, and taking the inner product of the second with $\mathbf{n}$ and dividing by $\|\mathbf{w}\|_2$, gives

$$\lambda_1 - V_\parallel = r\omega, \qquad \lambda_1 - \mu = \omega / r.$$

Because $\lambda_1 > V_\perp \geq \mu$, we have $\omega > 0$. Subtracting gives

$$V_\parallel - \mu = \omega(1/r - r) \geq D_A > 0.$$

Since $\omega > 0$, we first conclude $0 < r < 1$. Hence $\omega \leq \|\mathbf{m}_A\|_2 = \mathrm{Mix}$ may be substituted without reversing the inequality:

$$D_A \leq \mathrm{Mix}(1/r - r), \qquad \mathrm{Mix}\ r^2 + D_A r - \mathrm{Mix} \leq 0.$$

The positive root of this quadratic gives

$$r \leq \frac{2\,\mathrm{Mix}}{\sqrt{D_A^2 + 4\,\mathrm{Mix}^2} + D_A}.$$

For $S = \sqrt{D_A^2 + 4\,\mathrm{Mix}^2}$, the squared root bound is $r^2 \leq (S - D_A)/(S + D_A)$. Consequently,

$$\mathrm{Align} = x^2 = \frac{1}{1 + r^2} \geq \frac{S + D_A}{2S},$$

which is equation 3.5. □

The slack $\Omega_\perp$ records that the coefficient leakage and orthogonal remainder covariance need not attain their maximal variance in the same direction. It is not an additional task signal, and $M_{\mathrm{RA}} > 0$ is generally sufficient rather than necessary. Positive dominance alone guarantees alignment strictly greater than $1/2$. A small ratio $\mathrm{Mix}\,/D_A$ is a sufficient way to make the lower bound close to one. This sufficient bound makes no necessity claim about that ratio. Moreover, $\mathbf{\Sigma}_A \mathbf{u}_T = V_\parallel \mathbf{u}_T + \mathbf{m}_A$; under positive dominance the task axis itself is PC1 exactly when $\mathbf{m}_A = \mathbf{0}$.

A.2.2 Proof of Corollary 3.3

Proof. The sign convention and the alignment assumption imply $\mathbf{v}^\top \mathbf{u}_T \geq \sqrt{a}$. Since both vectors are unit,

$$\|\mathbf{v} - \mathbf{u}_T\|_2^2 = 2 - 2\mathbf{v}^\top \mathbf{u}_T \leq 2 - 2\sqrt{a} = \delta_a^2.$$

The Cauchy–Schwarz inequality then gives

$$\left|(\mathbf{v} - \mathbf{u}_T)^\top (\mathbf{z} - \mathbf{z}')\right| \leq \|\mathbf{v} - \mathbf{u}_T\|_2 \|\mathbf{z} - \mathbf{z}'\|_2 \leq \delta_a \|\mathbf{z} - \mathbf{z}'\|_2.$$

This proves equation 3.6. Subtracting its right-hand side from a strictly larger positive reference score difference proves the ordering assertion. At $a = 1$, the oriented vectors coincide and the score differences are identical. No probability or independence assumption on the compared vectors is needed. □

A.2.3 Proof of Corollary 3.5

Proof. The strict inequalities $G_0 > L_0 \geq 0$ and $R_0 > 0$, together with $\sigma_\tau^2 = \frac{1}{4} > 0$, give

$$D_0 = \sigma_\tau^2 (G_0^2 - L_0^2) + R_0 = \sigma_\tau^2 (G_0 - L_0)(G_0 + L_0) + R_0 > 0,$$

establishing equation 3.10. Equation equation 3.9 gives

$$M_{\mathrm{RA}} \geq \sigma_\tau^2 (G_0^2 - L_0^2) + R_0 = D_0.$$

Here $\Gamma_A^2 \geq G_0^2$ uses $G_0 > 0$, $\|\mathbf{e}_A^\perp\|_2^2 \leq L_0^2$ uses $L_0 \geq 0$, and (A-RD) supplies $D_\varepsilon = q_\varepsilon - \nu_\varepsilon \geq R_0 > 0$. Theorem 3.2 gives $D_A \geq D_0 > 0$ and the eigengap bound. When an additional upper bound $\mathrm{Mix} \leq B_0$, $B_0 \geq 0$, is available, $0 \leq \mathrm{Mix}/D_A \leq B_0/D_0$, so

$$\frac{1}{2}\left(1 + \frac{1}{\sqrt{1 + 4(\mathrm{Mix}/D_A)^2}}\right) \geq \frac{1}{2}\left(1 + \frac{1}{\sqrt{1 + 4(B_0/D_0)^2}}\right).$$

These are exactly the bounds in equation 3.5 and equation 3.11, including $B_0 = 0$. □

In Section A.5, (A-RD) supplies $R_0 > 0$ and (A-M) supplies $G_0 > L_0$; together they establish the positive component margin used above.

The signed condition $\Gamma_A \geq G_0 > 0$ is convenient for adding blockwise lower bounds below. For the spectral conclusion alone it may be replaced by $|\Gamma_A| \geq G_0$, since only Γ_A^2 enters the margin.

A.3 Exact QKVO head–token decomposition

Let $\mathcal{A}$ be the complete nonempty finite set of token slots for every model query head at the fixed layer and readout, indexed by $a = (h, t)$. Token slots follow one fixed indexing rule across samples. The attention weights are those used at inference, without attention dropout, and satisfy

$$\alpha_{i,h,t} \geq 0, \qquad \sum_{t:(h,t)\in\mathcal{A}} \alpha_{i,h,t} = 1. \tag{A.16}$$

Each head has at least one eligible token in every sample. Missing or masked atoms are assigned zero attention weight and, by convention, zero projected value. Let $KV(h)$ identify the KV head used by query head h; ordinary multihead attention has $KV(h) = h$. For a value dimension d_v, define

$$\begin{aligned} &\mathbf{v}_{i,KV(h),t} \in \mathbb{R}^{d_v}, \qquad \mathbf{W}_h^O \in \mathbb{R}^{d \times d_v}, \\ &\mathbf{y}_{i,a} = \begin{cases} \mathbf{W}_h^O \mathbf{v}_{i,KV(h),t}, & \text{if atom } a \text{ is eligible in sample } i, \\ \mathbf{0}, & \text{otherwise,} \end{cases} \\ &\mathbf{z}_{i,a} = \alpha_{i,a} \mathbf{y}_{i,a}, \qquad \mathbf{A}_i = \sum_{a\in\mathcal{A}} \mathbf{z}_{i,a}. \end{aligned} \tag{A.17}$$

The output matrices are fixed. A fixed output bias may be omitted since centering removes it. The last identity describes the linear attention write; it does not include a residual-stream

addition or a subsequent normalization or gate. Assume $\mathbb{E}\|\mathbf{y}_{i,a}\|_2^2 < \infty$ for each atom. Together with $0 \le \alpha_{i,a} \le 1$ and finite $\mathcal{A}$, this guarantees all moments used below, including those of the interaction terms. Indeed, $|\delta\alpha_{i,a}| \le 1$, so $\mathbb{E}\|\delta\alpha_{i,a}\delta\mathbf{y}_{i,a}\|_2^2 \le \mathbb{E}\|\delta\mathbf{y}_{i,a}\|_2^2 < \infty$; no fourth-moment or independence assumption is needed.

Write $\bar{\alpha}_a = \mathbb{E}[\alpha_{i,a}]$, $\bar{\mathbf{y}}_a = \mathbb{E}[\mathbf{y}_{i,a}]$, $\delta\alpha_{i,a} = \alpha_{i,a} - \bar{\alpha}_a$, and $\delta\mathbf{y}_{i,a} = \mathbf{y}_{i,a} - \bar{\mathbf{y}}_a$. For $\widetilde{\mathbf{z}}_{i,a} = \mathbf{z}_{i,a} - \mathbb{E}[\mathbf{z}_{i,a}]$, define

$$\mathbf{d}_a = \frac{\mathbb{E}[\tau_i \widetilde{\mathbf{z}}_{i,a}]}{\sigma_\tau^2}, \qquad d_a^\alpha = \frac{\mathbb{E}[\tau_i \delta\alpha_{i,a}]}{\sigma_\tau^2}, \qquad \mathbf{d}_a^y = \frac{\mathbb{E}[\tau_i \delta\mathbf{y}_{i,a}]}{\sigma_\tau^2}. \tag{A.18}$$

For every query head, normalization also implies $\sum_t \bar{\alpha}_{h,t} = 1$ and $\sum_t d_{h,t}^\alpha = 0$, where the sums include all its atoms. Sharing KV heads does not change these per-query-head identities. Separate value, routing, and interaction contributions as

$$\mathbf{d}_a^V = \bar{\alpha}_a \mathbf{d}_a^y, \qquad \mathbf{d}_a^R = d_a^\alpha \bar{\mathbf{y}}_a, \qquad \mathbf{d}_a^X = \frac{\mathbb{E}[\tau_i \delta\alpha_{i,a}\delta\mathbf{y}_{i,a}]}{\sigma_\tau^2}. \tag{A.19}$$

Here routing contains the effect of queries and keys through the attention weights, and the values already include their output projections.

