# THE UNEMBEDDING BRIDGE: FROM HIDDEN-STATE MOVEMENT TO IN-CONTEXT LEARNING PREDICTIONS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Mechanistic accounts of in-context learning (ICL) characterize internal representations and causal components, whereas output-level accounts measure demonstration-induced changes in the predictive distribution. We argue that the unembedding bridges these levels: ICL reflects the interaction of a context-induced hidden-state movement with training-shaped output geometry. Across seven open-weight language models and seven classification tasks, ICL final states generally have higher mean unembedding alignment than paired zero-shot states, against a frequency-organized, spectrally concentrated output geometry. Synthetic-model training, synthetic–LLM comparison, and theory show how cross-entropy attraction and repulsion can create this geometry when the optimizer rescales rows individually. Splitting the movement at the label-contrast space of each task, we find that its label part lowers mean alignment, because the demonstrated labels point against the mean unembedding direction, whereas its orthogonal part raises it and carries the vocabulary-wide shift; remapping the labels to non-outlier symbols removes the label part but leaves the orthogonal shift. Finally, targeted ablation, attribution, probability shifts, and steering identify induction heads as a principal causal source of the movement. Thus training supplies the geometry, demonstrations the movement, and their interaction the behavior.

## 1 INTRODUCTION

A defining capability of **L**arge **L**anguage **M**odels (LLMs) is performing tasks from input demonstrations without parameter updates. Known as **I**n-**c**ontext **L**earning (ICL) (Brown et al., 2020; Radford et al., 2019), it has drawn substantial attention given its contested consistency with supervised learning (de Wynter, 2026). Understanding its mechanisms has thus become a central academic pursuit.

Two complementary approaches dominate this pursuit. *Internal-level accounts* track demonstration-induced hidden-state changes and their carriers: induction heads localize causal components, task and function vectors summarize activation shifts, and geometry traces movement across layers (Elhage et al., 2021; Olsson et al., 2022; Hendel et al., 2023; Todd et al., 2024; Kirsanov et al., 2025). *Output-level accounts* measure how demonstrations change candidate- or ground-truth-token probabilities relative to zero-shot or alternative prompts, summarized by decoded labels, accuracy, and entropy (Lu et al., 2022; Zhao et al., 2021; Hashimoto et al., 2024; Cho et al., 2025b). One level localizes internal computation; the other characterizes its predictive effects. Neither maps state changes to token-probability changes.

We propose the unembedding as this missing map. Let $U = [u_1, \ldots, u_V]^\top$ be the token unembeddings and $x$ the normalized final state at the answer position. For ICL and **Z**ero-**S**hot **L**earning (ZSL) prompts sharing a query,

$$\Delta z = U \Delta x, \qquad \Delta x = x_{\mathrm{ICL}} - x_{\mathrm{ZSL}}. \tag{1}$$

Every prediction-relevant internal change must therefore reach the final state and pass through $U$. Motivated by the vocabulary-wide map $\Delta z \in \mathbb{R}^V$, we compare paired ICL and ZSL states against all token unembeddings in $U$ (Figure 1B). ICL final states generally show higher mean unembedding alignment than paired ZSL states. This is surprising because a common account emphasizes that demonstrations define a small label space and concentrate prediction within it (Min et al., 2022), whereas our signature spans the vocabulary. To explain this global effect, we examine $U$ itself. The

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.

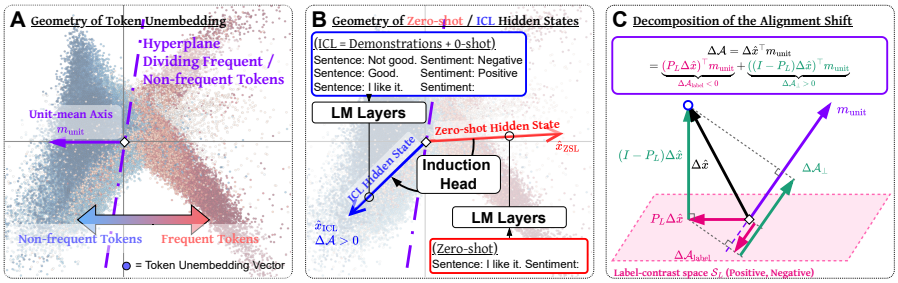



Figure 1: **A.** Pretrained unembedding rows split by token frequency: a hyperplane separates a small set of frequent tokens (red) from a coherent majority of non-frequent tokens (blue); this majority sets the direction of the unit-mean axis $m_{\text{unit}}$, the average of the normalized rows. **B.** Zero-shot final states point toward the frequent side; with demonstrations, the state moves toward the non-frequent side and its mean unembedding alignment rises ($\Delta\mathcal{A} > 0$), with induction heads generating a principal component of this movement. **C.** The ICL-induced movement $\Delta\hat{x}$ splits into its part in the label-contrast space $\mathcal{S}_L$, spanned here by the contrast directions of Positive and Negative, and its orthogonal part. Projected onto $m_{\text{unit}}$, the first points backward and lowers alignment ($\Delta\mathcal{A}_{\text{label}} < 0$), because the labels point against $m_{\text{unit}}$, while the second points forward and raises it ($\Delta\mathcal{A}_\perp > 0$).

unembedding is highly structured rather than isotropic: its singular spectrum is concentrated, and its raw-mean direction induces a sparse frequency-based bipartition (Figure 1A).

Starting from $\Delta z = U\Delta x$, we first show that ICL induces a vocabulary-wide increase in mean unembedding alignment, exposed by a frequency-organized, spectrally concentrated $U$ (Figure 1A,B). We then trace this geometry to cross-entropy attraction and softmax repulsion through theory, synthetic-model training, and comparison with LLM unembeddings. Next, we decompose the movement at the task's label-contrast space: its component within this space lowers mean alignment because the demonstrated labels point against the unit-mean axis, whereas the orthogonal component raises it; label remapping and prompt controls identify the source of each (Figure 1C). Neither component requires a correct input–label mapping, consistent with accounts in which demonstrations act mainly through task recognition rather than task learning (Min et al., 2022; Pan et al., 2023). Finally, attribution, ablation, probability-distribution recovery, and steering identify induction heads as a principal causal source of the movement and yield a training-free steering method competitive with training-based alternatives (Figure 1B). Together, we contribute a unified unembedding bridge, a cross-entropy account of its learned geometry, and a causal account of the ICL-induced state movement it reveals.

## 2 Related Work

**Internal-level accounts of ICL** Internal-level work asks how demonstrations change the LLM representations and components. Induction heads implement an ICL copying pattern—in `[w] [y] ... [w]` $\rightarrow$ `[y]`, retrieving the label after a matching precedent (Olsson et al., 2022). Task vectors summarize distributed activations for steering (Hendel et al., 2023), while geometric analyses track prompt-induced reorganization (Kirsanov et al., 2025). Recent work revisits larger-model circuits, compares induction and function-vector heads, localizes task recognition and task learning to attention heads, and relates heads to task vectors via layerwise geometry (Cho et al., 2025a; Yin & Steinhardt, 2025; Yang et al., 2026; 2025). These accounts localize computation but lack a common readout from final-state movement to vocabulary-wide probability change.

**Output-level accounts of ICL** Output-level work compares predictions with and without demonstrations. Prompt-order and calibration studies find redistributed answer-token probability and changed decoded labels relative to zero-shot or content-free baselines (Lu et al., 2022; Zhao et al., 2021). Incremental analysis tracks candidate- and ground-truth-token changes as demonstrations accumulate (Hashimoto et al., 2024). Related accounts invoke contextual reinforcement, label verbalization, or predictive confidence (Yan et al., 2024; Tao et al., 2024; Li et al., 2025a). These studies track a few tokens that the prompt singles out, not how the final-state change moves the whole vocabulary through the unembedding.

