PREDICTION TRAINING, REPRESENTATION GEOMETRY, AND HALLUCINATION DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Prediction training shapes the relative geometry of hidden representations. For an attention model trained by gradient descent, we characterize how representations of contexts associated with the same prediction approach one another. Their remaining differences follow a limiting direction determined by prediction costs, context composition, and the parameter update. This connects prediction training to pairwise representation geometry. We use this component-mixture perspective to develop a pre-generation hallucination detector. For each labeled outcome, the detector reconstructs a test state by reweighting nearby reference states. Its energy penalizes concentrated mixture weights and reconstruction error. It requires no classifier training. The detector achieves a mean AUROC of 79.26% across 21 language and vision-language model–dataset combinations. On visual tasks, it reaches 81.65%, exceeding a supervised pre-generation probe by 2.35 points on the same independent test sets.

1 INTRODUCTION

Large language models (LLMs) can produce fluent answers that contain factual errors, sometimes with high confidence (Simhi et al., 2025). For users who rely on these answers for information and decisions, such hallucinations make it important to assess whether a response can be trusted. Once a response is available, its claims can be examined and checked against evidence, but an indication of risk before generation could help a system decide whether to seek additional information before answering. Obtaining such an indication requires working without the answer itself, raising the question of what information is available at this earlier stage. By the time generation begins, the model has already processed the input into hidden representations from which it will produce its response. We therefore ask what these pre-generation states can reveal about the reliability of the answer that follows.

Existing research has examined several sources of evidence for assessing answer reliability. SelfCheckGPT studies factual consistency across sampled responses (Manakul et al., 2023), while semantic entropy measures uncertainty over their meanings (Farquhar et al., 2024). The internal representations of these responses also offer evidence: INSIDE examines consistency in representation space (Chen et al., 2024), and Yin et al. (2024) use local intrinsic dimension to distinguish correct from incorrect generations. The latter provides evidence that local geometry can carry information about answer correctness, but its measurements depend on representations of already generated content. It therefore remains unclear from these results whether such geometric information is available before the answer itself has been produced. At this earlier stage, Ji et al. (2024) show that a probe trained on question states can predict subsequent hallucination risk, establishing that predictive information is already present. This leaves a more specific question: can the local geometry of pre-generation states be used to estimate hallucination risk?

To develop a component-level understanding of local representation geometry, we examine how prediction training changes the weights with which representations are formed. Our analysis uses an attention model with fixed value vectors and a fixed prediction head to isolate how these weight changes affect representation differences. For contexts with the same prediction target and a shared conditional mixture, the shared contribution cancels from their representation difference. The remaining components can therefore determine the entire difference even as their attention weights vanish during training. We characterize this behavior under gradient descent, the difference con-

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.

tracts, while its rescaled vector converges to a limit determined by prediction costs, context composition, and the parameter update. The resulting component description also gives an interpretation of local reconstruction through weighted combinations of reference states, connecting the analysis to the detection method developed in Section 5.

Building on this view, we introduce local mixture energy for pre-generation hallucination detection. For a new question, we reconstruct its pre-generation state using separate groups of nearby reference states from questions that the same model answered correctly or incorrectly. Our working hypothesis is that greater compatibility with incorrect-answer references relative to correct-answer references signals a greater risk of hallucination. Pre-generation states reflect how the model processes a question, so resemblance to states preceding known errors could indicate recurring conditions under which the model struggles to answer correctly. To make this intuition operational, we assess how well a weighted mixture of references reconstructs the new state and how strongly that reconstruction depends on a few references. Local mixture energy balances reconstruction error with a penalty for departures from uniform mixture weights, favoring accurate reconstructions supported by multiple nearby states. We compare the two class energies to estimate hallucination risk, assigning a higher risk when the incorrect-answer references reconstruct the state at lower energy than the correct-answer references.

We evaluate both the component-mixture perspective and its use in hallucination detection. Synthetic experiments test the predicted attention dynamics and examine the role of shared components in a fully trained two-layer Transformer. A separate experiment with Qwen2.5-7B finds partial reuse of attention components across paraphrases, providing empirical support for a local fixed-component approximation in a pretrained LLM. For hallucination detection, we evaluate 21 language and vision-language model–dataset combinations on independent test questions. Local mixture energy outperforms nearest-neighbor label voting across all combinations and improves on the supervised pre-generation probe on average, with the strongest results on visual tasks. These comparisons examine whether reconstructing a test state from labeled neighbors improves error ranking beyond using their labels alone.

2 Related Work

Attention training and representation geometry. Attention training dynamics offer a way to study the structure of language-model representations. Ataee Tarzanagh et al. (2023b;a) connect attention optimization to maximum-margin token selection. Li et al. (2024) analyze retrieval and composition in next-token prediction, showing how gradient descent selects components according to token-priority relations and forms convex combinations for prediction. These results characterize component selection and weight allocation during training. We study the resulting representation differences across contexts and derive their limiting direction as they contract. Prediction costs, context composition, and the parameter update jointly determine this direction.

Representation geometry in pretrained models also varies with context and layer. Ethayarajh (2019) examines anisotropy and contextual variation across layers. Geva et al. (2022) decompose feed-forward outputs into weighted parameter vectors and study their contributions to vocabulary predictions. Such decompositions make representation changes more concrete by identifying the contributing vectors and their weights. Attention value vectors themselves depend on context, making the extent of component reuse across nearby inputs an empirical question that connects fixed-component analysis to pretrained models.

Local geometry and reconstruction. Neighborhood reconstruction has a long history in manifold learning. Locally linear embedding (LLE) reconstructs each sample from its neighbors and uses the resulting weights to construct a low-dimensional embedding (Roweis & Saul, 2000). The weights describe a sample's relation to its neighborhood, and the residual measures how well that neighborhood can represent it. Neighbors' relative positions determine which variations their combinations can express. In class-conditional reconstruction, reference labels specify the samples available for each mixture, while weight constraints and regularization determine the admissible combinations.

Hallucination detection using internal representations. Hallucination detectors draw on answer probabilities, consistency across generations, and internal representations. SelfCheckGPT compares

factual consistency across sampled responses (Manakul et al., 2023), while semantic entropy measures uncertainty over their meanings (Farquhar et al., 2024). INSIDE extends consistency analysis to the internal states of generated answers (Chen et al., 2024). These methods use repeated generations for the same question to assess answer reliability.

Other studies examine truthfulness signals in representation space. Yin et al. (2024) use the local intrinsic dimension of model activations to detect incorrect answers. HaloScope identifies a hallucination-related subspace from unlabeled generations (Du et al., 2024). HARP extracts subspaces through singular value decomposition of the unembedding matrix and uses projected features for detection (Hu et al., 2026). Orgad et al. (2025) study truthfulness information in answer-token representations. The detection stage determines the evidence available: Ji et al. (2024) train a risk probe on prompt states before answer generation. The probe learns a global decision rule from labeled prompt states.

3 Training Dynamics and Representation Geometry

The theoretical analysis considers an attention model with fixed value vectors $b_1, \ldots, b_m$ and a fixed prediction head. A context contains these components in different counts, and shared attention scores determine their weights. Its representation is $h_i = r + \sum_a \pi_i(a) b_a$, where $\pi_i(a)$ is the attention weight of component a and r is a common residual. Training minimizes prediction cross-entropy across contexts. Fixing the values and prediction head isolates how changes in attention weights alter representation geometry. The full model is specified in Appendix A.

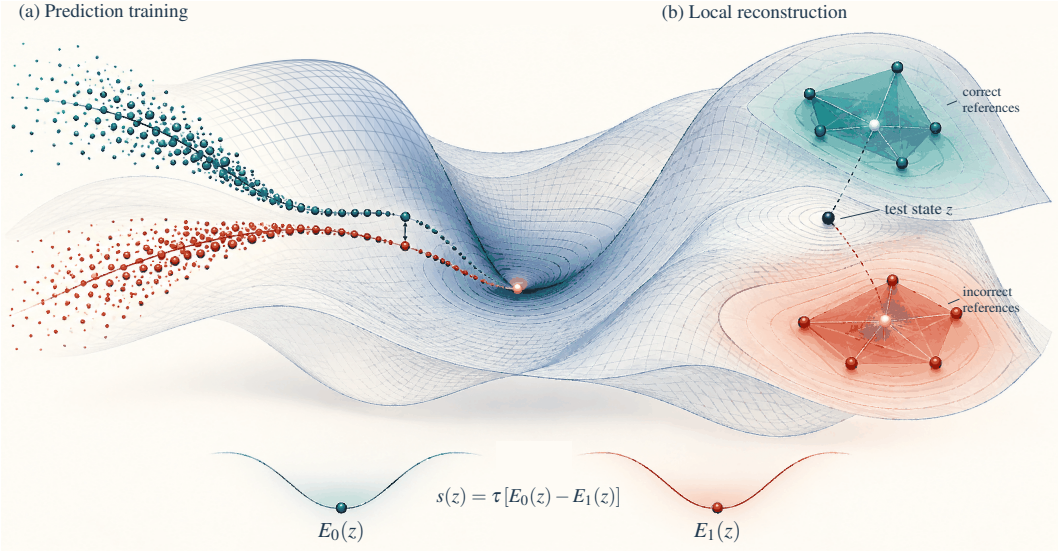


Figure 1: **Prediction training and local reconstruction.** (a) Training contracts representations with a shared optimal mixture, leaving a structured residual difference. (b) Colored vertices are nearby correct- and incorrect-answer reference states. White points denote the optimized mixtures used to reconstruct the test state z, and dashed segments show the reconstruction residuals. Each class energy combines reweighting and reconstruction costs. Lower energy for the incorrect-answer class gives a higher error-risk score. The surface and energy curves are schematic.