A.3.1 Proof of Proposition 3.4

Proof. The centered product has the exact expansion

$$\widetilde{\mathbf{z}}_{i,a} = \bar{\alpha}_a \delta\mathbf{y}_{i,a} + \bar{\mathbf{y}}_a \delta\alpha_{i,a} + \delta\alpha_{i,a}\delta\mathbf{y}_{i,a} - \mathbb{E}[\delta\alpha_{i,a}\delta\mathbf{y}_{i,a}]. \tag{A.20}$$

Multiply by τ_i, take expectations, and divide by σ_τ^2. The constant term vanishes because $\mathbb{E}[\tau_i] = 0$, giving the local identity

$$\mathbf{d}_a = \mathbf{d}_a^V + \mathbf{d}_a^R + \mathbf{d}_a^X. \tag{A.21}$$

Since the atom set is finite and complete, centering and the task cross moment commute with its sum. Thus $\mathbf{d}_A = \sum_a \mathbf{d}_a$, which proves equation 3.8. □

These are algebraic attributions under the specified representation and centering convention, not independent causal effects. In particular, changing the value assigned to a zero-weight token leaves $\mathbf{z}_{i,a}$ unchanged but can redistribute its V, R, X terms; the zero-extension convention must therefore be held fixed. In particular, V measures changes of the zero-extended projected value, which can include changes in token eligibility.

The interaction coefficient also has a precise within-condition interpretation. Write $\alpha_{a,\pm} = \mathbb{E}[\alpha_{i,a} \mid \tau_i = \pm\frac{1}{2}]$ and $\mathbf{y}_{a,\pm} = \mathbb{E}[\mathbf{y}_{i,a} \mid \tau_i = \pm\frac{1}{2}]$, and define

$$\mathbf{K}_{a,\pm} = \mathbb{E}[(\alpha_{i,a} - \alpha_{a,\pm})(\mathbf{y}_{i,a} - \mathbf{y}_{a,\pm}) \mid \tau_i = \pm\tfrac{1}{2}].$$

Balance gives $\alpha_{a,\pm} - \bar{\alpha}_a = \pm d_a^\alpha/2$ and $\mathbf{y}_{a,\pm} - \bar{\mathbf{y}}_a = \pm\mathbf{d}_a^y/2$. Expanding the conditional product therefore yields

$$\mathbb{E}[\delta\alpha_{i,a}\delta\mathbf{y}_{i,a} \mid \tau_i = \pm\tfrac{1}{2}] = \mathbf{K}_{a,\pm} + \tfrac{1}{4} d_a^\alpha \mathbf{d}_a^y.$$

The common last term cancels in the class difference, so $\mathbf{d}_a^X = \mathbf{K}_{a,+} - \mathbf{K}_{a,-}$. Thus X is a contrast of within-condition routing–value cross-covariances; it is not a measure of every possible interaction between routing and values.

A.4 Head blocks and bounds on the task coefficient

Fix a subset $\mathcal{H}_\star$ of the model's query heads, and partition all atoms into

$$\mathcal{C} = \{(h,t) \in \mathcal{A} : h \in \mathcal{H}_\star\}, \qquad \mathcal{B} = \mathcal{A} \setminus \mathcal{C}.$$

Both sets are independent of the sample index and contain all tokens of their respective heads. Either set may be empty; its contributions are then zero. No semantic token labels are required. Define the six blocks $\mathfrak{B} = \{\mathcal{C}, \mathcal{B}\} \times \{V, R, X\}$, with

$$\mathbf{d}_{(\mathcal{Q},\kappa)} = \sum_{a \in \mathcal{Q}} \mathbf{d}_a^\kappa, \qquad \mathbf{e}_b^\perp = \mathbf{P}_T^\perp \mathbf{d}_b \quad (b \in \mathfrak{B}).$$

All sums over b, c below run over these six blocks. The parallel contributions are

$$V_{\mathcal{C}}^{\parallel} = \mathbf{u}_T^\top \mathbf{d}_{(\mathcal{C},V)}, \quad R_{\mathcal{C}}^{\parallel} = \mathbf{u}_T^\top \mathbf{d}_{(\mathcal{C},R)},$$
$$X_{\mathcal{C}}^{\parallel} = \mathbf{u}_T^\top \mathbf{d}_{(\mathcal{C},X)}, \quad B^{\parallel} = \mathbf{u}_T^\top \sum_{\kappa \in \{V,R,X\}} \mathbf{d}_{(\mathcal{B},\kappa)}.$$

Proposition 3.4 gives

$$\Gamma_A = V_{\mathcal{C}}^{\parallel} + R_{\mathcal{C}}^{\parallel} + X_{\mathcal{C}}^{\parallel} + B^{\parallel}, \qquad \mathbf{e}_A^{\perp} = \sum_b \mathbf{e}_b^{\perp}. \tag{A.22}$$

In particular, $V_{\mathcal{C}}^{\parallel} = \sum_{a \in \mathcal{C}} \bar{\alpha}_a \mathbf{u}_T^\top \mathbf{d}_a^y$ retains both the average routing weight and the projected value change. Each selected head has total attention mass one, so this partition introduces no token-coverage normalization. The total mass across selected query heads is $|\mathcal{H}_\star|$; their writes are summed, without division by the head count or KV sharing multiplicity.

The drive and leakage inequalities in equation 3.9 follow from the explicit assumptions:

$$\begin{aligned} &\text{(A-S)} \quad V_{\mathcal{C}}^{\parallel} \geq \underline{V}, \quad R_{\mathcal{C}}^{\parallel} \geq \underline{R}, \quad X_{\mathcal{C}}^{\parallel} \geq \underline{X}, \quad B^{\parallel} \geq -\overline{C}_{\text{adv}}, \\ &\qquad\qquad G_0 := \underline{V} + \underline{R} + \underline{X} - \overline{C}_{\text{adv}} > 0, \qquad \overline{C}_{\text{adv}} \geq 0; \\ &\text{(A-CL)} \quad \sum_{b,c} H_{bc}^{\perp} \leq L_0^2, \qquad H_{bc}^{\perp} = (\mathbf{e}_b^{\perp})^\top \mathbf{e}_c^{\perp}, \qquad L_0 \geq 0. \end{aligned} \tag{A.23}$$

The three underlined bounds may be signed. Equation equation A.22 gives $\Gamma_A \geq G_0$, and $\|\mathbf{e}_A^{\perp}\|_2^2 = \sum_{b,c} H_{bc}^{\perp}$ gives the leakage bound. Negative cross inner products are retained: individual blocks need not align with the task axis. These inequalities must be established for the chosen blocks; the decomposition alone does not imply them.