**Unembedding-based links between internal and output accounts** Several interpretability methods use the unembedding to map internal computation into output coordinates. Logit and tuned lenses decode intermediate hidden states as token predictions, while direct logit attribution decomposes final logits into additive component contributions (nostalgebraist, 2020; Belrose et al., 2023; Chughtai et al., 2024). In ICL, task-subspace attribution evaluates head outputs against task-label unembeddings, linking components to label-space recognition and selection (Yang et al., 2026). Studies of token (un)embedding geometry further identify sparse log-linear probability directions and early frequency encoding (Cho et al., 2025c), shared cross-scale structure enabling steering-vector transfer (Lee et al., 2025), and semantic, syntactic, and frequency structure emerging during training, with faster convergence for frequent words (Papadimitriou & Prince, 2026). Controlled pretraining also shows that effective rank varies with training hyperparameters (Kulkarni et al., 2026). These works either use $U$ mainly as a fixed decoder or study its geometric regularities and training correlates. We connect these views by treating $U$ as both a learned geometric object and a functional readout: we derive how cross-entropy attraction and repulsion organize it by frequency, then test how this geometry maps paired ICL–ZSL state movement to vocabulary-wide alignment and task-label probability changes.

## 3 Background

**Transformer readout and attribution** The circuit formulation treats Transformer components as reading from and additively writing to the residual stream. For an $n$-token input, let $h_i^l \in \mathbb{R}^d$ be the state at position $i$ after layer $l$. Unrolling a Transformer with $H$ heads per layer gives the answer-position state

$$h_n^{L_{\text{lay}}} = h_n^0 + \sum_{l=1}^{L_{\text{lay}}} \left( \sum_{k=1}^{H} a_{n,k}^l + m_n^l \right), \tag{2}$$

where $a_{n,k}^l$ and $m_n^l$ are head $(l,k)$ and MLP outputs. The final RMS-normalized state $x = \text{Norm}(h_n^{L_{\text{lay}}})$ yields logits $z = Ux$; write $\widehat{x} = x/\|x\|_2$. Direct logit attribution (DLA) assigns a head $U\,\text{Norm}(a_{n,k}^l)$, its contribution to every vocabulary logit, where Norm applies the final-normalization weight and frozen prompt-specific RMS scale (Olsson et al., 2022; Chughtai et al., 2024; Yu & Ananiadou, 2024; Lieberum et al., 2023). Exact scaling is in Section E.2.

**ICL and induction heads** An ICL prompt prepends $m$ text–label demonstrations $(w_1, y_1), \ldots, (w_m, y_m)$ to query $q$, while its paired ZSL prompt omits them, inducing movement $\Delta x = x_{\text{ICL}} - x_{\text{ZSL}}$ and directional change $\Delta\widehat{x} = \widehat{x}_{\text{ICL}} - \widehat{x}_{\text{ZSL}}$. Induction heads (IHs) use prefix-matching, copy-like attention to route labels following earlier matching contexts toward the answer position (Olsson et al., 2022); we later test whether their outputs form a principal component of $\Delta x$.

## 4 Experiments

**Models, datasets, and ICL setting** We test Llama3.2-1B/3B, Llama3-8B/70B, Llama3.1-8B (Grattafiori et al., 2024), and GLM4-9B/32B (GLM Team, 2024); unless noted, main-text LLM results use Llama3.1-8B. We evaluate SST-2 (Socher et al., 2013), SNLI (Bowman et al., 2015), TREC (Li & Roth, 2002), RTE (Dagan et al., 2006), MR (Pang & Lee, 2005), SUBJ (Pang & Lee, 2004), and MNLI (Williams et al., 2018). Unless specified, we compare 10-shot ICL with a same-query ZSL prompt; other models and full sampling, label, template, split, and sample-count details are in the section-matched appendices and Sections B and C.

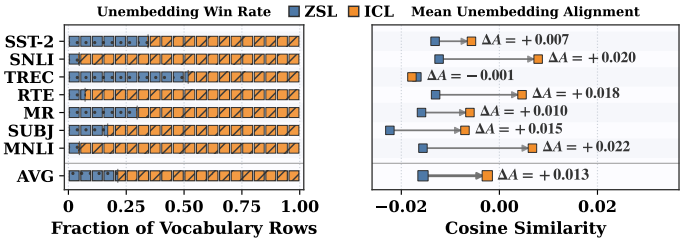



Figure 2: **Left:** Most tasks and the equal-dataset average align a majority of token unembeddings more strongly with ICL than ZSL. **Right:** Positive mean differences show a broad vocabulary-wide shift rather than an effect driven by a few extreme token unembeddings.

### 4.1 THE UNEMBEDDING REVEALS A VOCABULARY-WIDE SIGNATURE OF ICL MOVEMENT

**Vocabulary-wide readout** By Equation (1), ICL changes every logit, $\Delta z_v = u_v^\top \Delta x$, whereas output-level accounts read only a few (Section 2). We average all $V$, normalized to cosines so every token counts equally:

$$\mathcal{A}(x) = \frac{1}{V} \sum_{v \in \mathcal{V}} \cos(x, u_v). \tag{3}$$

We call this *mean unembedding alignment*. Figure 2 shows a broad shift: under equal-dataset averaging, ICL raises Llama3.1-8B mean alignment by 0.013, and 78.6% of token unembeddings align more with ICL than with ZSL (the ICL win rate), so the increase is shared across the vocabulary rather than confined to the few tokens that output-level accounts read. As we increase the demonstrations from one to ten, both the shift and the win rate grow (Section F.2), and the shift holds in most of the six other LLMs (Sections F.1 and G.2).

**A structured global axis** The shift cannot be interpreted without the geometry of its readout. Writing $\widehat{u}_v = u_v / \|u_v\|_2$ and $m_{\text{unit}} = \frac{1}{V} \sum_v \widehat{u}_v$ gives

$$\Delta\mathcal{A} = (\Delta\widehat{x})^\top m_{\text{unit}}. \tag{4}$$

By Equation (4), a shift requires a nonzero mean direction $m_{\text{unit}}$; isotropic unembeddings would make $m_{\text{unit}} \approx 0$ and $\mathcal{A}$ constant. We therefore ask how training creates this mean direction.

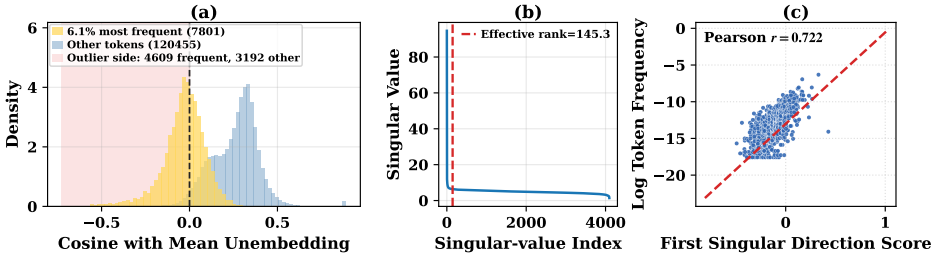



Figure 3: **Left:** Frequent tokens concentrate on the sparse negative side of the mean partition. **Center:** Rapid singular-value decay confines this geometry to a small subspace. **Right:** Log token frequency tracks the leading singular coordinate. Replications and centering controls are in Sections F.4 to F.6.