To isolate the representation variation that remains after shared predictive content is held fixed, we compare contexts with the same prediction target and a common optimal mixture. Let T be the set of components retained by this mixture. Proportional counts of the components in T across contexts yield the same conditional mixture on T at every shared score vector; write its value as v_T. The difference between two representations is exactly

$$h_i - h_j = \sum_{a \notin T} \big(\pi_i(a) - \pi_j(a)\big)(b_a - v_T). \tag{1}$$

Only components outside T contribute to pairwise representation differences.

To describe how training changes the remaining weights, assign each component a prediction cost: the first-order increase in loss when part of an optimal mixture is replaced by that component. An exact loss decomposition bounds the total weight on components with positive cost by the excess loss. As training approaches the optimum, these weights vanish. The decomposition and weight bound appear in Appendix B.

Write $\delta_t(a)$ for the prediction cost and $n_i(a)$ for the count of component a in context i. The proportionality factor for counts on T is c_i. Let $\bar{n}_a = \sum_\ell \rho_\ell n_\ell(a)/c_\ell$, where ρ_ℓ is the training weight of context ℓ. For score gradient descent with fixed positive-definite preconditioner P, assume positive costs outside T and positive curvature along mixing directions within T. With a sufficiently small fixed step size η and an initialization in a low-excess-loss region,

$$\eta k\,(h_i^k - h_j^k) \longrightarrow g_{ij},$$
$$g_{ij} = \sum_{a \notin T} \chi_a \frac{n_i(a)/c_i - n_j(a)/c_j}{\delta_t(a)\bar{n}_a}(r + b_a - h^*). \tag{2}$$

Here h^* is the common optimal representation. The coefficients χ_a depend on P; for ordinary score gradient descent, $P = I$ gives $\chi_a = |T|/m$ for every $a \notin T$. The formula gives the residual directions and their coefficients at order $1/k$. Larger prediction costs suppress a component's contribution, while differences in normalized counts determine how strongly it separates two contexts. The update-dependent coefficients χ_a determine which components remain at this order. Theorem 1 in Appendix D characterizes χ_a and gives the full conditions and proof, including the attention-matrix update.

4 Synthetic Experiments

Predicting the remaining representation difference. We train the query matrix of a single-head attention model with width 32, keeping the keys, values, and prediction head fixed. Two contexts share the optimal support and differ in their residual-component counts. The limiting coefficients and difference vector are computed from the model parameters before training. Figure 2 shows the trajectories approaching these predictions. The scaled difference vector approaches its predicted limit with 3.8% relative error, and the late-training distance slope is -0.973, close to the predicted -1. The agreement supports the predicted vector limit and contraction rate. Appendix G reports the remaining numerical details.

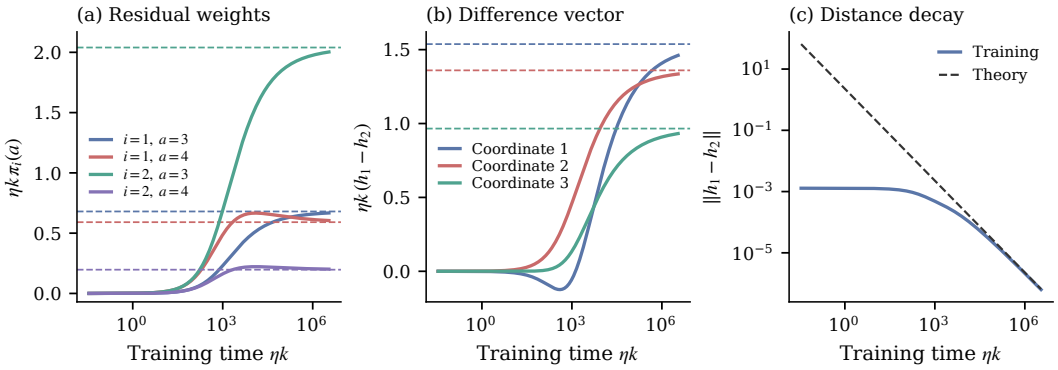


Figure 2: **Training approaches the predicted residual geometry.** Solid curves show query-matrix gradient descent; dashed curves show predictions computed before training. The panels display scaled residual weights, scaled representation differences, and distance decay. The horizontal axis is effective training time ηk.

Shared components after full-parameter training. We also train a two-layer Transformer to return the value of a queried field. Each value has two token encodings, allowing us to compare

inputs with the same answer but different target tokens. On held-out field-value combinations, we pair each anchor with two variants. One preserves the target token and changes the counts of the other components. The other replaces the target token with its alternative encoding, preserving the remaining input. After training, the shared-token pair is closer in 99.5% of comparisons (Figure 3). The resulting geometry distinguishes how inputs encode the same answer. Appendix G gives both experimental settings.

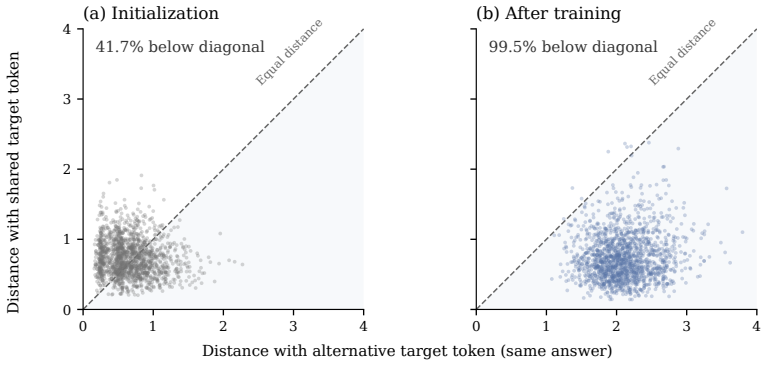


Figure 3: **Shared target components bring same-answer representations closer.** Each point compares two variants of one anchor. The horizontal axis changes the target token's encoding; the vertical axis changes non-target counts while preserving that token. Points below the diagonal have a closer shared-token variant. Both variants retain the answer.

5 Pre-Generation Hallucination Detection

Component mixtures and local reconstruction. Equation 1 identifies the component weights that control representation differences. Write $O = T^c$, $G_T = [b_a - v_T]_{a \in O}$, and $\pi_{i,O} = (\pi_i(a))_{a \notin T}$. Under the shared-mixture condition of Section 3, each representation has the form $h_i = r + v_T + G_T \pi_{i,O}$. For k reference contexts, let $\bar{h} = k^{-1} \sum_i h_i$ and $\bar{w}_i = 1/k$. Any probability vector $w \in \Delta_k = \{w \geq 0 : \sum_i w_i = 1\}$ then satisfies

$$\sum_i w_i h_i - \bar{h} = G_T \sum_i (w_i - \bar{w}_i) \pi_{i,O}. \tag{3}$$

Equation 3 expresses reference-state reweighting as a change in the aggregate component weights. The training analysis characterizes the scales and directions of representation differences generated by these weights. Local reconstruction measures the reweighting needed to explain a test state and the residual left outside the resulting mixture.

The weight cost also has a component interpretation. Let $\pi(w) = \sum_i w_i \pi_i$ and $\bar{\pi} = \sum_i \bar{w}_i \pi_i$ be the induced component distributions. The log-sum inequality gives

$$\mathrm{KL}\big(\pi(w) \| \bar{\pi}\big) \leq \mathrm{KL}(w \| \bar{w}). \tag{4}$$

The reference-weight penalty therefore bounds the change in the induced component distribution. Together, Eqs. 3 and 4 connect the reconstruction and weight cost to the component structure of the analytical model. The assumptions in Section 3 make the representation dynamics analytically tractable. The detector does not require them. It uses only the weaker qualitative mechanism exposed by the analysis: local representation variation can be modeled through reweighting nearby reference states, with outcome labels providing the class-specific reference sets. Appendix E proves these identities and derives their local form after representation normalization.