A.5 Exact remainder closure and directional bounds

Define atom-level regression remainders

$$\begin{aligned} \eta_{i,a}^R &= \delta\alpha_{i,a} - \tau_i d_a^\alpha, \\ \boldsymbol{\eta}_{i,a}^V &= \delta\mathbf{y}_{i,a} - \tau_i \mathbf{d}_a^y, \\ \boldsymbol{\eta}_{i,a}^X &= \delta\alpha_{i,a}\delta\mathbf{y}_{i,a} - \mathbb{E}[\delta\alpha_{i,a}\delta\mathbf{y}_{i,a}] - \tau_i \mathbf{d}_a^X. \end{aligned} \tag{A.24}$$

To verify their moment properties, let X_i be any centered scalar or vector with finite second moment, set $d_X = \mathbb{E}[\tau_i X_i]/\sigma_\tau^2$, and write $\eta_i = X_i - \tau_i d_X$. Then

$$\mathbb{E}[\eta_i] = \mathbb{E}[X_i] - \mathbb{E}[\tau_i] d_X = 0, \qquad \mathbb{E}[\tau_i \eta_i] = \mathbb{E}[\tau_i X_i] - \sigma_\tau^2 d_X = 0,$$

where 0 denotes the scalar or vector zero as appropriate. Applying this to the three centered terms in equation A.24 proves their zero means and zero task cross moments. In the interaction case, centering does not change d_X because $\mathbb{E}[\tau_i] = 0$. The binary calculation in Lemma A.1 also gives $\mathbb{E}[X_i \mid \tau_i] = \tau_i d_X$, hence $\mathbb{E}[\eta_i \mid \tau_i] = 0$. For $\mathcal{Q} \in \{\mathcal{C}, \mathcal{B}\}$, set

$$\begin{aligned} \boldsymbol{\xi}_{i,(\mathcal{Q},V)} &= \sum_{a \in \mathcal{Q}} \bar{\alpha}_a \boldsymbol{\eta}_{i,a}^V, \\ \boldsymbol{\xi}_{i,(\mathcal{Q},R)} &= \sum_{a \in \mathcal{Q}} \bar{\mathbf{y}}_a \eta_{i,a}^R, \\ \boldsymbol{\xi}_{i,(\mathcal{Q},X)} &= \sum_{a \in \mathcal{Q}} \boldsymbol{\eta}_{i,a}^X. \end{aligned} \tag{A.25}$$

Lemma A.2 (Six-block remainder closure)**.** *The complete block sum satisfies*

$$\boldsymbol{\varepsilon}_{A,i} = \sum_b \boldsymbol{\xi}_{i,b}, \qquad \mathbb{E}[\boldsymbol{\xi}_{i,b}] = \mathbf{0}, \qquad \mathbb{E}[\tau_i \boldsymbol{\xi}_{i,b}] = \mathbf{0}. \tag{A.26}$$

In fact, each block is conditionally centered: $\mathbb{E}[\boldsymbol{\xi}_{i,b} \mid \tau_i] = \mathbf{0}$.

Proof. Subtracting $\tau_i \mathbf{d}_a$ from equation A.20 and using equation A.21 gives

$$\widetilde{\mathbf{z}}_{i,a} - \tau_i \mathbf{d}_a = \bar{\alpha}_a \boldsymbol{\eta}^V_{i,a} + \bar{\mathbf{y}}_a \eta^R_{i,a} + \boldsymbol{\eta}^X_{i,a}.$$

Sum over the fixed, complete partition. The moment identities follow from equation A.24 and linearity. □

Let $s_{i,b} = \mathbf{u}_T^\top \boldsymbol{\xi}_{i,b}$ and $\boldsymbol{\zeta}_{i,b} = \mathbf{P}_T^\perp \boldsymbol{\xi}_{i,b}$. By Lemma A.2, $s_i = \sum_b s_{i,b}$ and $\boldsymbol{\zeta}_i = \sum_b \boldsymbol{\zeta}_{i,b}$. Therefore

$$\begin{aligned} q_\varepsilon &= \sum_{b,c} \mathbb{E}[s_{i,b} s_{i,c}], \\ \mathbf{C}_\varepsilon &= \sum_{b,c} \mathbb{E}[\boldsymbol{\zeta}_{i,b} \boldsymbol{\zeta}_{i,c}^\top], \\ \mathbf{m}_\varepsilon &= \sum_{b,c} \mathbb{E}[\boldsymbol{\zeta}_{i,b} s_{i,c}]. \end{aligned} \tag{A.27}$$

The sums include every ordered pair and all diagonal terms. Equivalently, using any fixed ordering of the blocks,

$$\begin{aligned} q_\varepsilon &= \sum_b \mathbb{E}[s_{i,b}^2] + 2 \sum_{b<c} \mathbb{E}[s_{i,b} s_{i,c}], \\ \mathbf{C}_\varepsilon &= \sum_b \mathbb{E}[\boldsymbol{\zeta}_{i,b} \boldsymbol{\zeta}_{i,b}^\top] \\ &\quad + \sum_{b<c} \mathbb{E}[\boldsymbol{\zeta}_{i,b} \boldsymbol{\zeta}_{i,c}^\top + \boldsymbol{\zeta}_{i,c} \boldsymbol{\zeta}_{i,b}^\top]. \end{aligned} \tag{A.28}$$

The scalar cross terms can be negative, and the symmetrized matrix cross terms need not be positive semidefinite. Only their complete covariance sum is necessarily positive semidefinite. No independence between blocks is assumed.

The remainder condition imposes a positive lower bound on the difference between axial variance and the strongest orthogonal variance. We impose it together with a comparison of task-axis drive and orthogonal leakage:

$$\begin{aligned} &\text{(A-RD)} \quad D_\varepsilon = q_\varepsilon - \nu_\varepsilon \\ &\qquad\qquad = \sum_{b,c} \mathbb{E}[s_{i,b} s_{i,c}] - \lambda_{\max}\left(\sum_{b,c} \mathbb{E}[\boldsymbol{\zeta}_{i,b} \boldsymbol{\zeta}_{i,c}^\top] \right) \geq R_0 > 0, \\ &\text{(A-M)} \quad G_0 > L_0. \end{aligned} \tag{A.29}$$

Here (A-RD) is the *remainder directional contrast* condition, with a fixed positive lower bound R_0; (A-M) is the *drive–leakage margin* condition, requiring the task-axis drive bound to exceed the net orthogonal leakage bound. Since (A-CL) gives $L_0 \geq 0$ and $\sigma_\tau^2 = \frac{1}{4} > 0$, (A-RD) and (A-M) jointly yield

$$D_0 := \sigma_\tau^2 (G_0^2 - L_0^2) + R_0 = \sigma_\tau^2 (G_0 - L_0)(G_0 + L_0) + R_0 > 0.$$

Combining this strict positivity with (A-S), (A-CL), and (A-RD) gives

$$D_A \geq M_{\mathrm{RA}} \geq D_0 > 0.$$

Consequently, Corollary 3.5 guarantees directional dominance and

$$\lambda_1 - \lambda_2 \geq D_0 > 0,$$

so PC1 is unique up to sign. Theorem 3.2 also gives $\mathrm{Align}(\mathbf{v}_1, \mathbf{u}_T) > \frac{1}{2}$. Its quantitative alignment bound in equation 3.5 uses the actual Mix defined in equation A.10. A separately supplied upper bound $\mathrm{Mix} \leq B_0$, $B_0 \geq 0$, yields equation 3.11; this additional information is distinct from (A-M).

Condition (A-RD) controls the net contrast directly, without requiring separate bounds on its two variance terms. No independence between those terms is assumed. Summing the signed cross terms before taking a norm or a largest eigenvalue preserves cancellation and directional information. Bounding every term separately can be much more conservative.

B Benchmark Construction and Preparation

B.1 CAC Construction

Source data and question construction. CAC is constructed from CUB-200-2011, FGVC-Aircraft, and Stanford Cars, comprising 200, 100, and 196 classes, respectively. We randomly select three initial gold images per class, yielding 1,488 source records. After shuffling records within each dataset, we sample cyclically across the three sources to obtain 500 question instances: 167 from CUB-200-2011, 167 from FGVC-Aircraft, and 166 from Stanford Cars. Each instance receives an attribute question generated by inserting its class name into one of five source-specific templates. These templates concern bird morphology, aircraft configuration, or vehicle appearance, such as leg colour, engine count, and door count. Template selection is deterministic given the sampling seed and record identifier. The resulting instances contain 474 distinct question texts.

Gold images and graded distractors. Each question is associated with five gold images: the initial gold image and four additional images sampled from the same class. Five distractors are constructed across three levels:

- **L1: Cross-dataset negatives.** One image is sampled uniformly from a randomly selected alternative source dataset.
- **L2: Within-dataset random negatives.** One image is sampled uniformly from the source dataset after excluding the target class.
- **L3: Within-dataset hard negatives.** Images outside the target class are ranked by cosine similarity to the initial gold image using CLIP ViT-L/14 embeddings. Three images are selected in descending similarity order subject to the constraint that they belong to three distinct classes.

Target-class exclusion precedes distractor sampling and retrieval. Thus, L3 difficulty is operationalized through visual similarity under a distinct-class constraint. Relevance labels follow the source annotations: images belonging to the class named in the question are positive, and distractors are negative. The resulting benchmark contains 500 question instances with ten candidates each, yielding 5,000 image–question pairs indexed by question instance.