**Frequency-organized unembedding geometry** To test for isotropy, we locate each row relative to the raw mean $\bar{u} = \frac{1}{V} \sum_i u_i$: under isotropy, about half the rows would fall in $\mathcal{O}_{\text{raw}} = \{i : u_i^\top \bar{u} \le 0\}$. Instead, only 6.1% do (Left), a sharply asymmetric bipartition. Isotropy would also spread the rows over all $d$ dimensions, yet the participation-rank effective dimension $\text{EDim}(U) = (\sum_j \sigma_j^2(U))^2 / \sum_j \sigma_j^4(U)$, normalized by $d$, is only 0.035 (Center). We then ask where tokens of different frequency, estimated from WikiText-103 (Merity et al., 2017) (Section F.3), lie in this geometry. The outlier set overlaps 59.1% with the set $F_k$ of the $k = |\mathcal{O}_{\text{raw}}|$ most frequent tokens, well above chance. Moreover, the projection $u_v^\top \psi_1$ onto the leading right singular vector $\psi_1$, nearly antiparallel to $m_{\text{unit}}$ (Section H.1), tracks log frequency with $r = 0.722$ (Right), an approximately linear frequency encoding. Cho et al. (2025c) likewise find that pretrained token unembeddings sparsely encode token probability and acquire corpus-frequency information early; here, the same frequency structure organizes the ICL–ZSL readout. Centering and normalization controls exclude norm or parameterization artifacts (Section F.6). Figures 2 and 3 thus show ICL movement read out through a frequency-organized geometry, whose origin we examine next.

### 4.2 CROSS-ENTROPY SHAPES FREQUENCY-ORGANIZED UNEMBEDDING GEOMETRY

**A synthetic analogue of language-model training** To investigate this geometry's origin, we train a synthetic Transformer with three ingredients retained from LLM training: long-tailed token frequency, contextual co-occurrence, and next-token cross-entropy. **Language-modeling sequences.** The vocabulary is partitioned into 30 clusters, each a block of one major token followed by its 133 minor tokens. For each topic $\gamma$, we randomly order the blocks, so the leading block's major token ranks first and its minor tokens follow, and set $p_\gamma(v) = r_\gamma(v)^{-2} / \sum_{w \in \mathcal{V}} r_\gamma(w)^{-2}$ from the resulting rank $r_\gamma(v)$. Each of 20,000 sequences samples $\gamma$ uniformly and draws 128 tokens independently from

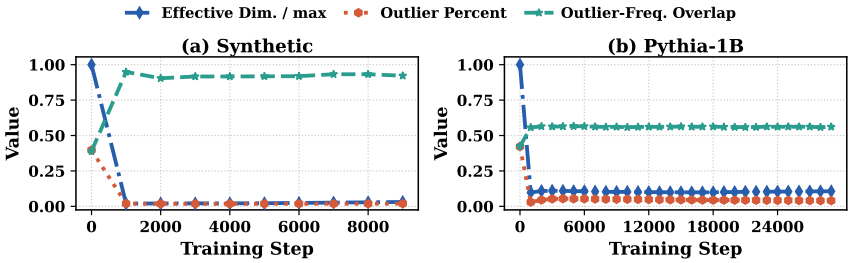



Figure 4: **Left (synthetic model):** Spectral contraction, formation of a sparse mean partition, and its increasing correspondence with token frequency emerge together during training. **Right (Pythia-1B):** The same coordinated transition occurs during language-model pretraining, turning the endpoint similarity into process-level evidence for a shared training origin.

$p_\gamma$. The heavy-tailed marginal approximates natural-language word frequency (Zipf, 1949), while topic-specific ranks create co-occurrence and unequal update frequencies. **ICL sequences.** Following evidence that bursty, parallel pretraining structures facilitate ICL (Chan et al., 2022; Chen et al., 2024), we add 1,000 10-shot ICL sequences. Each draws a cluster pair $(a, b)$ and, independently, two distinct labels $(\ell_a, \ell_b)$ from a pool of 30 label tokens, so labels are independent of cluster identity across prompts. Each demonstration draws a minor token $w_j$ from cluster $a$ or $b$ and emits $w_j$ : $\ell_a$ | or $w_j$ : $\ell_b$ | accordingly; the query repeats a demonstrated token, followed by :, and targets its cluster's label. Training and evaluation use 100 and 20 disjoint cluster pairs, preventing memorization of evaluation pairings.

The six-layer, width-256 causal Transformer has 16 attention heads and MLP intermediate size 512 per layer; width/depth ablations are in Section G.1. It uses next-token cross-entropy, matching the LLM autoregressive objective. We test whether it develops LLM-like frequency geometry and functional ICL; full details are in Section D.

**The synthetic model reproduces the LLM geometry and readout** Table 1 compares the synthetic model with Llama3.1-8B. Both have low normalized effective rank and a small negative raw-mean group that overlaps strongly with an equal-size frequent-token set.[1] Also, their ICL accuracies confirm the ICL function. More importantly, both have ICL win rates above 0.7 and positive $\Delta A$. Despite different vocabularies and widths, both exhibit the same frequency organization and ICL alignment, implicating **shared next-token cross-entropy training**. Other LLMs appear in Section G.2.

| Metric | Synthetic | Llama3.1-8B |
|---|---|---|
| **Unembedding geometry** | | |
| Eff. rank / dim. | 0.032 | 0.035 |
| Outlier rate | 0.021 | 0.061 |
| Outlier–freq. overlap | 0.953 | 0.591 |
| **ICL–ZSL readout** | | |
| ICL accuracy | 0.518 | 0.768 |
| ICL win rate | 0.734 | 0.786 |
| Mean-align. change $\Delta A$ | **+0.054** | **+0.013** |

Table 1: Synthetic and Llama3.1-8B share frequency geometry and paired ICL–ZSL readout; normalized rank permits cross-width comparison.

**Training dynamics support a shared origin** The agreement between the fully trained models suggests a common origin, but a stronger test is whether the geometry develops during next-token training. We therefore track normalized effective dimension, outlier fraction, and outlier–frequency overlap during training of the synthetic model and Pythia-1B. Figure 4 shows effective dimension contracting as an asymmetric mean partition emerges and increasingly matches the frequent set. Parallel synthetic and pretrained trajectories provide process-level evidence for cross-entropy shaping this geometry; further dynamics are in Section G.4.

**Cross-entropy explains the emerging global geometry** Figure 4 shows when, but not why, this geometry forms; we trace it to the cross-entropy gradient. Let $\pi_i = \Pr(y = i)$ be token $i$'s frequency in the training data and $\mu_i = \mathbb{E}[x \mid y = i]$ its conditional mean state. Through the output layer, population cross-entropy gives $u_i$ the gradient

$$-\nabla^{\text{out}}_{u_i} \mathcal{L}_{\text{CE}} = \pi_i \mu_i - \mathbb{E}[p_i(x)x], \tag{5}$$

an attraction toward states where token $i$ is the target and a repulsion proportional to its predicted probability $p_i(x) = \text{softmax}(Ux)_i$. Let $\Phi_t$ denote the optimizer update map—the identity for

[1] In every experiment, each model's frequent set contains its top $\text{len}(\mathcal{O}_{\text{raw}})$ tokens by empirical frequency, matching the outlier-set size.

gradient descent and a row-dependent per-coordinate rescaling for Adam-family optimizers. Then $\Delta u_i = u_i^{(K)} - u_i^{(0)} = \sum_{t<K} \eta_t \, \Phi_t(\pi_i \mu_i^{(t)} - \mathbb{E}[p_i^{(t)}(x) x^{(t)}])$. We focus on the initial training phase, where the geometry quickly forms and stabilizes (Figure 4) and predictions are nearly uniform, so $\mathbb{E}[p_i^{(t)}(x) x^{(t)}] \approx \bar{x}^{(t)}/V$, with $\bar{x}^{(t)} = \mathbb{E}[x^{(t)}]$. Thus, for $K$ within this phase, each row's accumulated update is

$$\Delta u_i \approx \sum_{t<K} \eta_t \, \Phi_t\big(\pi_i \mu_i^{(t)} - \bar{x}^{(t)}/V\big), \tag{6}$$

a token-specific pull scaled by $\pi_i$ against a vector shared across tokens. For rare rows, this pull is smaller, so the shared vector dominates their accumulated update and becomes a common component, around which the rows spread according to their pulls and initialization relative to it (Section A.3); this coherent majority concentrates the spectrum into the low-rank structure of Figure 3.