Local component reuse in a pretrained LLM. We test whether reweighting fixed value vectors can explain changes in the attention contribution of aligned tokens across paraphrases. The experiment uses 411 groups of TriviaQA questions with Qwen2.5-7B-Instruct, each containing an original question and two paraphrases. At layer 24, we align shared content tokens and fix their projected value vectors to those of the original question. Let $B_A^{(0)}$ contain these vectors across attention heads,

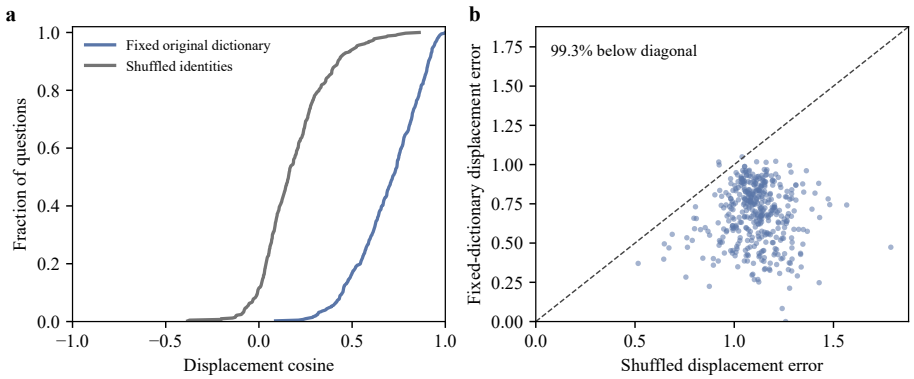


Figure 4: **Predicting attention changes with a fixed component dictionary.** The original question supplies the value vectors, and the two paraphrases supply the attention-weight difference. **(a)** Distribution of directional cosine across question groups. **(b)** Relative prediction error for each group. Points below the diagonal favor the aligned dictionary. The control averages 16 component identity shuffles per group.

and let $\alpha_A^{(1)}, \alpha_A^{(2)}$ be their attention weights in the two paraphrases at the final prompt token. We predict the change in the aligned attention contribution by

$$\widehat{\Delta o_A} = B_A^{(0)}\big(\alpha_A^{(2)} - \alpha_A^{(1)}\big). \tag{5}$$

The fixed dictionary explains 49.8% of the observed change energy. Shuffling component identities substantially increases prediction error (Figure 4), indicating that the correspondence between weights and value vectors matters. The result supports partial component reuse in these attention contributions. Appendix E.2 gives the protocol, confidence intervals, and full results.

Detection task and labeled references. Fix a trained model and an answer-generation procedure. For a question x, let $h(x)$ denote a hidden state computed before the first answer token. Let $Y(x) = 0$ indicate a correct answer and $Y(x) = 1$ an incorrect answer under the benchmark's answer-matching rule. These outcome labels partition a reference set into correct-answer and incorrect-answer classes. Each class supplies the reference states for a local reconstruction. Section 6.1 specifies the labels and evaluation partitions for text-only and visual questions.

Local mixture energy. We extract the final prompt-token state after the model's final normalization, subtract the reference-state mean, and normalize to unit length. For a test state z, cosine distance selects k reference neighbors from each class $c \in \{0, 1\}$. Write their states as $z_{c,1}, \ldots, z_{c,k}$ and $Z_c = [z_{c,1} \ \cdots \ z_{c,k}]$. Let τ be the mean pairwise cosine distance within the reference set and set $\varepsilon = \lambda_{\text{rec}}\tau$, with $\lambda_{\text{rec}} > 0$. Define

$$E_c(z) = \min_{w \in \Delta_k} \left\{ 2 \sum_{i=1}^{k} w_i \log(k w_i) + \frac{1}{\varepsilon} \left\| z - Z_c w \right\|_2^2 \right\}. \tag{6}$$

The first term penalizes departure from the uniform local mixture. The second measures the part of the test state that the mixture cannot reconstruct. Nonnegative weights that sum to one preserve the mixture interpretation in Eq. 3. The parameter ε sets the relative cost of reconstruction error.

The error-risk score is

$$s(z) = \tau\big[E_0(z) - E_1(z)\big]. \tag{7}$$

A positive score means that the incorrect-answer class reconstructs the test state at lower cost than the correct-answer class. We solve the convex optimization problem separately for each test state and class. The model is frozen, and the detector requires no classifier training. Appendix E.3 gives the solver and scoring procedure.

Role of the reference labels. Outcome labels anchor the local mixtures to observed answer correctness. TruthfulQA shows that language models can reproduce false beliefs learned from human

text (Lin et al., 2022). Farquhar et al. (2024) distinguish such consistent errors from confabulations that vary across sampled answers. An incorrect answer can receive high probability and recur across generations, so confidence and agreement alone can miss it. Labeled references place states associated with these errors in the incorrect-answer class. The detector can then assess a test state against these observed error patterns through its class-specific reconstruction energy.

6 Hallucination Detection Experiments

The experiments test whether local geometry of pre-generation states distinguishes hallucinated from non-hallucinated answers across language and vision-language tasks. Comparisons include answer-based scores, k-nearest-neighbor (kNN) label voting, and a supervised pre-generation probe.

6.1 Experimental setup

Datasets and models. The language tasks are HotpotQA in its distractor setting (Yang et al., 2018), MMLU (Hendrycks et al., 2021), TriviaQA (Joshi et al., 2017), and TruthfulQA in its binary multiple-choice setting (Lin et al., 2022). The language models are Qwen2.5-3B-Instruct, Qwen2.5-7B-Instruct (Yang et al., 2024), and Llama-3.1-8B-Instruct (Grattafiori et al., 2024). The visual tasks are CV-Bench (Tong et al., 2024), GQA (Hudson & Manning, 2019), and the adversarial split of POPE (Li et al., 2023). The vision-language models are Qwen2.5-VL-3B-Instruct, Qwen2.5-VL-7B-Instruct (Bai et al., 2025), and InternVL3-9B (Zhu et al., 2025). This gives 21 model–dataset combinations. Dataset sizes and generation settings appear in Appendix F.

Evaluation protocol. We use 400 labeled reference questions for TruthfulQA, 1,002 for POPE-Adv, and 1,000 for each other task. The remaining questions form an independent test set. Visual reference and test partitions are disjoint by image. We use a fixed neighbor count for each dataset, shared across its three models, and set the residual scale using the labeled reference set. Each test state is scored using only the reference bank. Appendix F gives the split sizes and neighborhood construction.

Baselines. The answer-based baselines are Perplexity, length-normalized entropy (LN-Entropy), Semantic Entropy, Lexical Similarity (Kuhn et al., 2023), and EigenScore (Chen et al., 2024). Perplexity uses the greedy answer. The other four baselines share five sampled answers per question. We include kNN with $k = 5$, which scores a test state by the fraction of incorrect-answer labels among its five nearest references. It uses the same centered, unit-normalized states and cosine distance as our method. The supervised probe of Ji et al. (2024) uses the same reference labels and independent test questions. All methods in Table 1 are evaluated on the same test sets.

Hallucination labels. All methods predict the hallucination label of the same greedy response. Answers failing the following task-specific matching rules are labeled hallucinated. The matching rules use parsed option letters for multiple-choice tasks and yes/no answers for POPE. HotpotQA uses ROUGE-L F1 with stemming against the reference answer, with a score above 0.5 counted as correct. TriviaQA and GQA use normalized exact match against the available reference answers.

Metric. Hallucination-detection AUROC (%) measures ranking performance. Higher values are better. Averages weight model–dataset combinations equally.

6.2 Main results

Table 1 reports AUROC on the independent test sets. Local mixture energy averages 79.26 across the 21 combinations, exceeding the supervised probe by 1.81 points. It also improves on five-neighbor label voting in every combination. This consistent gain shows that local reconstruction uses information beyond the labels of the nearest references.

The strongest results are on visual tasks, with a mean AUROC of 81.65 and a 2.35-point gain over the probe. The method leads on every CV-Bench and POPE-Adv evaluation, and its average advantage over the probe holds across all three vision-language models. GQA is more mixed, with Perplexity and the probe leading for two of the models.

Table 1: **Hallucination detection on independent test questions.** AUROC (%). Ours, kNN, and the Ji probe use the same labeled reference set. All methods are evaluated on the same remaining questions. Resp. is the number of generated answers used for scoring. Bold marks the highest value in each column within a model block.