Data partitioning. The question instances are partitioned into 250 fit, 20 development, and 230 test instances, around the ratio of $10 : 1 : 9$. All candidates associated with an instance remain in the same partition. Partitioning uses question-instance identifiers, preserving separate instances when their question texts coincide. The fit split supports direction estimation, the development split supports head selection, and the test split supports held-out analysis and theoretical validation. Estimated directions and selected head subsets are fixed before test evaluation.

B.2 Visual-RAG Preparation

Source questions and candidate pools. We use all 374 questions from Visual-RAG v2. Each question concerns a species attribute and is accompanied by a pool of 152–300 images from iNat21 depicting the same species. Binary annotations indicate whether each image is relevant to the particular question. The original natural-language questions are retained and inserted directly into the relevance-judgment template.

First-stage retrieval. Following the INQUIRE-Rerank protocol, we construct an initial ranking within each question's supplied candidate pool using CLIP ViT-L/14-336. Question and image embeddings are normalized, and candidates are ranked by their cosine similarity. The top 100 images form the reranking pool. Every evaluated method operates on the same candidates for each query, so differences in reranking performance are assessed under a common retrieval setting.

Data partitioning and evaluation set. Questions are randomly partitioned using seed 0 into 75 calibration queries and 299 test queries, corresponding approximately to the $1:4$ allocation used in the main text. Calibration and configuration choices use the calibration partition, while reported results use the test partition. The initial test set contains 29,900 image–question pairs. Fifteen test queries have no positive image among their retrieved candidates and are excluded consistently across all methods. The final evaluation therefore comprises 284 queries, each containing both positive and negative candidates. Directions estimated on CAC remain fixed when applied to Visual-RAG.

C Methodological Details

C.1 Readout Estimation and Development-Split Head Selection

Standardization and direction estimation. For each model, layer, and activation location, preprocessing and direction estimation use the 2,500 balanced image–question pairs from the 250 CAC fit questions. For an activation vector $\mathbf{x}_i$, we estimate the coordinate-wise mean $\widehat{\boldsymbol{\mu}}$ and sample standard deviation using denominator $n_{\text{fit}} - 1$. A numerical floor is applied to the standard deviations. Coordinates with zero variance or a zero-entry frequency exceeding the configured sparsity threshold are masked. Let $\mathbf{S}$ be diagonal, with inverse floored standard deviations on retained coordinates and zeros elsewhere. The standardized activation and readout are

$$\widetilde{\mathbf{x}}_i = \mathbf{S}(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}), \qquad g(i) = \widehat{\mathbf{v}}_1^\top \widetilde{\mathbf{x}}_i. \tag{C.1}$$

We compute the leading principal direction of the standardized fit activations through their Gram matrix. Its sign is oriented to have positive correlation between fit readout scores and binary relevance labels. The standardizer, direction, and orientation are then frozen. Equivalently, the coefficient in the original activation coordinates is $\mathbf{w} = \mathbf{S}^\top \widehat{\mathbf{v}}_1$, giving $g(i) = \mathbf{w}^\top(\mathbf{x}_i - \widehat{\boldsymbol{\mu}})$.

Donor construction and intervention. Head selection uses the FFN-product activation $\mathbf{H}_k$ at the input to the FFN down-projection. All layer and head indices are zero-based. Selection uses the 200 balanced pairs from the 20 CAC development questions. For each recipient, a donor is chosen from the same question with the opposite relevance label. We construct eight donor maps by randomized greedy assignment with seed 60, allowing each donor to be reused at most four times within a map. For a candidate head h at the selected layer k, we replace its final-token attention output slice, before the attention output projection, with the corresponding donor slice. The patched multi-head output is passed through the native output projection and FFN to obtain the intervened $\mathbf{H}_k$. Intervention effects are measured with the frozen, standardized $\mathbf{H}_k$ readout fitted on CAC fit data.

Effect aggregation and subset selection. Let $d_m(i)$ be the donor for recipient i in map m, $y_i \in \{0,1\}$ its relevance label, and $g_i^{(h \leftarrow d_m(i))}$ the intervened readout. We compute

$$e_h = \frac{1}{8N_{\text{dev}}} \sum_{m=1}^{8} \sum_{i=1}^{N_{\text{dev}}} \operatorname{sign}(y_i - y_{d_m(i)}) \big(g_i - g_i^{(h \leftarrow d_m(i))}\big), \tag{C.2}$$

where $N_{\text{dev}} = 200$. The four heads with the largest $|e_h|$ form $\mathcal{H}_\star$. Thus, signed effects are averaged before their magnitudes are ranked, and either effect sign can contribute to the selected subset. The resulting subsets are frozen for subsequent analysis. The prespecified Qwen3-VL-8B subset coincides exactly with the development-based selection.

C.2 Activation Extraction and Component Readouts

Activation locations. Under the inference protocol in Appendix C.3, activations are extracted at the final prompt position, before answer generation. At decoder block ℓ, let $\mathbf{R}_\ell = \mathbf{Z}_{\ell-1}$ denote the incoming residual stream and $\bar{\mathbf{R}}_\ell$ its RMS-normalized attention input. We extract the attention update $\mathbf{A}_\ell$ after the attention output projection, the post-attention

Table 4: Selected transition layers and head subsets.

Model	k	$\mathcal{H}_\star$
Qwen3-VL-8B	19	$\{10, 13, 16, 31\}$
Qwen2.5-VL-7B	17	$\{1, 3, 21, 26\}$
InternVL2.5-8B	14	$\{0, 1, 3, 28\}$
MiniCPM-V-4.5	19	$\{10, 13, 16, 31\}$
Ovis2.5-9B	20	$\{1, 8, 9, 15\}$

residual $\mathbf{M}_\ell = \mathbf{R}_\ell + \mathbf{A}_\ell$, the FFN-product activation $\mathbf{H}_\ell$ at the input to the FFN down-projection, and the block output $\mathbf{Z}_\ell$. The residual and attention quantities have hidden dimension d, whereas $\mathbf{H}_\ell$ has FFN intermediate dimension d_{ff}.

Separate component readouts. Each of $\mathbf{Z}_{k-1}$, $\bar{\mathbf{R}}_k$, $\mathbf{A}_k$, $\mathbf{M}_k$, and $\mathbf{Z}_k$ in Table 2 has its own fitted standardizer, direction, sign, and probability mapping. The $\mathbf{Z}_\ell$ readout in Table 3 is likewise fitted independently. Table 1 ranks candidates directly by the $\mathbf{A}_k$ readout; all attention coordinates are retained in that analysis. These standardized readouts are distinct from the raw-coordinate covariance analysis in Table 4, whose leading eigenvector is computed from the test covariance and whose reference task direction is derived from fit class means. The respective coordinate systems and directions are kept separate.

C.3 VLM-Based Evidence Reranking

Image–question inputs and judgment prompt. For each Visual-RAG query q, the candidate set $\mathcal{I}_q$ comprises the 100 images retrieved by CLIP, as described in Appendix B. Each candidate image is evaluated independently together with the same query. We use the following judgment prompt, inserting the original question verbatim:

```
Question: {question}

Is this image sufficient to answer the question above correctly?
Answer with only Yes or No.
```

The image precedes the text in a single user turn. Inputs are formatted through each checkpoint's native multimodal chat template, including the assistant generation cue; thinking mode is disabled where supported. The same judgment template is used for CAC fitting and Visual-RAG evaluation. Activations are read at the final token of the generation cue, immediately before the first answer token. Native Yes/No probabilities are obtained from the next-token logits at this position.

Candidate scoring. Reranking uses the model's unmodified forward activations and the frozen readouts estimated on CAC. For the direct readout in Table 1, each candidate receives its standardized attention-based score $g_k^A(q, I)$. For the probability-based readout in Table 3, its score is obtained by applying the CAC-fitted logistic map to the independently fitted block-output readout $g_\ell^Z(q, I)$ at the calibration-selected layer. Specifically,

$$\begin{aligned} s_{\text{direct}}(q, I) &= g_k^A(q, I), \\ s_{\text{prob}}(q, I) &= \text{sigmoid}\left(a_\ell + b_\ell \frac{g_\ell^Z(q, I) - \mu_{g,\ell}}{s_{g,\ell}}\right). \end{aligned} \tag{C.3}$$

Here the superscripts identify the activation location; each readout uses its own standardizer and direction. The Behavioral baseline scores the same image–question inputs by the model's native $P(\text{Yes})$. The logistic fitting and layer-selection procedures are specified in Appendix C.5.