**Frequency-based separation depends on the optimizer** The partition in Figure 3 is defined relative to the raw mean, $\mathcal{O}_{\text{raw}} = \{i : u_i^\top \bar{u} \le 0\}$, asking whether a row's accumulated update agrees with the average update $\sum_j \Delta u_j / V$. Cross-entropy keeps this average at zero because the row gradients sum to $(\sum_i p_i(x) - 1)x^\top = 0$. Thus, for an untied output matrix, any update map applying the same $D_t$ to every row, as gradient descent and momentum SGD do, leaves $\bar{u}$ at its random initialization $\bar{u}^{(0)}$, inducing only random rather than frequency-based partitions (weight tying changes this through the input-embedding gradient; Section A.8). Adam-family optimizers, used for both Pythia-1B and our synthetic model, instead apply token-specific $D_{i,t}$, so the sum no longer cancels. To track row $u_i$, we project Equation (6) onto $e$, the direction of $\bar{x}^{(t)}$, assumed fixed while the geometry forms, obtaining $\Delta u_i^\top e \approx \pi_i M_i - S_i/V$, where $M_i = \sum_{t<K} \eta_t \, \mu_i^{(t)\top} D_{i,t} e$ is the attraction and $S_i = \sum_{t<K} \eta_t \, \bar{x}^{(t)\top} D_{i,t} e$ the repulsion. Since $\bar{x}^{(t)} = \sum_j \pi_j \mu_j^{(t)}$, assuming comparable conditional means $\mu_j^{(t)} \approx \mu_i^{(t)}$ gives $S_i \approx M_i$ and hence $(\pi_i - 1/V) M_i$: rows with $\pi_i > 1/V$ move toward $e$, while the rest—the vast majority under a Zipf marginal—drift along $-e$. Because their step size is largely frequency-independent, this majority dominates the average, giving $\bar{u} \approx -\kappa e$ with $\kappa > 0$, so $u_i^\top \bar{u} \approx -\kappa \, \Delta u_i^\top e$ is negative exactly for the frequent rows, yielding the small outlier set in Figure 3.

**Theorem 1** (Frequency-separated low-rank geometry under cross-entropy and an adaptive optimizer; informal)**.** *Let $F = \{i : \pi_i M_i > S_i/V\}$ and $N = \mathcal{V} \setminus F$. Suppose **(P1)** the mean row moves along $-e$, $\kappa = -(\bar{u}^{(K)} - \bar{u}^{(0)})^\top e > 0$, and on row $i$ the rest $\bar{u}_{\text{rest}} = \bar{u}^{(K)} + \kappa e$ does not reverse the sign this movement gives $u_i^\top \bar{u}$, $\operatorname{sign}(u_i^\top e)\, u_i^\top \bar{u}_{\text{rest}} < \kappa |u_i^\top e|$, and **(P2)** the rows of a subset $G \subseteq N$ holding a share $\omega_{\text{g}}$ of the squared row norms deviate from their mean $\bar{u}_G$ by $\bar{r}\|\bar{u}_G\|_2$ in root mean square, with $(1+\bar{r}^2)/\omega_{\text{g}} \ll \sqrt{d}$. Then $\mathrm{EDim}(U)/d \le (1+\bar{r}^2)^2/(d\omega_{\text{g}}^2) \ll 1$, and $\mathcal{O}_{\text{raw}}$ coincides, to leading order, with $F$ on every row that satisfies (P1) and $|\Delta u_i^\top e| > |u_i^{(0)\top} e|$.*

The theorem formalizes how the attraction–repulsion imbalance in Equations (5) and (6) produces the two global signatures observed above. Low target frequency weakens row-specific attraction, so the shared repulsion gives the non-frequent majority a common component around which their rows cluster (P2); the tighter they cluster, the more the full singular spectrum concentrates, with effective dimension capped at $(1+\bar{r}^2)^2/\omega_{\text{g}}^2$ whatever the width $d$. Once the optimizer also moves the mean row along $-e$, each row lies on the side of the learned mean boundary set by its own update along $e$, unless the rest $\bar{u}_{\text{rest}}$ reverses that side, which (P1) excludes, or its initialization outweighs the update. Rows whose attraction outweighs the repulsion, $\pi_i M_i > S_i/V$, therefore fall across the boundary while the rest stay on its shared side, producing a small outlier set with high overlap with frequent tokens (Figure 1A). The result thus connects CE update moments to normalized effective dimension, outlier rate, and outlier–frequency overlap. Conditions and proofs, including the derivation of (P2) from the early-phase drift, are in Section A. Section G.6 measures both premises across optimizers and finds the frequency-separated low-rank geometry only when both hold. In untied runs, momentum SGD violates (P1), keeping the mean row at initialization and the partition near chance, whereas Adam satisfies both and its outliers overlap the frequent set by 0.905–0.933.

**Data modulate CE geometry** Theorem 1 places conditions on the updates, not on the data, so we vary how the training data are generated. In our synthetic setting, topic-plus-Zipf sampling of the language-modeling sequences combines unequal $\pi_i$, separating frequent and non-frequent attraction regimes, with topic-conditioned co-occurrence: $\mu_i^{(t)} = \sum_\gamma \Pr(\gamma \mid y = i)\, \mathbb{E}[x^{(t)} \mid \gamma, y = i] = \sum_\gamma \Pr(\gamma \mid y = i)\, \mathbb{E}[x^{(t)} \mid \gamma]$. Conditional independence of $x^{(t)}$ and $y$ given $\gamma$ gives the second

equality. Every $\mu_i^{(t)}$ thus lies in the span of 30 topic-state means, which further bounds the rank of the token-specific updates. **Removing topic structure only** samples every token independently from the same marginal: each $\pi_i$ is unchanged, but $y$ no longer depends on $x^{(t)}$, so $\mu_i^{(t)} = \bar{x}^{(t)}$ and the drift in Equation (6) becomes $(\pi_i - 1/V)\,\bar{x}^{(t)}$, moving every row along one direction with a sign set by $\pi_i$ against $1/V$. With the conditional means now exactly comparable, **the spectrum should stay concentrated and the outliers should match** $\{i : \pi_i > 1/V\}$, up to tokens with $\pi_i \approx 1/V$ and ICL-only tokens. **Keeping the topics but removing the Zipf skew within them** equalizes $\pi_i$, turning the drift into $(\mu_i^{(t)} - \bar{x}^{(t)})/V$, a topic deviation that ignores frequency and averages to zero, so the **frequency-based mean partition should disappear** and the spectrum spread over the topic directions. **Removing both the topics and the Zipf skew** gives $\pi_i = \frac{1}{V}$ and $\mu_i^{(t)} = \bar{x}^{(t)}$, so the drift in Equation (6) vanishes and **both the separation due to Zipf skew and the low rank sustained by topics should disappear**.