Method	Resp.	Language tasks					Visual tasks			
		HotpotQA	MMLU	TriviaQA	TruthfulQA	Avg.	CV-Bench	GQA	POPE-Adv	Avg.
Qwen2.5-3B-Instruct							*Qwen2.5-VL-3B-Instruct*			
Perplexity	1	66.54	78.34	83.61	57.05	71.39	71.69	**77.53**	73.47	74.23
LN-Entropy	5	66.70	74.98	**83.85**	52.98	69.63	71.39	77.45	74.98	74.61
Semantic Entropy	5	51.23	51.04	74.72	48.39	56.34	55.75	61.45	66.88	61.36
Lexical Similarity	5	55.59	54.40	76.62	51.01	59.40	65.67	66.52	66.47	66.22
EigenScore	5	53.29	50.41	75.88	48.78	57.09	63.58	73.52	62.96	66.69
kNN ($k = 5$)	0	59.32	76.76	70.74	79.21	71.51	75.08	73.18	73.44	73.90
Ji probe	0	63.45	78.73	76.23	84.00	75.60	74.37	76.27	82.13	77.59
Ours	0	**66.92**	**80.63**	76.55	**84.35**	**77.11**	**79.31**	76.26	**83.36**	**79.64**
Qwen2.5-7B-Instruct							*Qwen2.5-VL-7B-Instruct*			
Perplexity	1	68.09	79.22	**82.49**	65.01	73.70	79.52	70.48	74.98	74.99
LN-Entropy	5	68.15	79.20	**82.49**	65.01	73.71	78.66	70.06	74.73	74.48
Semantic Entropy	5	51.73	50.76	70.11	50.86	55.86	57.86	60.59	58.53	58.99
Lexical Similarity	5	55.56	52.69	73.54	51.73	58.38	69.65	64.49	58.51	64.22
EigenScore	5	53.19	44.30	70.31	50.65	54.61	63.38	68.65	64.85	65.63
kNN ($k = 5$)	0	66.30	76.68	72.61	77.08	73.17	78.92	73.17	79.75	77.28
Ji probe	0	66.63	80.03	75.21	83.03	76.22	82.08	74.64	83.17	79.96
Ours	0	**70.12**	**81.63**	77.31	**86.56**	**78.91**	**84.00**	**77.15**	**85.32**	**82.16**
Llama-3.1-8B-Instruct							*InternVL3-9B*			
Perplexity	1	66.68	74.19	79.40	58.82	69.78	84.03	75.95	74.55	78.18
LN-Entropy	5	66.79	73.90	**82.02**	55.68	69.60	83.22	75.73	75.10	78.02
Semantic Entropy	5	53.43	52.62	71.45	45.33	55.71	61.23	64.69	65.59	63.84
Lexical Similarity	5	64.83	66.59	80.10	53.17	66.17	70.58	65.88	63.57	66.68
EigenScore	5	60.61	55.83	72.12	45.96	58.63	67.80	73.51	67.94	69.75
kNN ($k = 5$)	0	63.06	76.61	69.17	81.11	72.49	82.50	72.86	75.76	77.04
Ji probe	0	**69.30**	**80.36**	74.71	81.22	76.39	81.38	**78.26**	81.42	80.35
Ours	0	68.98	80.22	74.99	**81.39**	**76.40**	**87.25**	77.04	**85.19**	**83.16**

On language tasks, performance varies by dataset. Local mixture energy leads on TruthfulQA for all three models and on HotpotQA and MMLU for both Qwen models. The probe leads on HotpotQA and MMLU with Llama-3.1-8B, while likelihood-based scores lead on TriviaQA. Thus the relative usefulness of reference geometry and answer likelihood depends on the task.

TruthfulQA is informative because it targets errors associated with common misconceptions (Lin et al., 2022). Local mixture energy averages 84.10, compared with 60.29 for the strongest answer-based baseline, Perplexity. This gap supports the value of outcome-labeled references for questions where confidence and agreement can miss errors. The method also exceeds both kNN and the supervised probe for each model, supporting the use of local reconstruction with these reference labels.

Value of local reconstruction. The three pre-generation methods use the same prompt states and labeled reference questions. Their differences in performance therefore reflect how they use these states and labels. The gains over kNN show that class-specific mixture reconstruction provides a more informative error ranking than five-neighbor label voting across these evaluations. The comparison with the supervised probe shows that optimizing local mixtures is effective with the available reference budgets.

7 Discussion

Prediction training can contract representations without erasing the structure of their differences. In the attention model, shared predictive components cancel, while the leading difference records the residual components that training suppresses, their counts across contexts, and the parameter update. These structured differences draw attention to how nearby states combine to represent a new state. The pretrained-LLM experiment finds partial reuse of attention components across paraphrases, supporting a local component description beyond the analytical model. On held-out detection tasks, class-conditional reconstruction ranks errors more effectively than neighbor-label voting on the same

states and labels. The advantage is strongest on visual questions. The useful signal lies in how outcome-labeled states jointly represent a new question before its answer is generated.

Reproducibility statement

The theoretical assumptions and proofs are provided in Appendices A–D. Experimental settings are described in Appendices F, E.2, and G. Code and reproduction materials are available at `https://anonymous.4open.science/r/prediction-geometry-FE5E/`.

AI use statement

Generative AI assisted with manuscript editing and computational checks.

References

Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023a.

Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. In *Advances in Neural Information Processing Systems*, volume 36, pp. 48314–48362. Curran Associates, Inc., 2023b.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL technical report. *CoRR*, abs/2502.13923, 2025.

Arindam Banerjee, Srujana Merugu, Inderjit S. Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6(58):1705–1749, 2005.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *International Conference on Learning Representations*, 2024.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, New York, 2 edition, 1998.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. HaloScope: Harnessing unlabeled LLM generations for hallucination detection. In *Advances in Neural Information Processing Systems*, volume 37, pp. 102948–102972. Curran Associates, Inc., 2024.

Kawin Ethayarajh. How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 55–65. Association for Computational Linguistics, 2019.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630:625–630, 2024.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 30–45. Association for Computational Linguistics, 2022.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021.

Junjie Hu, Gang Tu, ShengYu Cheng, Jinxin Li, Jinting Wang, Rui Chen, Zhilong Zhou, and Dongbo Shan. HARP: Hallucination detection via reasoning subspace projection. In *International Conference on Learning Representations*, 2026.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019.

Ziwei Ji, Delong Chen, Etsuko Ishii, Samuel Cahyawijaya, Yejin Bang, Bryan Wilie, and Pascale Fung. LLM internal states reveal hallucination risk faced with a query. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 88–104. Association for Computational Linguistics, 2024.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611. Association for Computational Linguistics, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations*, 2023.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023.

Yingcong Li, Yixiao Huang, Muhammed E. Ildiz, Ankit Singh Rawat, and Samet Oymak. Mechanics of next token prediction with self-attention. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 685–693. PMLR, 2024.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252. Association for Computational Linguistics, 2022.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017. Association for Computational Linguistics, 2023.

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*, 2025.

Sam T. Roweis and Lawrence K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, I'm wrong: LLMs hallucinate with certainty despite knowing the answer. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 14665–14688. Association for Computational Linguistics, 2025.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 87310–87356. Curran Associates, Inc., 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *CoRR*, abs/2412.15115, 2024.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380. Association for Computational Linguistics, 2018.

Fan Yin, Jayanth Srinivasa, and Kai-Wei Chang. Characterizing truthfulness in large language model generations with local intrinsic dimension. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57069–57084. PMLR, 2024.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

A Attention Model

Shared attention. Let $B = [b_1, \dots, b_m] \in \mathbb{R}^{d\times m}$ be a dictionary of value vectors. Context x_i contains $n_i(a) \in \{0, 1, \dots\}$ copies of component a, with at least one component present. For shared scores $s \in \mathbb{R}^m$, attention forms the representation

$$\pi_i(a; s) = \frac{n_i(a)e^{s_a}}{\sum_{\ell=1}^m n_i(\ell)e^{s_\ell}}, \qquad h_i(s) = r + B\pi_i(s), \tag{8}$$

where r is a fixed residual common to the contexts. The multiplicity $n_i(a)$ accounts for repeated occurrences of a component. In particular, an absent component receives zero attention, and contexts containing different numbers of the same components can have different weights under the same s. Because the score vector is shared, changing s_a affects component a in every context where it occurs. We study score updates with fixed positive-definite preconditioning. Appendix D relates these updates to training an attention matrix with fixed keys and query. We suppress the argument s when it is clear. The common residual isolates the representation differences produced by the attention mixture.

Prediction objective. We fix an output matrix $C \in \mathbb{R}^{K\times d}$ and bias $\beta \in \mathbb{R}^K$. Let $t \in \mathbb{R}^K$ be a target probability vector, and use $\|\cdot\|$ for the Euclidean norm. Set $p(h) = \mathrm{softmax}(Ch + \beta)$ and $f_t(h) = -\sum_{o=1}^K t(o) \log p_o(h)$. The training objective is $L(s) = \sum_i \rho_i f_{t_i}(h_i(s))$, where $\rho_i > 0$ and $\sum_i \rho_i = 1$. The entries of CB specify the contribution of each value vector to each output logit. We allow linearly dependent columns of B and arbitrary alignment between B and C. Output classes and dictionary components are indexed separately, so a target can correspond to an output class absent from the context.

Every h_i lies in $\mathcal{H} = r + \mathrm{conv}\{b_1, \dots, b_m\}$. For each target, choose $h_t^* \in \arg\min_{h\in\mathcal{H}} f_t(h)$ and set $p_t^* = p(h_t^*)$. The hull and output layer together determine h_t^*; the components available in an individual context determine which mixtures that context can form.