Within-query ranking and evaluation. For each query, its 100 candidates are sorted in descending order of the selected score, producing the reranked list. Higher direct scores indicate stronger alignment with the fitted task direction, while higher mapped probabilities

indicate greater predicted support for the model's Yes response. The positive slopes of the Table 3 mappings make their rankings identical to those of the corresponding block-output scores. All methods operate on the same retrieved candidates, and their ranked lists are evaluated against the same relevance annotations. Preprocessing, directions, and probability mappings remain fixed during this evaluation; Visual-RAG calibration labels are used only to select the Table 3 layer. The reported downstream task is visual evidence reranking.

C.4 Transition-Layer Localization

Transition-layer localization. Layerwise localization uses standardized $\mathbf{H}_\ell$ readouts fitted on CAC fit data. For the half-rise criterion, pooled AUROC is evaluated on the paired configuration of all 250 CAC split-B questions, comprising 1,500 pairs. Let $\bar{a}_\ell$ denote this AUROC curve after three-point smoothing. The transition layer is the first layer satisfying

$$\begin{aligned} k_{\text{half}} &= \min\left\{\ell : \bar{a}_\ell \geq a_{\text{floor}} + \tfrac{1}{2}(a_{\max} - a_{\text{floor}})\right\}, \\ a_{\text{floor}} &= \max\{\min_\ell \bar{a}_\ell,\, 0.5\}, \qquad a_{\max} = \max_\ell \bar{a}_\ell. \end{aligned} \tag{C.4}$$

Two additional localization procedures are used in the analyses: the largest adjacent-layer AUROC increase on the unsmoothed fit-split curve, and the modal layer of the largest increase across five grouped out-of-fold profiles. All three procedures yield the layers in Table 4. Tables 1–2 and Figure 2 use these frozen layers; head selection and Table 4 use the grouped-profile modal layer, while Figure 1 uses the fit-split maximum-increase layer. The half-rise profile includes all split-B questions, including the 230 CAC test questions, and is therefore not independent of that test split.

C.5 Output-Probability Mapping and Calibration-Based Layer Selection

Fitting to native answer probabilities. For each model, layer, and activation location, a logistic map converts the scalar readout into a Yes-probability estimate. Its target is the model's normalized native answer probability

$$t_i = \frac{P_i(\text{Yes})}{P_i(\text{Yes}) + P_i(\text{No})}, \qquad p_i = \text{sigmoid}\left(a + b\frac{g_i - \mu_g}{s_g}\right), \tag{C.5}$$

where μ_g and s_g are the scalar-score mean and standard deviation on the corresponding CAC fitting subset. We minimize soft-label cross-entropy,

$$\mathcal{L}(a, b) = -\frac{1}{N}\sum_{i=1}^{N}\left[t_i \log p_i + (1 - t_i)\log(1 - p_i)\right], \tag{C.6}$$

using L-BFGS-B without regularization or sign constraints. This is equivalent to an affine logistic map of the unnormalized score with intercept $a' = a - b\mu_g/s_g$ and slope $b' = b/s_g$.

Fitting subsets and transfer. For Table 3, the logistic parameters are fitted on the 1,500 CAC paired-fit entries: the initial gold image and five distractors for each of the 250 fit questions. The activation standardizer and direction still use all 2,500 balanced-fit entries. For Table 2 and the stage and layerwise profiles in Figure 2, each logistic map uses the 2,500 balanced-fit entries. All targets are native soft answer probabilities. The fitted preprocessing, directions, and logistic parameters remain fixed on Visual-RAG; its calibration split is used only for layer selection.

Selecting the layer for Table 3. For every layer, we evaluate the mapped $\mathbf{H}_\ell$ scores on the 75 Visual-RAG calibration questions and select the layer maximizing calibration mAP against relevance labels. Table 3 then uses the independently fitted block-output readout $\mathbf{Z}_\ell$ at that selected layer, together with its own CAC-fitted logistic map. Its selection criterion differs from the transition-layer criterion described in Appendix C.4. Visual-RAG test questions participate in neither parameter fitting nor layer selection. The fitted block-output slopes are positive for all five models, so their probability estimates and scalar scores induce identical rankings. The Behavioral comparison uses the model's native $P(\text{Yes})$.

Token-level correspondence. For the model's actual binary output $\widehat{y}_i$, the assigned probability is p_i when $\widehat{y}_i = \text{Yes}$ and $1 - p_i$ when $\widehat{y}_i = \text{No}$. Averaging this quantity gives the output-token agreement reported in Table 2 and Figure 2. It measures probabilistic correspondence with native decisions.

D Experimental Details and Supplementary Analyses

This section collects the evaluation metrics, statistical inference procedures, and supplementary analyses for the experiments. Readout estimation and inference procedures are specified in Appendix C.

D.1 Reranking Metrics and Statistical Comparisons

D.1.1 Query-Level Ranking Metrics

Notation and evaluation set. Let $\mathcal{Q}_{\text{test}}$ denote the 299 Visual-RAG test queries. Query q has a retrieved candidate set $\mathcal{I}_q = \{I_{q,i}\}_{i=1}^{n_q}$, where $n_q = 100$. Each candidate has a binary relevance label $y_{q,i} \in \{0, 1\}$ and a method-specific score $s_{q,i}$. Let π_q denote the permutation returned by the evaluator when scores are sorted in descending order, and write $y_{q,(r)} = y_{q,\pi_q(r)}$ for the relevance label at rank r. Define the numbers of positive and negative candidates as

$$n_q^+ = \sum_{i=1}^{n_q} y_{q,i}, \qquad n_q^- = n_q - n_q^+. \tag{D.1}$$

Ranking metrics are evaluated on the common query set

$$\mathcal{Q}_{\text{rank}} = \{q \in \mathcal{Q}_{\text{test}} : n_q^+ > 0,\ n_q^- > 0\}, \qquad N_{\text{rank}} = |\mathcal{Q}_{\text{rank}}| = 284. \tag{D.2}$$

The same 15 queries without a retrieved positive are excluded for every method. For any query-level ranking metric m_q, the reported point estimate is its unweighted query mean,

$$\widehat{M} = \frac{1}{N_{\text{rank}}} \sum_{q \in \mathcal{Q}_{\text{rank}}} m_q. \tag{D.3}$$

All positive counts and recall denominators refer to the retrieved candidate pool. The definitions below take values in $[0, 1]$; percentage reporting multiplies both point estimates and interval endpoints by 100.

Mean average precision. Let $c_q(r) = \sum_{j=1}^{r} y_{q,(j)}$ and $P_q(r) = c_q(r)/r$ denote the number of positives and precision within the first r ranks. For distinct scores, the INQUIRE-Rerank truncated average precision is

$$\text{AP}_q\,@K = \frac{1}{\min(n_q^+, K)} \sum_{r=1}^{K} y_{q,(r)} P_q(r), \qquad 1 \le K \le n_q. \tag{D.4}$$

The implementation multiplies the top-K average precision by $c_q(K)/\min(n_q^+, K)$. It returns zero when $c_q(K) = 0$. Mean average precision is explicitly defined by

$$\text{mAP}\,@K = \frac{1}{N_{\text{rank}}} \sum_{q \in \mathcal{Q}_{\text{rank}}} \text{AP}_q\,@K. \tag{D.5}$$

The reported mAP = mAP @100 uses the full reranking pool, $K = 100 = n_q$, for which the implementation's multiplicative factor equals one.

For tied scores, the implementation uses scikit-learn's threshold-based average precision. Its exact full-pool form can be written using the J_q distinct-score groups $\mathcal{G}_{q,1}, \ldots, \mathcal{G}_{q,J_q}$ ordered by decreasing score, each containing the indices of candidates with that score. Let $a_{q,j} = \sum_{i \in \mathcal{G}_{q,j}} y_{q,i}$, $A_{q,j} = \sum_{h=1}^{j} a_{q,h}$, and $N_{q,j} = \sum_{h=1}^{j} |\mathcal{G}_{q,h}|$. Then

$$\text{AP}_q\,@100 = \frac{1}{n_q^+} \sum_{j=1}^{J_q} a_{q,j} \frac{A_{q,j}}{N_{q,j}}. \tag{D.6}$$

This reduces to Equation D.4 when all scores are distinct.