**Controlled interventions test these implications** We test whether the predictions of the theory for the three controls hold, with the architecture and optimizer fixed. To check the claim that removing topic structure makes the outliers the tokens with $\pi_i > 1/V$ (4.9% of the vocabulary under the Zipf marginal), Table 2 also reports the threshold match, the Jaccard index between the outlier set and that set; in the two controls without Zipf skew every $\pi_i$ is close to $1/V$, so no threshold set is predicted and the match is not reported. Removing topic structure only keeps the spectrum as concentrated as with topics (0.033 against 0.032) and moves the outliers onto the threshold set, raising the threshold match from 0.43 to 0.93. Removing Zipf skew only destroys the sparse frequency-based partition, whose overlap falls to its chance level, the outlier rate, and costs part of the low rank, raising effective dimension by 0.231; removing the topics as well costs far more, raising it by 0.802. All variants preserve functional ICL accuracy (0.488–0.546), and untied controls exclude weight tying as the sole source (Section G.5). Cross-entropy under an Adam-family optimizer thus leaves a coherent non-frequent majority and a frequency-organized mean partition.

| Zipf | Topic | Eff. rank / dim. ↓ | Outlier rate | Outlier–freq. overlap ↑ | Threshold match ↑ | ICL acc. ↑ |
|---|---|---|---|---|---|---|
| ✓ | ✓ | **0.032** | **0.021** | **0.953** | **0.43** | **0.518** |
| ✓ | × | 0.033 ↑0.001 | 0.049 ↑0.028 | 0.970 ↑0.017 | 0.93 ↑0.50 | 0.494 ↓0.024 |
| × | ✓ | 0.263 ↑0.231 | 0.192 ↑0.171 | 0.191 ↓0.762 | – | 0.546 ↑0.028 |
| × | × | 0.834 ↑0.802 | 0.203 ↑0.182 | 0.197 ↓0.756 | – | 0.488 ↓0.030 |

Table 2: Removing topic structure preserves spectral concentration and places outliers at the threshold $\{\pi_i > 1/V\}$; removing Zipf skew destroys the frequency partition, while removing both also destroys spectral concentration.

### 4.3 Decomposing the mean-alignment shift relative to the task-label subspace

**From learned geometry to mean alignment** We next ask how CE-shaped geometry produces the shift in Figure 2, separating formation of a global axis from prompt-time movement along it.

**Corollary 1** (A coherent non-frequent majority defines a global alignment axis)**.** *Let $G \subseteq N$ be the set of* (P2) *in Theorem 1, $E = N \setminus G$ the remaining non-frequent rows, and $v_i$ the normalized rows. If $\|\sum_{i\in G} v_i\|_2 > \|\sum_{i\in E} v_i\|_2 + \|\sum_{i\in F} v_i\|_2$, then $m_{\text{unit}} = \frac{1}{V}\sum_i v_i \neq 0$.*

For paired ICL and ZSL states, $\Delta\mathcal{A} = (\Delta\widehat{x})^\top m_{\text{unit}}$, so Theorem 1 and Corollary 1 connect CE moments to the global ICL signature: when the coherent non-frequent majority outweighs aggregate cancellation, $m_{\text{unit}}$ is a nonzero axis along which prompt-time movement determines $\Delta\mathcal{A}$ (proof and condition audit in Sections A.7 and G.3; relation to ICL in Section G.6). This axis, which measures movement against the whole vocabulary, is also nearly antiparallel to the leading singular vector $\psi_1$ encoding log frequency in Figure 3 (right; oriented toward frequent tokens), with cosine $-0.997$ in Llama3.1-8B and at most $-0.989$ across LLMs (Section H.1). A positive $\Delta\mathcal{A}$ thus moves the final state away from the frequent-token side, but whether it stems from the ICL state aligning with the demonstration label unembeddings or from other factors remains open.

**Task-label decomposition** To study this, we zoom in on the demonstration label unembeddings. Let $L \subset \mathcal{V}$ be the demonstration label set, $\bar{L} = \mathcal{V} \setminus L$, and $\bar{u}_S$ the mean unembedding over $S$; we consider the span of unembedding contrast directions $\mathcal{S}_L = \text{span}\{u_\ell - \bar{u}_{\bar{L}} : \ell \in L\}$, with orthogonal projector $P_L$. Projecting the hidden-state movement onto $\mathcal{S}_L$ captures its component along all label-related unembedding directions. As $(u_\ell - \bar{u}_{\bar{L}})^\top x$ is $\ell$'s logit minus the mean background logit, movement outside $\mathcal{S}_L$ leaves every label's standing intact. We thus decompose the signature $\Delta\mathcal{A} = (\Delta\widehat{x})^\top m_{\text{unit}}$ as

$$\Delta\mathcal{A} = (P_L\Delta\widehat{x})^\top m_{\text{unit}} + ((I - P_L)\Delta\widehat{x})^\top m_{\text{unit}} = \Delta\mathcal{A}_{\text{label}} + \Delta\mathcal{A}_\perp, \tag{7}$$

where $\Delta\mathcal{A}_{\text{label}}$ is carried by movement within the label contrasts and $\Delta\mathcal{A}_{\perp}$ by movement outside them, separating alignment gained by moving toward the demonstration labels from alignment gained without changing their standing.

**The orthogonal part carries the shift** Figure 5 (left) shows the result of the decomposition in Equation (7): in all seven tasks the orthogonal part is positive and the label part negative, so the positive ICL shift comes from movement outside the label subspace, which the label part partly offsets and in TREC nearly cancels. To explain the negative label part, we compute for each demonstration label its *label depth* $\cos(u_\ell - \bar{u}_{\bar{L}}, m_{\text{unit}})$ and its frequency rank. Every label points against $m_{\text{unit}}$, and 18 of the 20 lie in the outlier set $\mathcal{O}_{\text{raw}}$ on the frequent side of the unembedding (Figure 3), 15 of them also in the size-matched frequent set $F_k$ (Figure 5, right; Sections H.3, H.5 and H.6 extend this to all seven LLMs). ICL movement toward these labels therefore also moves the state against $m_{\text{unit}}$, which lowers its alignment and makes the label part negative.

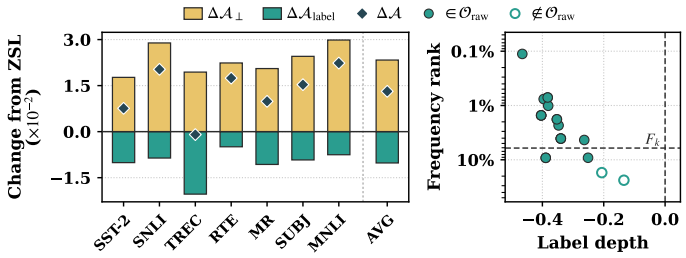



Figure 5: Movement outside the label subspace carries the ICL alignment shift. **Left:** In every task the orthogonal part is positive and the label part negative; diamonds mark their sum $\Delta\mathcal{A}$. **Right:** Every demonstration label points against $m_{\text{unit}}$ and ranks among the most frequent tokens; filled markers lie in $\mathcal{O}_{\text{raw}}$, and the dashed line bounds the size-matched frequent set $F_k$.

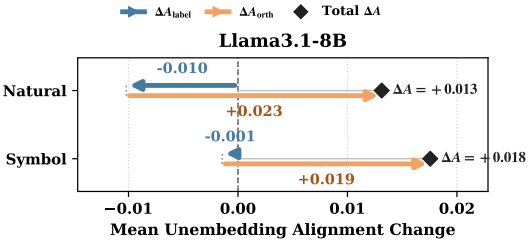



Figure 6: Under both label types, movement orthogonal to the task-label subspace supplies the positive alignment shift; the label-subspace contribution is small and negative.