B Loss Decomposition and Component Selection

Fix a target t. First-order optimality on $\mathcal{H}$ gives the component cost

$$\delta_t(a) = \left\langle C^\top(p_t^* - t), r + b_a - h_t^* \right\rangle \geq 0.$$

This is the first-order cost of replacing part of the optimal mixture by component a.

Proposition 1 (Component costs and excess loss)**.** *For every $\pi \in \Delta_m$ and $h = r + B\pi$,*

$$f_t(h) - f_t(h_t^*) = \sum_a \pi(a)\delta_t(a) + D_{\mathrm{KL}}(p_t^* \| p(h)). \tag{9}$$

If every positive cost is at least $\gamma_t > 0$, then

$$\sum_{a:\delta_t(a)>0} \pi(a) \leq \frac{f_t(h) - f_t(h_t^*)}{\gamma_t}, \qquad D_{\mathrm{KL}}(p_t^* \| p(h)) \leq f_t(h) - f_t(h_t^*). \tag{10}$$

Proof. The Bregman identity for the log-partition function (Banerjee et al., 2005) yields

$$f_t(h) - f_t(h_t^*) = \left\langle C^\top(p_t^* - t), h - h_t^* \right\rangle + D_{\mathrm{KL}}(p_t^* \| p(h)).$$

Substitute $h - h_t^* = \sum_a \pi(a)(r + b_a - h_t^*)$ to obtain (9). Both terms are nonnegative, and $\sum_a \pi(a)\delta_t(a) \geq \gamma_t \sum_{a:\delta_t(a)>0} \pi(a)$, which gives the bounds. □

Thus the total attention on positive-cost components vanishes with the excess loss.

C Shared Attention and Representation Differences

C.1 Shared conditional mixture and representation differences

Let $h^* = r + B\pi^*$ be a common optimum for target t, with support T. Assume that the component counts on T are proportional across contexts:

$$n_i(a) = c_i n_0(a), \qquad a \in T, \quad c_i, n_0(a) > 0. \tag{11}$$

Their conditional mixture on T is then

$$v_T = \frac{\sum_{a\in T} n_0(a) e^{s_a} b_a}{\sum_{a\in T} n_0(a) e^{s_a}},$$

independent of the context. Subtracting $h_i = r + v_T + \sum_{a\notin T} \pi_i(a)(b_a - v_T)$ for two contexts proves (1). If $\delta_t(a) \geq \gamma_t > 0$ outside T, Proposition 1 also gives

$$\|h_i - h_j\| \leq \frac{\operatorname{diam}(\mathcal{H})}{\gamma_t} \big[f_t(h_i) + f_t(h_j) - 2 f_t(h^*) \big]. \tag{12}$$

The excess losses therefore control the distance, while the surviving residual components determine its leading direction.

C.2 Attainability and the count condition

Differentiating the shared attention weights gives

$$\frac{\partial L}{\partial s_a} = \sum_i \rho_i \pi_i(a) \left\langle r + b_a - h_i, C^\top (p(h_i) - t_i) \right\rangle. \tag{13}$$

Proposition 2 (Common optimum and descent direction)**.** *Suppose all contexts have target t and satisfy* (11) *for an optimal mixture $h_t^* = r + B\pi^*$ with support T. Then $\inf_s L(s) = f_t(h_t^*)$, and every finite score vector with positive excess loss has a strict descent direction. Consequently, every finite stationary point attains the optimum. The same conclusions hold for $s_a = k_a^\top W q$ when $\Xi = [k_1, \ldots, k_m]$ has full column rank and $q \neq 0$.*

Proof. For $M \to \infty$, choose

$$s_a(M) = \begin{cases} \log\big(\pi^*(a)/n_0(a)\big), & a \in T, \\ -M, & a \notin T. \end{cases}$$

On T, the unnormalized weights equal $c_i \pi^*(a)$; all others vanish. Hence $\pi_i(s(M)) \to \pi^*$ for every context and $L(s(M)) \to f_t(h_t^*)$, proving the infimum claim.

At any finite score vector, set

$$d_a = \begin{cases} \pi^*(a)/(n_0(a) e^{s_a}), & a \in T, \\ 0, & a \notin T. \end{cases}$$

Direct differentiation gives

$$\frac{d}{d\epsilon} h_i(s + \epsilon d)\Big|_{\epsilon=0} = \frac{c_i}{\sum_a n_i(a) e^{s_a}} (h_t^* - h_i).$$

By convexity,

$$\begin{aligned} \langle \nabla_s L, d \rangle &= \sum_i \frac{\rho_i c_i}{\sum_a n_i(a) e^{s_a}} \langle \nabla f_t(h_i), h_t^* - h_i \rangle \\ &\leq -\sum_i \frac{\rho_i c_i}{\sum_a n_i(a) e^{s_a}} \big[f_t(h_i) - f_t(h_t^*) \big]. \end{aligned} \tag{14}$$

The right-hand side is strictly negative whenever the excess loss is positive. For full-column-rank Ξ, the map

$$W = \Xi (\Xi^\top \Xi)^{-1} s \, \frac{q^\top}{\|q\|^2}$$

realizes $\Xi^\top W q = s$ and maps d to a parameter direction with the same loss derivative. □

Necessity when the optimal attention is unique. If $L(s^\ell) \to f_t(h_t^*)$, positivity of the context weights and uniqueness imply $\pi_i(s^\ell) \to \pi^*$ for every context, so every component in T must be present. For $a, b \in T$,

$$\frac{\pi_i(a;s)}{\pi_i(b;s)} = \frac{n_i(a)}{n_i(b)} e^{s_a - s_b}.$$

Taking the ratio across two contexts cancels the shared scores. Both attention ratios converge to π_a^*/π_b^*, so the count ratios agree. Fixing one component as an anchor yields $n_i(a) = c_i n_0(a)$ on T, establishing necessity of (11).

D Preconditioned Gradient Descent and Exact Asymptotics

D.1 Assumptions and main theorem

We state the leading weights and representation difference. All contexts have target t. Assume $1 \le |T| < m$, the count condition (11), positive costs outside T, and occurrence of each dictionary component in at least one context. We require the loss to have positive curvature along nonzero mixing directions within T:

$$w^\top B_T^\top \nabla^2 f_t(h^*) B_T w > 0 \quad \text{for } w \neq 0,\ \mathbf{1}^\top w = 0. \tag{15}$$

This condition is empty for $|T| = 1$. Together these assumptions make π^* the unique optimal attention vector.

Consider gradient descent with a fixed positive-definite matrix P:

$$s^{k+1} = s^k - \eta P \nabla_s L(s^k), \qquad P = P^\top \succ 0. \tag{16}$$

The leading residual coefficients depend only on P and T. Define them by the following strictly convex quadratic problem:

$$\xi^* = \arg \min_{\substack{\mathbf{1}^\top \xi = 0 \\ \xi_a \ge 0\ (a \notin T)}} \left\{ \tfrac{1}{2} \xi^\top P \xi - \sum_{a \notin T} \xi_a \right\}, \qquad \chi_a = \xi_a^* \quad (a \notin T). \tag{17}$$

Here ξ is an auxiliary vector; its entries on T are unrestricted. Let $\bar{n}_a = \sum_i \rho_i n_i(a)/c_i$.

Theorem 1 (Residual weights and representation differences)**.** *Under these assumptions, there exist $\varepsilon_0, \eta_0 > 0$ such that any finite initialization with $L(s^0) - f_t(h^*) < \varepsilon_0$ and any fixed step size $0 < \eta \le \eta_0$ give $\pi_i(s^k) \to \pi^*$ and*

$$\eta k\, \pi_i(a; s^k) \longrightarrow \chi_a \frac{n_i(a)/c_i}{\delta_t(a) \bar{n}_a}, \qquad a \notin T. \tag{18}$$

For any two contexts,

$$\eta k\, (h_i^k - h_j^k) \longrightarrow \sum_{a \notin T} \chi_a \frac{n_i(a)/c_i - n_j(a)/c_j}{\delta_t(a) \bar{n}_a} (r + b_a - h^*). \tag{19}$$

A step-size bound and the low-loss convergence argument are given in Section D.2; the proof follows in the remaining subsections.

Equation (18) separates three contributions to the residual weights. Larger prediction costs suppress a component more strongly. Normalized counts determine its relative contribution to each context. The coefficient χ_a captures the effect of the parameter update: components with $\chi_a > 0$ retain weights of order $1/k$ wherever they occur, while those with $\chi_a = 0$ decay faster.

Equation (19) then gives the leading geometry. Differences in component counts weight the directions $r + b_a - h^*$. The representation difference shrinks at rate $O(1/(\eta k))$; the vector limit identifies the part that remains at this scale. For ordinary score gradient descent, $P = I$ gives $\chi_a = |T|/m$ for every $a \notin T$.

For an attention matrix W with scores $s = \Xi^\top W q$, the chain rule gives $\nabla_W L = \Xi(\nabla_s L) q^\top$. A gradient step on W therefore induces

$$s^{k+1} = s^k - \eta \|q\|^2 \Xi^\top \Xi \nabla_s L(s^k).$$

With fixed keys Ξ of full column rank and a fixed nonzero query q, the theorem applies with

$$P = \|q\|^2 \Xi^\top \Xi \succ 0. \tag{20}$$

Thus the key geometry enters through the same residual coefficients.