Normalized discounted cumulative gain. For binary relevance and distinct scores, discounted cumulative gain and its ideal value are

$$\mathrm{DCG}_q\,@K = \sum_{r=1}^{K} \frac{y_{q,(r)}}{\log_2(r+1)}, \qquad \mathrm{IDCG}_q\,@K = \sum_{r=1}^{\min(n_q^+,K)} \frac{1}{\log_2(r+1)}. \tag{D.7}$$

The normalized score is

$$\mathrm{nDCG}_q\,@K = \frac{\mathrm{DCG}_q\,@K}{\mathrm{IDCG}_q\,@K}. \tag{D.8}$$

The reported query-averaged metric is

$$\mathrm{nDCG}\,@K = \frac{1}{N_{\mathrm{rank}}} \sum_{q \in \mathcal{Q}_{\mathrm{rank}}} \mathrm{nDCG}_q\,@K, \qquad K = 10. \tag{D.9}$$

Following the INQUIRE-Rerank implementation, we supplement the top-K list with $n_q^+ - c_q(K)$ dummy positives at score -100. These entries preserve the full candidate-pool positive count in the ideal normalization. We then call `ndcg_score`, which averages over tied scores in the supplied list.

Mean reciprocal rank. Let $r_q^\star = \min\{r : y_{q,(r)} = 1\}$ be the rank of the first relevant candidate. Reciprocal rank and its query mean are

$$\mathrm{RR}_q = \frac{1}{r_q^\star}, \qquad \mathrm{MRR} = \frac{1}{N_{\mathrm{rank}}} \sum_{q \in \mathcal{Q}_{\mathrm{rank}}} \mathrm{RR}_q\,. \tag{D.10}$$

Recall. Recall at cutoff K is

$$\mathrm{Recall}_q\,@K = \frac{c_q(K)}{n_q^+}. \tag{D.11}$$

The reported query-averaged recall is

$$\mathrm{Recall}\,@K = \frac{1}{N_{\mathrm{rank}}} \sum_{q \in \mathcal{Q}_{\mathrm{rank}}} \mathrm{Recall}_q\,@K, \qquad K \in \{5, 10\}. \tag{D.12}$$

The table headings R@5 and R@10 denote these quantities. These quantify retrieval of positives within the reranking pool; positives omitted by first-stage retrieval are outside the denominator.

Area under the ROC curve. For each query, AUROC is the fraction of positive–negative pairs ordered correctly by the scores, assigning half credit to ties:

$$\mathrm{AUROC}_q = \frac{1}{n_q^+ n_q^-} \sum_{i:y_{q,i}=1} \sum_{j:y_{q,j}=0} \left[\mathbf{1}\{s_{q,i} > s_{q,j}\} + \tfrac{1}{2}\mathbf{1}\{s_{q,i} = s_{q,j}\}\right], \tag{D.13}$$

where $\mathbf{1}\{\cdot\}$ is the indicator function. The reported AUROC is the query-macro estimate

$$\widehat{\mathrm{AUROC}} = \frac{1}{N_{\mathrm{rank}}} \sum_{q \in \mathcal{Q}_{\mathrm{rank}}} \mathrm{AUROC}_q\,. \tag{D.14}$$

Thus, comparisons are formed within each query before aggregation. The same metric definitions apply to all scoring methods. A strictly increasing score transformation preserves these metrics when score ties are preserved; the positive logistic slopes in Table 3 permit evaluation using the underlying readout scores.

D.1.2 Confidence Intervals and Paired Comparisons

Ranking intervals in Tables 1 and 3. For each metric, we retain the fixed query-level values m_q on $\mathcal{Q}_{\mathrm{rank}}$ and resample queries with replacement. For example, $m_q = \mathrm{AP}_q\,@100$ gives $\widehat{M} = \mathrm{mAP}$. We use $B_{\mathrm{CI}} = 4{,}000$ replicates with seed 7, indexed by $b \in \{1, \ldots, B_{\mathrm{CI}}\}$.

Each replicate contains $N_{\text{rank}} = 284$ draws. Let $W^{*(b)}_{\text{rank},q}$ count how often query q is drawn in replicate b. Its multiplicity vector and the corresponding metric estimate satisfy

$$\mathbf{W}^{*(b)}_{\text{rank}} \sim \text{Multinomial}\Big(N_{\text{rank}}; \tfrac{1}{N_{\text{rank}}}, \ldots, \tfrac{1}{N_{\text{rank}}}\Big), \qquad \widehat{M}^{*(b)} = \frac{1}{N_{\text{rank}}} \sum_{q \in \mathcal{Q}_{\text{rank}}} W^{*(b)}_{\text{rank},q} m_q. \tag{D.15}$$

Thus $\sum_{q \in \mathcal{Q}_{\text{rank}}} W^{*(b)}_{\text{rank},q} = N_{\text{rank}}$. Let $\text{Quant}_p\{x_b\}$ denote the empirical p-quantile of a collection of bootstrap values. The two-sided 95% percentile interval reported alongside $\widehat{M}$ is

$$\text{CI}_{0.95} = \Big[\text{Quant}_{0.025}\{\widehat{M}^{*(b)}\}_{b=1}^{B_{\text{CI}}}, \text{Quant}_{0.975}\{\widehat{M}^{*(b)}\}_{b=1}^{B_{\text{CI}}}\Big]. \tag{D.16}$$

The point estimate remains the full-sample mean in Equation D.3. The same procedure is applied to every ranking metric, including columns for which only point estimates are displayed.

Paired comparisons. Methods share the same query set and bootstrap draws. For methods U and V, we compute the difference within each replicate, with point estimate $\widehat{\Delta} = \widehat{M}_U - \widehat{M}_V$:

$$\Delta^{*(b)} = \widehat{M}^{*(b)}_U - \widehat{M}^{*(b)}_V = \frac{1}{N_{\text{rank}}} \sum_{q \in \mathcal{Q}_{\text{rank}}} W^{*(b)}_{\text{rank},q} (m_{U,q} - m_{V,q}), \tag{D.17}$$

where $m_{U,q}$ and $m_{V,q}$ are the two methods' query-level metric values. Its interval is

$$\text{CI}^{U-V}_{0.95} = \Big[\text{Quant}_{0.025}\{\Delta^{*(b)}\}_{b=1}^{B_{\text{CI}}}, \text{Quant}_{0.975}\{\Delta^{*(b)}\}_{b=1}^{B_{\text{CI}}}\Big]. \tag{D.18}$$

A paired difference is marked significant when its two-sided 95% interval excludes zero. These are comparison-specific intervals; no simultaneous adjustment across all ranking metrics or method pairs is implied.

D.2 Geometric Analysis of Readout Directions

Directions and per-question quantities. Figure 1 uses the first ten principal directions of standardized $\mathbf{H}_k$ activations, estimated on the 2,500 CAC balanced-fit entries. Their signs are oriented using fit labels, and all directions and preprocessing constants are frozen. For each of the 250 split-B questions, let $\mathbf{X}_q \in \mathbb{R}^{10 \times d_{\text{ff}}}$ contain its standardized candidate activations. Let $\mathbf{C}_q$ be obtained by subtracting their within-question mean, and define

$$\boldsymbol{\delta}_q = \frac{1}{5} \sum_{i: y_i = 1} \mathbf{X}_{q,i} - \frac{1}{5} \sum_{i: y_i = 0} \mathbf{X}_{q,i}, \tag{D.19}$$

where $\mathbf{X}_{q,i}$ denotes a candidate activation as a column vector. For each direction $\widehat{\mathbf{v}}_j$, the plotted coordinates are

$$x_{qj} = \frac{\|\mathbf{C}_q \widehat{\mathbf{v}}_j\|_2^2}{\|\mathbf{C}_q\|_{\text{F}}^2}, \qquad y_{qj} = \frac{(\boldsymbol{\delta}_q^\top \widehat{\mathbf{v}}_j)^2}{\|\boldsymbol{\delta}_q\|_2^2}. \tag{D.20}$$

The horizontal coordinate is the fraction of within-question variance captured by the direction; the vertical coordinate is its share of the squared norm of the positive–negative mean difference. Each set of shares sums to one over a complete orthonormal basis when its denominator is nonzero.

Display and interpretation. The plot contains all 2,500 question–direction pairs, with the 250 first-direction points distinguished from directions two through ten. Both axes are logarithmic, with values clipped below 10^{-5} for display. The diagonal marks equality of variance share and class-separation share. A point above it indicates that the direction captures a greater share of the class-mean difference than of the overall within-question variation.