**Non-outlier labels remove the label part but not the orthogonal shift** We next test what happens when the labels are not outliers, remapping each task's labels to one-token symbols outside $\mathcal{O}_{\text{raw}}$ while keeping the queries, demonstrations, and their order fixed; for SST-2, negative/positive become ∀/←, and Section C lists every mapping. With these labels, whose contrasts are nearly orthogonal to $m_{\text{unit}}$, the negative label part nearly vanishes while the positive orthogonal part persists (Figure 6). The label part thus follows where the labels sit, whereas the orthogonal shift is a label-robust signature of ICL. Prompt controls (Section H.7) show that the orthogonal part needs demonstration records but not correct input–label pairs: in Llama3.1-8B, listing the labels, label-only records, and demonstrations with random labels reproduce 0.04, 0.48, and 1.07 of it, whereas the label part grows with how often the labels appear. Synthetic and other-LLM replications are in Sections H.2 and H.4.

### 4.4 Induction heads generate a principal component of ICL-induced hidden-state movement

Section 4.3 traces the alignment shift to movement outside the task-label subspace. **What components drive this movement?** Subtracting paired ICL–ZSL hidden states given in Equation (2) gives $\Delta h_n^{L_{\text{lay}}} = \Delta h_n^0 + \sum_{l,k} \Delta a_{n,k}^l + \sum_l \Delta m_n^l$ and thus $\Delta x = \text{Norm}(h_{n,\text{ICL}}^{L_{\text{lay}}}) - \text{Norm}(h_{n,\text{ZSL}}^{L_{\text{lay}}})$, enabling component-level analysis. IHs route information from earlier matching contexts and label positions to the query, making them natural

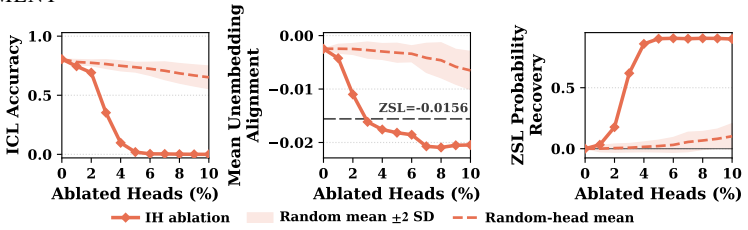



Figure 7: **Left:** IH ablation reduces accuracy more than matched random ablation. **Center:** Mean alignment returns toward and beyond ZSL. **Right:** Predictions recover toward ZSL. Dose effects exceed the random mean and ±2 SD, identifying IH outputs as a concentrated causal source. Other-model replications are in Section I.2.

candidates (Olsson et al., 2022; Cho et al., 2025a). We test the output of a selected IH set $\mathcal{I}$ in both directions, removing it from ICL and adding it to ZSL; identification is detailed in Section E.

**IH ablation reverses probability and alignment shifts** To test IH necessity, we remove progressively larger top-ranked IH sets and recompute the hidden-state movement. We evaluate the distribution over demonstrated labels and the remaining vocabulary using $p_L(x) = (p_\ell(x))_{\ell \in L}$, $p_{\bar{L}}(x) = \sum_{v \notin L} p_v(x)$, and $P_x = (p_L(x), p_{\bar{L}}(x))$. For ICL, ZSL, and ablated values, with $D_{\mathrm{TV}}$ the total variation distance, $R_{\mathrm{ZSL}} = 1 - D_{\mathrm{TV}}(P_{\mathrm{abl}}, P_{\mathrm{ZSL}})/D_{\mathrm{TV}}(P_{\mathrm{ICL}}, P_{\mathrm{ZSL}})$ rises as prediction approaches ZSL.

Across the 1–10% doses in Figure 7, accuracy and mean alignment fall toward and beyond ZSL, whereas ten equal-size random-head controls remain near unablated behavior. IH ablation therefore reverses both the prediction change and the global alignment shift rather than merely reducing accuracy, identifying IH outputs as a causal source of the paired movement.

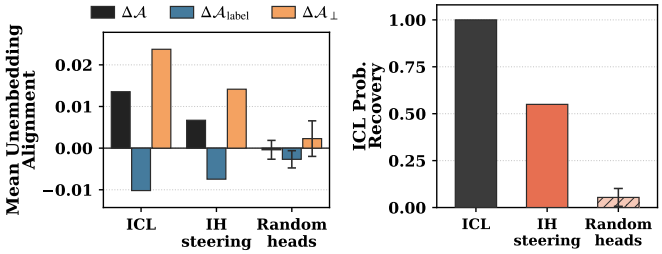



Figure 8: **Left:** IH steering recovers about half of the ICL alignment shift with the same composition, a positive orthogonal and a negative label part; matched random heads stay near zero. **Right:** IH steering moves the output distribution about halfway toward ICL; random heads barely move it.

**IH steering reconstructs the alignment shift and the ICL prediction** Ablation establishes necessity; the converse intervention tests whether IH output alone reproduces the ICL movement. We estimate the layerwise mean ICL–ZSL IH output $v^l_{\mathcal{I}} = \mathbb{E}[\sum_{k:(l,k)\in\mathcal{I}}(a^{l,\mathrm{ICL}}_{n,k} - a^{l,\mathrm{ZSL}}_{n,k})]$ on one split, add it to the final-position attention output of every layer $l$ on held-out ZSL prompts, and let later layers recompute. Steering raises mean alignment by about half of the paired ICL shift and reproduces its composition in Figure 6, a positive orthogonal part with a smaller negative label part, whereas matched random heads leave all three quantities near zero (Figure 8, left). For the prediction, $R_{\mathrm{ICL}} = 1 - D_{\mathrm{TV}}(P_{\mathrm{steer}}, P_{\mathrm{ICL}})/D_{\mathrm{TV}}(P_{\mathrm{ZSL}}, P_{\mathrm{ICL}})$ runs from 0 at ZSL to 1 at ICL; IH steering reaches 0.55 versus 0.05 for random heads (right), so IH output alone recovers about half of the ICL effect in both the hidden state and the output distribution. See Section I.5 for other LLMs and Section I.4 for direct IH attribution.

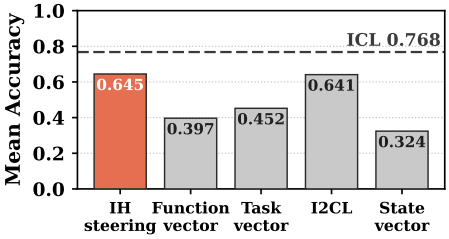



Figure 9: IH steering is the most accurate training-free steering method and matches trained I2CL in performance.

**IH steering is a competitive training-free method** The same layerwise vectors, estimated once and added without demonstrations, also serve as a practical steering method. Averaged over the seven tasks (Figure 9), IH steering reaches 0.645 accuracy against 0.768 for ICL, exceeds training-free function vectors (Todd et al., 2024), task vectors (Hendel et al., 2023), and state vectors (Li et al., 2024) by at least 0.19, and matches PEFT-based I2CL (Li et al., 2025b) (0.641), which trains its vectors. See Section I.5 for other models.

## 5 Conclusion

We establish the unembedding as the learned coordinate system linking internal- and output-level accounts of ICL. Across LLMs, ICL–ZSL movement increases vocabulary-wide alignment within frequency-organized, spectrally concentrated unembedding geometry; theory and synthetic Transformers connect this structure to asymmetric cross-entropy attraction and softmax repulsion. Splitting the movement at the label-contrast space shows that the vocabulary-wide shift lies outside the label subspace: the label part lowers alignment and the orthogonal part raises it, and label remapping and prompt controls trace the two parts to the label tokens and to the demonstration records. Attribution, ablation, and steering identify induction heads as a principal source of the movement.