The right-hand side of Equation (19) is the vector g_{ij} in Equation (2).

For the proof, write $\mathcal{E}(s) = L(s) - f_t(h^*)$, $\mathcal{E}_k = \mathcal{E}(s^k)$, and $O = T^c$.

D.2 Smoothness and convergence in the low-loss region

The loss decomposition forces every optimum to be supported on T. If a second optimal mixture existed, convexity would make the segment joining it to π^* optimal, contradicting the positive curvature in (15). Thus π^* is unique.

A uniform score-space smoothness bound is

$$\beta_s = \tfrac{1}{8}\|CB\|_2^2 + 3\sqrt{2}\,\mathrm{diam}(\mathcal{H})\,\|C\|_2\,, \qquad 0 < \eta \le \frac{1}{\beta_s \lambda_{\max}(P)}. \tag{21}$$

Indeed, with $J_i = \mathrm{diag}(\pi_i) - \pi_i \pi_i^\top$,

$$\begin{aligned}\nabla_s^2 f_t(h_i(s)) = J_i B^\top \nabla^2 f_t(h_i) B J_i + \mathrm{diag}\big(\nabla_s f_t(h_i(s))\big) \\ - \nabla_s f_t(h_i(s))\pi_i^\top - \pi_i \nabla_s f_t(h_i(s))^\top.\end{aligned}$$

The first term has norm at most $\|CB\|_2^2/8$, since both softmax covariance norms are at most $1/2$. By (13), each remaining term has norm at most $\sqrt{2}\,\mathrm{diam}(\mathcal{H})\,\|C\|_2$. Averaging these bounds gives (21). Gradient descent satisfies

$$\begin{aligned}\mathcal{E}_{k+1} &\le \mathcal{E}_k - \frac{\eta}{2}\nabla L(s^k)^\top P \nabla L(s^k) \\ &\le \mathcal{E}_k - \frac{\eta \lambda_{\min}(P)}{2}\left\|\nabla L(s^k)\right\|^2.\end{aligned} \tag{22}$$

Compactness and uniqueness give an $\varepsilon_0 > 0$ such that $\mathcal{E}(s) < \varepsilon_0$ implies $\pi_i(a) > \pi_a^*/2$ for all i and $a \in T$. Descent keeps the iterates in this region. Choose $d_a = \pi_a^*/\pi_1(a)$ on T and zero elsewhere. The common conditional mixture gives

$$J_i d = \frac{\sum_{a\in T}\pi_i(a)}{\sum_{a\in T}\pi_1(a)}(\pi^* - \pi_i), \qquad \|d\| \le 2\sqrt{|T|}.$$

The ratio is at least $1/2$, so convexity yields $\langle \nabla L, d\rangle \le -\mathcal{E}/2$ and $\|\nabla L\| \ge \mathcal{E}/(4\sqrt{|T|})$. Substitution into (22) gives

$$\mathcal{E}_k \le \left[\mathcal{E}_0^{-1} + \frac{\eta\lambda_{\min}(P)}{32|T|}k\right]^{-1}. \tag{23}$$

Every context has positive training weight, so its excess loss tends to zero and $\pi_i(s^k) \to \pi^*$. Appendix C constructs finite initializations with arbitrarily small excess loss.

D.3 Score coordinates and the residual coefficients

Let $F \in \mathbb{R}^{m\times(|T|-1)}$ have orthonormal columns spanning the zero-sum directions supported on T, and let $Q = [e_a - \mathbf{1}_T/|T|]_{a\in O}$. Define the metric blocks and their Schur complement by

$$A = F^\top P F, \qquad R = F^\top P Q, \qquad M = Q^\top P Q - R^\top A^{-1} R. \tag{24}$$

Since $[F, Q]$ spans $\mathbf{1}^\perp$, its induced metric and hence M are positive definite. When $|T| = 1$, the F coordinates are absent and $M = Q^\top P Q$. Eliminating the coordinates along F in (17) gives

$$\chi = \arg\min_{u\ge 0}\left\{\tfrac{1}{2}u^\top M u - \mathbf{1}^\top u\right\}, \qquad M\chi \ge \mathbf{1}, \quad \chi_a[(M\chi)_a - 1] = 0. \tag{25}$$

For $P = I$, $M = I + \mathbf{1}\mathbf{1}^\top/|T|$ and $\chi = (|T|/m)\mathbf{1}$.

Set $x = F^\top s$, $y = Q^\top s$, and $v_a = e^{y_a}$ for $a \in O$. Subtracting the mean score on T gives $s_T = (Fx)_T$ and $s_O = y$. The attention weights are

$$\begin{aligned} D_i(x,v) &= \sum_{a\in T} n_0(a) e^{(Fx)_a} + \sum_{a\in O} \frac{n_i(a)}{c_i} v_a, \\ \pi_i(a) &= \begin{cases} n_0(a) e^{(Fx)_a}/D_i(x,v), & a \in T, \\ n_i(a) v_a/[c_i D_i(x,v)], & a \in O. \end{cases} \end{aligned} \tag{26}$$

In these coordinates the excess loss $\mathcal{E}(x,v)$ is smooth near $(x^*,0)$, where x^* gives π^* on T. Attention convergence implies $x_k \to x^*$ and $v_k \to 0$. With $\nabla_y \mathcal{E} = \mathrm{diag}(v)\nabla_v \mathcal{E}$, the chain rule gives $\nabla_s L = F\nabla_x \mathcal{E} + Q\nabla_y \mathcal{E}$ and

$$\begin{aligned} x^+ - x &= -\eta(A\nabla_x \mathcal{E} + R\nabla_y \mathcal{E}), \\ y^+ - y &= -\eta(R^\top \nabla_x \mathcal{E} + Q^\top P Q \nabla_y \mathcal{E}). \end{aligned} \tag{27}$$

D.4 Eliminating the coupling to the optimal support

The aggregate residual costs satisfy

$$U_a = \delta_t(a) \sum_i \rho_i \pi_i(a) = \delta_t(a) v_a \sum_i \frac{\rho_i n_i(a)}{c_i D_i(x,v)}, \qquad a \in O. \tag{28}$$

The loss decomposition is

$$\mathcal{E} = \mathbf{1}^\top U + \sum_i \rho_i D_{\mathrm{KL}}(p_t^* \| p(h_i)). \tag{29}$$

Thus $U_k = O(k^{-1})$ by (23). Since $U_a/v_a \to \delta_t(a)\bar{n}_a/D_1(x^*,0) > 0$, also $v_k = O(k^{-1})$.

On the boundary $v = 0$, the curvature assumption and invertibility of the softmax Jacobian on zero-sum directions give $\nabla_x^2 \mathcal{E}(x^*,0) \succ 0$. The step-size bound ensures

$$\eta \lambda_{\max}\Big(A^{1/2} \nabla_x^2 \mathcal{E}(x^*,0) A^{1/2}\Big) \le \eta \lambda_{\max}(P) \beta_s \le 1.$$

Hence the boundary update is locally contractive in the A^{-1} norm. Smoothness and (27) give, for some $\vartheta < 1$ and all sufficiently large k,

$$\|x_{k+1} - x^*\|_{A^{-1}} \le \vartheta \|x_k - x^*\|_{A^{-1}} + C \|v_k\|.$$

Summing the geometric recursion yields $x_k - x^* = O(k^{-1})$.

To cancel the coupling to these coordinates, define the corrected costs

$$\widehat{U}_a = \frac{\delta_t(a)\bar{n}_a}{D_1(x^*,0)} \exp\big(y_a - [R^\top A^{-1}(x - x^*)]_a\big).$$

All $D_i(x^*,0)$ are equal. Together with the rates above and (28), this gives

$$U_{a,k}/\widehat{U}_{a,k} = 1 + O(k^{-1}), \qquad \widehat{U}_k = O(k^{-1}), \qquad U_k - \widehat{U}_k = O(k^{-2}).$$

Differentiating (29) gives

$$\nabla_y \mathcal{E} = U + O\big(\|v\|\,(\|x - x^*\| + \|v\|)\big) = \widehat{U} + O(k^{-2}).$$

Here the KL gradient vanishes at $(x^*,0)$, and differentiation with respect to y_a contributes a factor v_a. The two coordinate updates therefore yield

$$\log \frac{\widehat{U}_{a,k+1}}{\widehat{U}_{a,k}} = -\eta(M\nabla_y \mathcal{E})_a = -\eta(M\widehat{U}_k)_a + O(k^{-2}). \tag{30}$$