D.3 Output-Token Agreement

Native decisions and mapped probabilities. For candidate i of query q, define the normalized native Yes probability and native binary decision as

$$\rho_{q,i} = \frac{P_{q,i}(\text{Yes})}{P_{q,i}(\text{Yes}) + P_{q,i}(\text{No})}, \qquad \widehat{y}_{q,i} = \mathbf{1}\{\rho_{q,i} > 1/2\}. \tag{D.21}$$

Let $\widehat{p}_{q,i}$ be the readout-derived Yes probability obtained from the CAC-fitted logistic map described in Appendix C.5; $P_{q,i}(\text{Yes})$ and $P_{q,i}(\text{No})$ are the native next-token probabilities for the same image–question input. The mapped probability assigned to the native decision is

$$v_{q,i} = \widehat{y}_{q,i}\widehat{p}_{q,i} + (1 - \widehat{y}_{q,i})(1 - \widehat{p}_{q,i}). \tag{D.22}$$

We first average within each query and then across all test queries:

$$f_q = \frac{1}{n_q}\sum_{i=1}^{n_q} v_{q,i}, \qquad \widehat{F} = \frac{1}{N_{\text{all}}}\sum_{q\in\mathcal{Q}_{\text{test}}} f_q, \qquad N_{\text{all}} = |\mathcal{Q}_{\text{test}}| = 299. \tag{D.23}$$

This is the output-token agreement reported in Table 2 and Figure 2. It measures probabilistic correspondence with the native binary decision and does not require gold relevance labels. Consequently, queries without retrieved positives remain eligible for this metric.

Reference levels. For a given readout, let ρ_{fit} be the mean native Yes probability on its CAC probability-fitting subset. Writing $\mathcal{I}_{\text{fit}}$ for the indices of those pairs and $N_{\text{fit}} = |\mathcal{I}_{\text{fit}}|$, this is

$$\rho_{\text{fit}} = \frac{1}{N_{\text{fit}}}\sum_{i\in\mathcal{I}_{\text{fit}}} \rho_i, \tag{D.24}$$

where ρ_i is the normalized native Yes probability of fit pair i. The fixed-prior reference replaces every $\widehat{p}_{q,i}$ by ρ_{fit}. Define the query-averaged frequency of native Yes decisions as

$$\widehat{\eta} = \frac{1}{N_{\text{all}}}\sum_{q\in\mathcal{Q}_{\text{test}}} \frac{1}{n_q}\sum_{i=1}^{n_q} \widehat{y}_{q,i}. \tag{D.25}$$

The fixed-prior agreement is

$$F_{\text{prior}} = \widehat{\eta}\rho_{\text{fit}} + (1 - \widehat{\eta})(1 - \rho_{\text{fit}}). \tag{D.26}$$

The native-probability reference substitutes $\rho_{q,i}$ for $\widehat{p}_{q,i}$:

$$F_{\text{native}} = \frac{1}{N_{\text{all}}}\sum_{q\in\mathcal{Q}_{\text{test}}} \frac{1}{n_q}\sum_{i=1}^{n_q} \max\{\rho_{q,i}, 1 - \rho_{q,i}\}. \tag{D.27}$$

These provide empirical reference levels rather than mathematical lower and upper bounds. When $F_{\text{native}} > F_{\text{prior}}$, a descriptive normalization is

$$F_{\text{relative}} = \frac{\widehat{F} - F_{\text{prior}}}{F_{\text{native}} - F_{\text{prior}}}. \tag{D.28}$$

This relative position need not lie in $[0, 1]$. The tabulated agreement is the unnormalized value $\widehat{F}$.

D.4 Layerwise and Component-Level Profiles

Layerwise and within-layer profiles. Figure 2 evaluates the $\mathbf{H}_\ell$ readout at every layer on Visual-RAG test data, with normalized depth $(\ell + 0.5)/L$ for a model with L blocks. Experimentally, FFN-product and block-output readouts exhibit very similar layerwise behavior, supporting their use as approximately equivalent descriptions of the observed transition. The plotted layerwise values are computed from $\mathbf{H}_\ell$. At layer k, the inset expands the computation into $\mathbf{Z}_{k-1}$, $\bar{\mathbf{R}}_k$, $\mathbf{A}_k$, $\mathbf{M}_k$, $\mathbf{H}_k$, and $\mathbf{Z}_k$. The five hidden-dimensional stages

use the same independently fitted readouts as Table 2, and the $\mathbf{H}_k$ point is taken from the layerwise profile.

The task-level curve reports question-macro AUROC over the 284 test questions containing both positive and negative retrieved candidates. The token-level curve reports the readout-derived probability assigned to the model's actual Yes/No output, using the mappings described in Appendix C.5, over all 299 test questions. Its value is averaged over candidates within each question and then over questions. Confidence bands use 4,000 question-level bootstrap resamples with seed 7 and a 95% confidence level. Native-model and constant-prediction values provide reference lines.

D.4.1 Confidence Intervals for Layer and Component Profiles

Stage-profile intervals in Table 2. The stage-profile implementation generates multinomial query weights over all $N_{\text{all}} = 299$ test queries, again using 4,000 replicates and seed 7:

$$\mathbf{W}_{\text{all}}^{*(b)} \sim \text{Multinomial}\Big(N_{\text{all}}; \tfrac{1}{N_{\text{all}}}, \ldots, \tfrac{1}{N_{\text{all}}}\Big). \tag{D.29}$$

Let $W_{\text{all},q}^{*(b)}$ be the component of this vector corresponding to query q. Its components satisfy

$$\sum_{q \in \mathcal{Q}_{\text{test}}} W_{\text{all},q}^{*(b)} = N_{\text{all}}.$$

Using these shared weights, the bootstrap stage metrics are

$$\begin{aligned} \widehat{\text{AUROC}}^{*(b)} &= \frac{\sum_{q \in \mathcal{Q}_{\text{rank}}} W_{\text{all},q}^{*(b)} \, \text{AUROC}_q}{\sum_{q \in \mathcal{Q}_{\text{rank}}} W_{\text{all},q}^{*(b)}}, \\ \widehat{F}^{*(b)} &= \frac{1}{N_{\text{all}}} \sum_{q \in \mathcal{Q}_{\text{test}}} W_{\text{all},q}^{*(b)} f_q. \end{aligned} \tag{D.30}$$

AUROC weights are renormalized over the 284 eligible queries, whereas agreement uses all 299 queries. Both intervals use the percentile rule in Equation D.16. This sampling scheme differs from directly drawing 284 eligible queries: the total eligible-query multiplicity in a replicate is random. Its random draws also differ from those in Tables 1 and 3 despite the shared seed value. All prediction scores and fitted readout parameters remain fixed during these query-level bootstrap procedures.

Layer-profile confidence bands. Figure 2 likewise uses 4,000 query-level bootstrap replicates with seed 7 to obtain 95% confidence bands. Its task-level AUROC is evaluated on the 284 eligible queries, and its output-token agreement uses all 299 test queries. The bands summarize uncertainty at each displayed layer or stage; they are not simultaneous confidence bands across the entire profile.

D.5 Lower Confidence Bounds for Spectral Validation

Stratified question-cluster bootstrap. Table 4 uses the 230 CAC test questions, comprising 85 CUB, 61 Aircraft, and 84 Cars questions. Each question contains ten image–question pairs. Let $\mathcal{C}_s$ be the question set in source stratum $s \in \{1, 2, 3\}$, corresponding to CUB, Aircraft, and Cars, respectively, and let $n_s = |\mathcal{C}_s|$. Independently within each stratum, a bootstrap replicate draws n_s questions with replacement. Write $W_{\text{CAC},q}^{*(b)}$ for the multiplicity of question q in this resampling scheme:

$$\big(W_{\text{CAC},q}^{*(b)}\big)_{q \in \mathcal{C}_s} \sim \text{Multinomial}\Big(n_s; \tfrac{1}{n_s}, \ldots, \tfrac{1}{n_s}\Big), \qquad (n_1, n_2, n_3) = (85, 61, 84). \tag{D.31}$$

All ten candidates of a question inherit its multiplicity. We use $B_{\text{LCB}} = 20{,}000$ bootstrap replicates with seed 7101, indexed by $b \in \{1, \ldots, B_{\text{LCB}}\}$. The original sample, stored as replicate zero with $W_{\text{CAC},q}^{*(0)} = 1$ for every question, supplies the point estimates and is excluded from the bootstrap quantiles.