## AI USE STATEMENT

Generative AI tools assisted with refining hypotheses, the conceptual framework, research methodology, and experiments; formulating and editing mathematical claims and proofs; implementing and testing code; producing visualizations; interpreting results; reviewing relevant literature; and drafting and revising the manuscript. The authors reviewed all AI-assisted work, checked metric definitions and experimental outputs against executable code and source data, inspected mathematical claims and proofs, and verified cited literature. No evaluated model was used to assign ground-truth labels or judge the primary classification results. We take responsibility for all final content and artifacts.

## REFERENCES

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, 2023. URL https://arxiv.org/abs/2303.08112.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 2397–2430. PMLR, 2023. URL https://proceedings.mlr.press/v202/biderman23a.html.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 632–642, Lisbon, Portugal, 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL https://aclanthology.org/D15-1075/.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pp. 18878–18891, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/77c6ccacfd9962e2307fc64680fc5ace-Abstract-Conference.html.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. Parallel structures in pre-training data yield in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8582–8592, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.465. URL https://aclanthology.org/2024.acl-long.465/.

Hakaze Cho, Mariko Kato, Yoshihiro Sakai, and Naoya Inoue. Revisiting in-context learning inference circuit in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=xizpnYNvQq.

Hakaze Cho, Yoshihiro Sakai, Mariko Kato, Kenshiro Tanaka, Akira Ishii, and Naoya Inoue. Token-based decision criteria are suboptimal in in-context learning. 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. 5378–5401, Albuquerque, New Mexico, April 2025b. Association for Computational Linguistics. ISBN 979-8-89176-189-6. URL https://aclanthology.org/2025.naacl-long.278/.

Hakaze Cho, Yoshihiro Sakai, Kenshiro Tanaka, Mariko Kato, and Naoya Inoue. Understanding token probability encoding in output embeddings. In *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 10618–10633. Association for Computational Linguistics, 2025c. URL https://aclanthology.org/2025.coling-main.708/.

Bilal Chughtai, Alan Cooney, and Neel Nanda. Summing up the facts: Additive mechanisms behind factual recall in LLMs. *arXiv preprint arXiv:2402.07321*, 2024.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL recognising textual entailment challenge. In *Machine Learning Challenges: Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pp. 177–190. Springer, 2006. doi: 10.1007/11736790_9. URL https://link.springer.com/chapter/10.1007/11736790_9.

Adrian de Wynter. Is In-Context Learning learning? In *International Conference on Learning Representations*, volume 2026, pp. 101317–101355, 2026.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

GLM Team. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Kazuma Hashimoto, Karthik Raman, and Michael Bendersky. Take one step at a time to know incremental utility of demonstration: An analysis on reranking for few-shot in-context learning. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3973–3990. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.221. URL https://aclanthology.org/2024.naacl-long.221/.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Artem Kirsanov, Chi-Ning Chou, Kyunghyun Cho, and SueYeon Chung. The geometry of prompting: Unveiling distinct mechanisms of task adaptation in language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1855–1888, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-195-7. URL https://aclanthology.org/2025.findings-naacl.100/.

Atharva Kulkarni, Jacob Mitchell Springer, Arjun Subramonian, and Swabha Swayamdipta. Disentangling geometry, performance, and training in language models. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.20433.

Andrew Lee, Melanie Weber, Fernanda Viégas, and Martin Wattenberg. Shared global and local geometry of language model embeddings. In *Second Conference on Language Modeling*, 2025. URL https://openreview.net/forum?id=aJDykpJAYF.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. Large language models are miscalibrated in-context learners. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11575–11596. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.findings-acl.603. URL https://aclanthology.org/2025.findings-acl.603/.

Dongfang Li, Zhenyu Liu, Xinshuo Hu, Zetian Sun, Baotian Hu, and Min Zhang. In-context learning state vector with inner and momentum optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 7797–7820, 2024. doi: 10.52202/079017-0250.

Xin Li and Dan Roth. Learning question classifiers. In *COLING 2002: The 19th International Conference on Computational Linguistics*, 2002. URL https://aclanthology.org/C02-1150/.

Zhuowei Li, Zihao Xu, Ligong Han, Yunhe Gao, Song Wen, Di Liu, Hao Wang, and Dimitris Metaxas. Implicit in-context learning. In *International Conference on Learning Representations*, volume 2025, pp. 101644–101666, 2025b.

Tom Lieberum, Matthew Rahtz, János Kramár, Neel Nanda, Geoffrey Irving, Rohin Shah, and Vladimir Mikulik. Does circuit analysis interpretability scale? evidence from multiple choice capabilities in Chinchilla, 2023. URL https://arxiv.org/abs/2307.09458.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=Byj72udxe.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

nostalgebraist. Interpreting GPT: the logit lens, 2020. URL https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.527. URL https://aclanthology.org/2023.findings-acl.527/.

Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pp. 271–278, Barcelona, Spain, 2004. doi: 10.3115/1218955.1218990. URL https://aclanthology.org/P04-1035/.

Bo Pang and Lillian Lee. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pp. 115–124, Ann Arbor, Michigan, 2005. Association for Computational Linguistics. doi: 10.3115/1219840.1219855. URL https://aclanthology.org/P05-1015/.

Isabel Papadimitriou and Jacob Prince. Vocabulary embeddings organize linguistic structure early in language model training. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2510.07613.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In David Yarowsky, Timothy Baldwin, Anna Korhonen, Karen Livescu, and Steven Bethard (eds.), *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1170/.

Alessandro Stolfo, Ben Wu, Wes Gurnee, Yonatan Belinkov, Xingyi Song, Mrinmaya Sachan, and Neel Nanda. Confidence regulation neurons in language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 125019–125049. Curran Associates, Inc., 2024. doi: 10.52202/079017-3970.

Junyi Tao, Xiaoyin Chen, and Nelson F. Liu. Inference and verbalization functions during In-Context Learning. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16394–16421. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.957. URL https://aclanthology.org/2024.findings-emnlp.957/.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 17282–17333, 2024.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1112–1122. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-1101. URL https://aclanthology.org/N18-1101/.

Jianhao Yan, Jin Xu, Chiyu Song, Chenming Wu, Yafu Li, and Yue Zhang. Understanding In-Context Learning from repetitions. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=bGGYcvw8mp.

Haolin Yang, Hakaze Cho, Yiqiao Zhong, and Naoya Inoue. Unifying attention heads and task vectors via hidden state geometry in in-context learning. In *Advances in Neural Information Processing Systems*, volume 38, pp. 21701–21759. Curran Associates, Inc., 2025. doi: 10.52202/085713-0732.

Haolin Yang, Hakaze Cho, and Naoya Inoue. Localizing task recognition and task learning in in-context learning via attention head analysis. In *International Conference on Learning Representations*, volume 2026, pp. 93780–93839, 2026.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72428–72461. PMLR, 2025. URL https://proceedings.mlr.press/v267/yin25e.html.

Zeping Yu and Sophia Ananiadou. How do large language models learn in-context? query and key matrices of in-context heads are two towers for metric learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 3281–3292, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.192. URL https://aclanthology.org/2024.emnlp-main.192/.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

George Kingsley Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Addison-Wesley Press, Cambridge, MA, 1949.