D.5 Convergence of the residual coefficients

Rescale the corrected costs as $V_k = \eta k \widehat{U}_k$ for $k \geq 1$. This sequence is positive and bounded, and (30) gives

$$\begin{aligned} V_{a,k+1} - V_{a,k} &= \frac{V_{a,k}}{k}[1 - (MV_k)_a] + O(V_{a,k}k^{-2}), \\ \log \frac{V_{a,k+1}}{V_{a,k}} &= \frac{1}{k}[1 - (MV_k)_a] + O(k^{-2}). \end{aligned} \tag{31}$$

For the Lyapunov function

$$\mathcal{V}(V) = \sum_{\chi_a > 0} [V_a - \chi_a - \chi_a \log(V_a/\chi_a)] + \sum_{\chi_a = 0} V_a \geq 0,$$

these recurrences and (25) imply

$$\begin{aligned} \mathcal{V}(V_{k+1}) - \mathcal{V}(V_k) &= \frac{1}{k}(V_k - \chi)^\top (\mathbf{1} - MV_k) + O(k^{-2}) \\ &= -\frac{1}{k}(V_k - \chi)^\top M(V_k - \chi) - \frac{1}{k}V_k^\top (M\chi - \mathbf{1}) + O(k^{-2}) \\ &\leq -\frac{\lambda_{\min}(M)}{k} \left\| V_k - \chi \right\|^2 + O(k^{-2}). \end{aligned} \tag{32}$$

Hence $\sum_k \left\| V_k - \chi \right\|^2 / k < \infty$. Also $\|V_{k+1} - V_k\| \leq C/k$. An excursion between distances ϵ and $\epsilon/2$ from χ requires harmonic time $\sum 1/k \geq \epsilon/(2C)$, so infinitely many such excursions would contradict summability. Remaining outside the smaller ball would do so as well. Thus $V_k \to \chi$, and $U_{a,k}/\widehat{U}_{a,k} \to 1$ gives

$$\eta k U_k \longrightarrow \chi. \tag{33}$$

Finally, canceling v_a in (26) and (28) gives

$$\frac{\pi_i(a)}{U_a} \longrightarrow \frac{n_i(a)}{c_i \delta_t(a) \bar{n}_a}, \qquad a \in O,$$

which proves (18). Substitution into (1), with $v_T^k \to h^* - r$, proves (19). The rates $x_k - x^*, v_k = O(k^{-1})$ also give $h_i^k - h^* = O(k^{-1})$, so the output KL terms are $O(k^{-2})$ and

$$\eta k \mathcal{E}_k \longrightarrow \mathbf{1}^\top \chi.$$

E Local Mixture Energy

E.1 Connection to component mixtures

Under the shared-mixture condition, $h_i = r + v_T + G_T \pi_{i,O}$. Since $\sum_i (w_i - \bar{w}_i) = 0$,

$$\sum_i (w_i - \bar{w}_i) h_i = G_T \sum_i (w_i - \bar{w}_i) \pi_{i,O},$$

which proves (3). Applying the log-sum inequality over the reference index gives

$$\begin{aligned} \mathrm{KL}\left(\sum_i w_i \pi_i \,\middle\|\, \sum_i \bar{w}_i \pi_i \right) &\leq \sum_{i,a} w_i \pi_i(a) \log \frac{w_i}{\bar{w}_i} \\ &= \mathrm{KL}(w \| \bar{w}), \end{aligned}$$

proving (4). We use $0 \log 0 = 0$; $\bar{w}_i = 1/k > 0$ ensures that the uniform mixture covers the support of every reweighted mixture.

Normalized representations. Let ϕ be the normalization map with the reference mean held fixed. If its Jacobian is L-Lipschitz near $\bar{h}$ and $\left\| h_i - \bar{h} \right\| \leq \rho$, Taylor's theorem and (3) give

$$\left\| \sum_i (w_i - \bar{w}_i) \phi(h_i) - J_\phi(\bar{h}) G_T \sum_i (w_i - \bar{w}_i) \pi_{i,O} \right\| \leq \frac{L\rho^2}{2} \left\| w - \bar{w} \right\|_1 .$$

Thus reference reweighting retains a local component interpretation after normalization, with a remainder controlled by the neighborhood radius.

Table 2: Fixed-dictionary prediction of changes between two paraphrases. Entries are question-group means with 95% bootstrap confidence intervals. Explained energy is averaged per question.

Metric	Fixed original dictionary	Shuffled identities
Directional cosine ↑	$0.696\ [0.678, 0.713]$	$0.186\ [0.168, 0.203]$
Relative error ↓	$0.682\ [0.664, 0.700]$	$1.110\ [1.097, 1.123]$
Explained energy (%) ↑	$49.8\ [47.5, 52.3]$	$-25.6\ [-28.6, -22.8]$

E.2 Local component reuse in a pretrained LLM

Questions and activation measurements. We use Qwen2.5-7B-Instruct on TriviaQA validation questions. A fixed split contains 128 development question groups and 512 test groups. Each group contains an original question and two paraphrases generated greedily by the same model. The paraphrases pass a semantic-equivalence check by that model, with reference answers withheld from the check. This experiment uses the test groups. The prompt asks for a short answer, and all measurements are taken at the last prompt token before answer generation.

We record attention weights and value vectors at layer 24. We match case-folded content words that occur exactly once in each of the three question texts and have identical token-ID sequences. Matching is restricted to the user-question span. Requiring at least two matched words retains 411 of the 512 test groups. These positions carry 12.13% of total attention mass, averaged over heads and the two paraphrases.

Fixed-dictionary prediction. Index each component by an attention head h and aligned token position $t \in A$. For variant $j \in \{0, 1, 2\}$, let $b_{h,t}^{(j)} = W_O^{(h)} v_{h,t}^{(j)}$ be its projected value vector, where $j = 0$ denotes the original question. The measured contribution is

$$o_A^{(j)} = \sum_h \sum_{t \in A} \alpha_{h,t}^{(j)} b_{h,t}^{(j)}.$$

The observed change is $\Delta o_A = o_A^{(2)} - o_A^{(1)}$. Equation 5 fixes the dictionary to $\{b_{h,t}^{(0)}\}$ and uses the observed attention-weight difference between the two paraphrases. Both target value vectors contribute only to the measured change. The prediction uses the frozen original-question dictionary and involves no parameter fitting. This measures local reuse of aligned attention components across inputs. The output projection is applied before summing over heads and measuring errors.

Control and aggregation. For each question, we draw 16 seeded random cycles that permute the aligned token identities without fixed points. The same permutation is applied across all heads and both paraphrases. The control retains the dictionary vectors and attention-weight differences while changing their pairing. For a predicted change $\widehat{\Delta o_A}$, we report directional cosine, relative error $e_{\mathrm{rel}} = \|\widehat{\Delta o_A} - \Delta o_A\| / \|\Delta o_A\|$, and explained change energy $1 - e_{\mathrm{rel}}^2$. Negative explained energy means the prediction error exceeds that of predicting zero change. We average each control metric across the 16 permutations within a question, then average across question groups. All 95% confidence intervals use 5,000 bootstrap resamples of the 411 question groups.

Table 2 gives each question group equal weight. The aligned dictionary predicts the changes more accurately than the shuffled control, supporting the local fixed-component approximation.

E.3 Convex solver and scoring procedure

For a neighborhood matrix $Z = [z_1, \ldots, z_k]$, consider

$$\mathcal{F}(v) = \log\left(\frac{1}{k}\sum_{i=1}^{k} e^{z_i^\top v}\right) - z^\top v + \frac{\varepsilon}{2}\|v\|^2.$$

Fix the retrieved neighborhood. Let R be uniform on its k columns and let $G \sim \mathcal{N}(0, \varepsilon I)$ be independent of R. The log moment-generating function of $R+G$ is $\mathcal{F}(v)+z^\top v$. The corresponding

Cramér rate function (Dembo & Zeitouni, 1998) satisfies

$$\Lambda^*_{R+G}(z) = -\min_v \mathcal{F}(v) = \frac{1}{2}E_c(z).$$

The factor 2 in Eq. 6 therefore makes the local energy twice this rate function. The neighborhood is selected using z before this identity is applied.

This strictly convex dual objective has a unique minimizer. Define $w_i(v) = e^{z_i^\top v}/\sum_j e^{z_j^\top v}$. Its derivatives are

$$\nabla\mathcal{F}(v) = Zw(v) - z + \varepsilon v, \qquad \nabla^2\mathcal{F}(v) = Z\big(\mathrm{diag}(w) - ww^\top\big)Z^\top + \varepsilon I.$$

At the optimum, $\varepsilon v = z - Zw(v)$. The simplex stationarity condition for Eq. 6 gives $w_i \propto \exp\big(z_i^\top(z - Zw)/\varepsilon\big)$, so $w(v)$ is the optimal mixture.

We use damped Newton steps in the span of the centered neighbor states. The residual orthogonal to this span remains in the primal energy. Computation uses float64 arithmetic and at most 100 iterations. We stop when the primal–dual gap divided by $1 + |E_c|$ is below 10^{-9}. Eigenvalues of the centered neighbor Gram matrix below 10^{-9} are treated as numerical zeros when forming the optimization coordinates. Final energies are evaluated in the primal objective.