Each replicate recomputes the sample-dependent weights, means, regression coefficients, residuals, covariance matrices, spectral quantities, and component certificates. The reference direction $\widehat{\mathbf{u}}_T$, output projection $\mathbf{W}_O$, selected head subset, layer, model parameters, and fit preprocessing remain fixed. Thus, nonlinear statistics are recomputed from the resampled clusters rather than obtained by averaging previously computed certificate values.

Nominal one-sided lower bounds. For a statistical endpoint r, let $\widehat{\theta}_r$ be its estimate on the original test sample and $\widehat{\theta}_r^{*(b)}$ its value in replicate b. Define

$$\widehat{\mathrm{se}}_r = \mathrm{SD}_{b=1}^{B_{\mathrm{LCB}}}\left(\widehat{\theta}_r^{*(b)}\right), \qquad T_r^{*(b)} = \frac{\widehat{\theta}_r^{*(b)} - \widehat{\theta}_r}{\widehat{\mathrm{se}}_r}. \tag{D.32}$$

Here SD is the empirical standard deviation across bootstrap replicates, and the standardized deviation is defined for endpoints with $\widehat{\mathrm{se}}_r > 0$. Define the endpoint-specific critical value as

$$c_r = \mathrm{Quant}_{0.95}\{T_r^{*(b)}\}_{b=1}^{B_{\mathrm{LCB}}}. \tag{D.33}$$

The unadjusted nominal 95% lower confidence bound is

$$\begin{aligned} \mathrm{LCB}_{r,\mathrm{nom}} &= \widehat{\theta}_r - c_r \widehat{\mathrm{se}}_r \\ &= 2\widehat{\theta}_r - \mathrm{Quant}_{0.95}\{\widehat{\theta}_r^{*(b)}\}_{b=1}^{B_{\mathrm{LCB}}}. \end{aligned} \tag{D.34}$$

This is a reflected, or basic, bootstrap lower bound. It differs from the fifth percentile of the bootstrap estimates. A single bootstrap standard deviation is used for each endpoint across all replicates; the procedure does not perform nested studentization within each resample.

Simultaneous lower bounds. For a specified endpoint family $\mathcal{F}$, define the maximum standardized deviation and its critical value by

$$T_{\mathcal{F}}^{*(b)} = \max_{r \in \mathcal{F}} T_r^{*(b)}, \qquad c_{\mathcal{F}} = \mathrm{Quant}_{0.95}\{T_{\mathcal{F}}^{*(b)}\}_{b=1}^{B_{\mathrm{LCB}}}. \tag{D.35}$$

The simultaneous lower bound for each family member is

$$\mathrm{LCB}_{r,\mathrm{sim}} = \widehat{\theta}_r - c_{\mathcal{F}} \widehat{\mathrm{se}}_r. \tag{D.36}$$

The common critical value provides bootstrap-based approximate one-sided 95% simultaneous coverage within the specified family.

Correction families and reporting. For each model, Table 4 uses two families in the raw activation coordinates:

$$\begin{aligned} \mathcal{F}_{\mathrm{outcome}} &= \{D_A\}, \\ \mathcal{F}_{\mathrm{mechanism}} &= \{M_{\mathrm{RA}}, D_0, G_0, G_0 - L_0, 1 - \kappa_d, D_\varepsilon\}, \end{aligned} \tag{D.37}$$

using the theoretical quantities defined in the main text, with $\kappa_d = \|\mathbf{e}_A^{\perp}\|_2 / \Gamma_A$. The endpoint $1 - \kappa_d$ participates in the six-endpoint mechanism correction even though it is not displayed in the table. The outcome family contains a single endpoint, so its simultaneous and nominal bounds coincide. Critical values are computed separately for each model and family; the stated simultaneity does not extend across both families or all models.

For an endpoint estimate, the notation

$$\widehat{\theta}_r \, [\mathrm{LCB}_{r,\mathrm{sim}}, \mathrm{LCB}_{r,\mathrm{nom}}] \tag{D.38}$$

reports two one-sided lower bounds, whereas the brackets for ranking and agreement metrics contain the endpoints of a two-sided confidence interval. A positivity check is marked as satisfied when $\mathrm{LCB}_{r,\mathrm{sim}} > 0$. These bootstrap statements are conditional on the frozen reference direction, selected layer, and head subset. They quantify test-sample variation with those specifications held fixed and do not propagate uncertainty from their fitting or selection.

E Additional Results

This appendix supplements the results reported in the main text with the corresponding per-model figures for all five judges. Figure 4 shows each judge's layer-wise profile: the readout is at chance through the early layers and switches on abruptly at a single mid-depth layer, the jump layer analysed in the main text. Figure 5 shows, for each held-out question, the share of overall variation carried by each principal axis against the share of the relevant–irrelevant shift it carries: the task-relevance axis (blue) lies systematically above the diagonal, whereas the remaining axes (orange) concentrate around it, i.e. its distribution is clearly separated from that of the other axes.

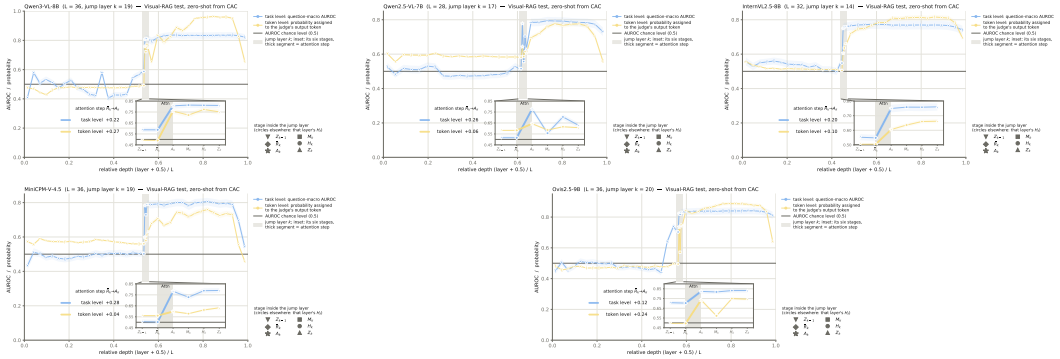

Figure 4: Per-model layer-wise profiles for the five judges, complementing the main text: each judge exhibits a sharp single-layer jump at its mid-depth jump layer.

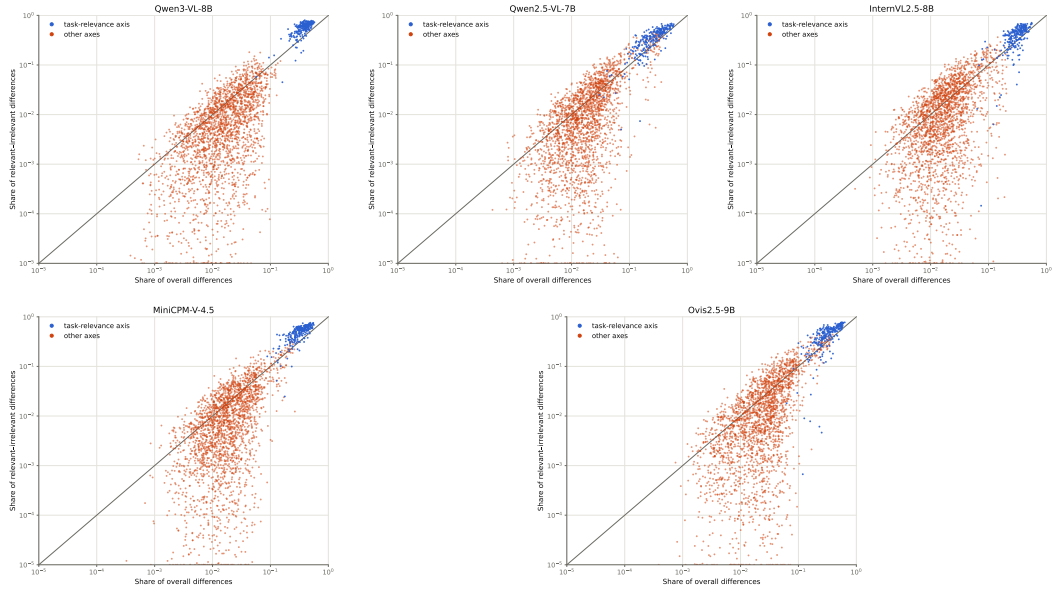

Figure 5: Per-question variance share vs. label-shift share for the five judges, complementing the main text: the task-relevance axis is distributed well above the diagonal, unlike the remaining axes.