The scoring procedure is to normalize the frozen prompt state using the reference mean, retrieve the nearest neighbors in each labeled class, solve the two local mixture problems, and return $\tau(E_0 - E_1)$. All model parameters remain fixed. The optimization variables are the mixture weights for the current test state.

F Hallucination Detection Experiments

F.1 Questions and generation

Visual questions use their original benchmark images.

Table 3 records the question collections used for each model. HotpotQA, MMLU, TriviaQA, and GQA use sampled subsets.

Table 3: Question counts per model and maximum generated answer lengths. Reference and test sets partition the complete collection.

Dataset	Questions	Reference	Test	Maximum tokens
HotpotQA	3,000	1,000	2,000	64
MMLU	3,000	1,000	2,000	16
TriviaQA	3,000	1,000	2,000	64
TruthfulQA (binary)	790	400	390	64
CV-Bench	2,638	1,000	1,638	16
GQA	3,000	1,000	2,000	64
POPE-Adv	3,000	1,002	1,998	16

For HotpotQA, we sample from the official development distractor split. All three language models share the reference/test partition in Table 3. Each prompt includes all ten original titled paragraphs in their original order, followed by the question and a request for the shortest answer phrase.

For TruthfulQA, each question presents the supplied best answer and best incorrect answer in randomized A/B order. We randomly partition the questions into reference and test sets before inference. Correctness is determined from the option selected by greedy decoding.

We use the models' official chat templates and bfloat16 inference. Prompts request an option letter for MMLU, TruthfulQA, and CV-Bench, a yes/no answer for POPE, and a short answer for HotpotQA, TriviaQA, and GQA.

F.2 Neighborhood construction

We center final normalized prompt states by the reference-set mean and normalize each vector to unit length. This preprocessing is fixed within each model–dataset evaluation. A test state retrieves

Table 4: Fixed neighbor counts shared across models within each dataset, with model-specific residual multipliers λ_{rec}. Each class supplies k neighbors. The three λ_{rec} columns correspond to Qwen2.5-3B, Qwen2.5-7B, and Llama-3.1-8B for language tasks, and Qwen2.5-VL-3B, Qwen2.5-VL-7B, and InternVL3-9B for visual tasks.

Dataset	k	$\lambda_{\mathrm{rec},1}$	$\lambda_{\mathrm{rec},2}$	$\lambda_{\mathrm{rec},3}$
HotpotQA	200	0.1	0.1	0.1
MMLU	200	0.1	0.1	0.1
TriviaQA	100	1	1	1
TruthfulQA	50	0.01	0.01	0.01
CV-Bench	100	0.01	0.1	0.01
GQA	200	0.1	0.1	0.1
POPE-Adv	50	0.1	0.1	0.1

k nearest reference states from each outcome class by cosine distance. The scale is

$$\tau = \frac{1}{n^2} \sum_{i,j=1}^{n} (1 - z_i^\top z_j), \qquad \varepsilon = \lambda_{\mathrm{rec}} \tau,$$

where the sum uses the reference bank, including zero diagonal distances. We then compute Eq. 6 with the solver in Appendix E.3.

The neighbor count is fixed per dataset and shared across all three models. The residual multiplier λ_{rec} is set separately for each model using the labeled reference set. Table 4 lists the settings used for independent test evaluation.

F.3 Baseline scores

The kNN baseline retrieves the five closest states from the complete reference set, using the same reference-centered unit vectors and cosine distance as local mixture energy. Its score is

$$s_{\mathrm{kNN}}(z) = \frac{1}{5} \sum_{i \in N_5(z)} Y_i,$$

where $N_5(z)$ is the set of five nearest references and $Y_i = 1$ denotes an incorrect answer. This setting is shared across all models and datasets.

The four sampling baselines share five answers generated with temperature 0.5, top-k sampling 10, and top-p 0.99. Likelihoods are evaluated under the original model distribution. Perplexity is the exponential of the greedy answer's mean negative token log likelihood. LN-Entropy averages the length-normalized negative log likelihood over sampled answers.

Semantic Entropy groups answers using DeBERTa-v2-xlarge-MNLI. A pair shares a group when neither direction predicts contradiction and at least one predicts entailment. We prepend the question and options to each answer and cap paired NLI inputs at 512 tokens. For HotpotQA, the full passage-and-question prompt is prepended, and left truncation preserves the question and answer at the end of each sequence. Cluster masses are formed from normalized, exponentiated mean token log likelihoods. Lexical Similarity uses one minus the mean pairwise ROUGE-L F1 score with stemming. EigenScore uses the five middle-layer representations at the last answer token, computes their response covariance, and averages its base-10 log eigenvalues after adding $0.001I$. We follow the authors' implementations of semantic uncertainty and EigenScore (Kuhn et al., 2023; Chen et al., 2024) with this common sampling budget.

F.4 Supervised pre-generation probe

The probe is trained on the reference questions in Table 3 and evaluated on the corresponding test partition.

We retain the probe configuration reported by Ji et al. (2024). The hidden width is 11008. The input is the final prompt-token state after the checkpoint's final RMSNorm, without feature standardization or unit normalization. Training uses cross-entropy, AdamW, learning rate 10^{-5}, batch size 128,

and 10 epochs, with linear learning-rate decay, no warmup, and zero weight decay. The epoch-10 checkpoint is used without validation-based selection. The difference between the hallucinated and non-hallucinated logits is the AUROC score. Results appear in Table 1.

G Synthetic Experiments

G.1 Shared components in a trained Transformer

The task in Figure 3 has four fields, each with eight possible values. A token identifies a field, its value, and one of two alternative encodings. There are 64 field-value tokens, 16 nuisance tokens, and four query tokens. An input contains one token type for each field and four distinct nuisance-token types. Each of these eight types occurs at least once. Twelve further occurrences are drawn from a uniform multinomial distribution, giving a context of length 20. Token positions are randomly permuted, and a query token is appended. The target is the value of the queried field.

We randomly partition the 8^4 field-value combinations into 3276 training combinations and 820 held-out combinations. Training batches independently sample encodings, nuisance types, multiplicities, and queries. The split and evaluation inputs are fixed across model seeds. The model has two pre-LayerNorm Transformer blocks with width 64, four attention heads, and feed-forward width 128. A final LayerNorm and linear head map the query state to eight classes. All positions attend to the 20 context positions. The model has no positional embeddings or dropout. All parameters are trained with cross-entropy using AdamW, learning rate 3×10^{-4}, weight decay 0.01, batch size 256, and 4000 updates.

For each of 512 held-out evaluation groups, we construct an anchor and two same-answer variants. The shared-component variant preserves the target token and its count, along with all other token identities, and redistributes the remaining counts among the seven non-target types. Each type remains present, and the sequence length is unchanged. The alternative-encoding variant replaces the target token with the other encoding of the same field value while preserving all counts and other identities. Queries are identical within each group.

We measure Euclidean distance between final query states after LayerNorm and before the output head. For each pair type, we average the 512 distances within each model seed. The final means across seeds are 0.750 ± 0.022 for shared target tokens and 2.110 ± 0.078 for alternative encodings, where the spread is the sample standard deviation across seeds.

G.2 Quantitative attention dynamics

Figure 2 uses a one-layer, single-head attention model of width 32. Four context-token embeddings and the query embedding are distinct basis vectors; the query attends only to the context tokens. Full-batch gradient descent updates the 32×32 query matrix. The embeddings, keys, values, identity attention output projection, and prediction head are fixed. The model has no feed-forward block or normalization layer.

The two equally weighted contexts have component counts $(1, 1, 1, 3)$ and $(2, 2, 6, 2)$, and both target the first of three output classes. The prediction head reads the first three value coordinates, with

$$B_{1:3,:} = \begin{pmatrix} 2 & 2 & 0 & -1 \\ 2 & 0 & 0 & 1 \\ 0 & 2 & 0 & 0 \end{pmatrix}.$$

The residual is the query embedding, which lies outside these coordinates. The optimal mixture is $\pi^* = (1/2, 1/2, 0, 0)$, and both components outside its support have positive prediction costs. With $1/\sqrt{32}$ attention scaling and a unit-norm query embedding, the fixed keys are chosen so that $\Xi^\top \Xi / 32 = P$, where

$$P = \begin{pmatrix} 1 & 0.20 & 0.10 & 0.05 \\ 0.20 & 1.30 & 0.05 & 0.15 \\ 0.10 & 0.05 & 0.80 & 0.10 \\ 0.05 & 0.15 & 0.10 & 1.10 \end{pmatrix}.$$

Starting from scores $(0.03, -0.03, -7, -7)$, we use step size $\eta = 0.03475044$ and run 10^8 updates in double precision. Figure 2 compares the scaled residual weights and representation difference with their predicted limits. The distance slope is fitted over the final 10^7 to 10^8 updates. At the final update, the scaled difference vector has relative error 3.8% with its predicted limit. The fitted distance slope is -0.973, compared with the predicted -1.