# Appendices

## A Theory: full statements and proofs

This appendix provides the full statements and proofs for the theory in Sections 4.2 and 4.3: it derives how cross-entropy updates, together with the optimizer, produce the frequency-organized unembedding geometry and its global alignment axis. Section A.1 writes the accumulated cross-entropy update of an output row under the optimizers used in practice. Section A.2 shows that the mean output row is conserved when every row shares its rescaling. Section A.3 derives premise (P2) and the frequency threshold from the early-phase drift, computes, to leading order, the movement of the mean row along $-e$ that premise (P1) assumes, turns (P1) into the frequency partition, and predicts the direction of the unit-mean axis and the frequency ordering along it. Section A.4 discusses which optimizers realize the two premises, Section A.5 states and proves the formal version of Theorem 1, Section A.6 extends it to group-level geometry, and Section A.7 proves Corollary 1. The remaining subsections treat weight tying and directional retention.

Let $\mathcal{D}_{\text{train}}$ denote the token-level next-token distribution whose output-side updates are analyzed, and define token frequency by its target marginal

$$\pi_i = \Pr_{(x,y)\sim\mathcal{D}_{\text{train}}}(y = i).$$

The frequency that scales target attraction below and the frequency that defines the frequent set are therefore the same quantity. In the synthetic experiments, empirical counts from the language-modeling training corpus estimate this marginal directly. For pretrained LLMs the pretraining distribution is unavailable, so the WikiText-103 counts in Section F.3 serve as an empirical proxy for its ranking; the statements below concern the true training marginal. The theoretical frequent set is the threshold set $F$ of Theorem 1, which for comparable conditional means is $\{i : \pi_i > 1/V\}$, a top set of the marginal; the experiments substitute the top-$|\mathcal{O}_{\text{raw}}|$ tokens by empirical frequency so that the two compared sets have equal size. For dynamical statements, $U^{(0)}, \ldots, U^{(K)}$ denotes the sequence of output matrices, an unsuperscripted row is the final row $u_i = u_i^{(K)}$, and $\Delta u_i = u_i^{(K)} - u_i^{(0)}$. For every nonzero row, $v_i = u_i/\|u_i\|_2$ denotes its normalized direction.

### A.1 Gradient identity and accumulated updates

A token unembedding is pulled toward hidden states when its token is the target, and pushed by the softmax term on every example. Target frequency controls how often the first force acts; the second is shared across the vocabulary and can therefore dominate rows with little target-specific evidence.

Let $y \in \mathcal{V}$ be the next-token target and let $x \in \mathbb{R}^d$ be the final state supplied to the output matrix. Define

$$z_i = u_i^\top x, \qquad p_i(x) = \frac{e^{z_i}}{\sum_j e^{z_j}}, \qquad \mu_i = \mathbb{E}[x \mid y = i], \qquad \bar{x} = \mathbb{E}[x] = \sum_{j\in\mathcal{V}} \pi_j \mu_j.$$

When $\pi_i = 0$, the product $\pi_i\mu_i$ is defined to be zero.

**Lemma 1** (Cross-entropy attraction–repulsion identity)**.** *For one example with target distribution $q \in \Delta^{V-1}$ (one-hot or smoothed) and loss $\ell_{\text{CE}} = -\sum_i q_i \log p_i(x)$, the gradient reaching row $u_i$ through the output logits is $\nabla_{u_i}\ell_{\text{CE}} = (p_i(x) - q_i)\,x$. For population cross-entropy $\mathcal{L}_{\text{CE}} = \mathbb{E}[-\log p_y(x)]$ this gives*

$$-\nabla_{u_i}^{\text{out}}\mathcal{L}_{\text{CE}} = \pi_i\mu_i - \mathbb{E}[p_i(x)x].$$

*Proof.* Since $\sum_i q_i = 1$, $\ell = -\sum_i q_i z_i + \log\sum_j e^{z_j}$, so $\partial\ell_{\text{CE}}/\partial z_i = p_i(x) - q_i$, and $\partial z_i/\partial u_i = x$. For one-hot $q$, taking expectations and using $\mathbb{E}[\mathbb{1}\{y = i\}x] = \pi_i\mu_i$ gives the population identity. The negative gradient has inner product $(1 - p_i)\|x\|_2^2 > 0$ with $x$ on target examples and $-p_i\|x\|_2^2 < 0$ otherwise, which justifies "attraction" and "repulsion." □

The identity does not yet imply a geometric partition: both terms depend on the evolving conditional states and predictions, and the optimizer decides how the gradients are turned into parameter changes. We therefore first make the update map explicit.

**Definition 1** (Row-diagonal optimizers)**.** An optimizer acting on the output matrix is *row-diagonal* if, apart from decoupled weight decay, the update of row $i$ at step $\tau$ is

$$u_i^{(\tau+1)} - u_i^{(\tau)} = \eta_\tau \, P_{i,\tau} \sum_{t \le \tau} w_{\tau,t} \, g_i^{(t)}, \qquad g_i^{(t)} = -\nabla_{u_i} \mathcal{L}^{(t)},$$

where each $P_{i,\tau}$ is a diagonal matrix with positive diagonal entries, and the weights $w_{\tau,t} \ge 0$ do not depend on the row and satisfy $w_{\tau,\tau} > 0$. It is *row-shared* if $P_{i,\tau} = P_\tau$ for every row.

Gradient descent ($P_{i,\tau} = I$, $w_{\tau,t} = \mathbb{1}\{t = \tau\}$), heavy-ball and Nesterov momentum ($w_{\tau,t} = \beta^{\tau-t}$), and global gradient-norm clipping ($P_{i,\tau} = c_\tau I$) are row-shared. Adam and AdamW are row-diagonal with $P_{i,\tau} = \operatorname{diag}(\sqrt{\hat{v}_{i,\tau}} + \epsilon_{\mathrm{A}})^{-1}$, where $\epsilon_{\mathrm{A}}$ is Adam's stabilizing constant, and the bias-corrected first-moment weights $w_{\tau,t} = (1-\beta_1)\beta_1^{\tau-t}/(1-\beta_1^{\tau+1})$; RMSProp is the case $\beta_1 = 0$. signSGD is row-diagonal with $P_{i,\tau} = \operatorname{diag}|g_i^{(\tau)}|^{-1}$ on the coordinates where the gradient is nonzero and any positive value elsewhere, where the update is zero regardless. These three are not row-shared. Nesterov momentum evaluates the gradient of the whole output matrix at a lookahead point; this is immaterial below, because Lemma 1 and the zero-sum identity hold at every parameter value, so $g_i^{(t)}$ may be read as the gradient at whichever point the optimizer evaluates it.

**Lemma 2** (Accumulated update)**.** *For a row-diagonal optimizer without weight decay,*

$$\Delta u_i = \sum_{t<K} \eta_t \, D_{i,t}\big(\pi_i \mu_i^{(t)} - \mathbb{E}[p_i^{(t)}(x) x^{(t)}]\big), \qquad D_{i,t} = \frac{1}{\eta_t} \sum_{\tau=t}^{K-1} \eta_\tau \, w_{\tau,t} \, P_{i,\tau}, \tag{8}$$

*where each $D_{i,t}$ is diagonal with positive diagonal entries, and $D_{i,t} = D_t$ for every row when the optimizer is row-shared. Decoupled weight decay at rate $\lambda$ adds $-\sum_{\tau<K} \eta_\tau \lambda \, u_i^{(\tau)}$.*

*Proof.* Summing the updates over $\tau < K$ and exchanging the order of summation gives $\Delta u_i = \sum_{t<K} \big(\sum_{\tau=t}^{K-1} \eta_\tau w_{\tau,t} P_{i,\tau}\big) g_i^{(t)}$, and Lemma 1 supplies $g_i^{(t)}$. Each $D_{i,t}$ is a nonnegative combination of positive diagonal matrices containing the term $w_{t,t}P_{i,t}$ with $w_{t,t} > 0$, so it is diagonal with positive entries; if every $P_{i,\tau}$ equals $P_\tau$, so does every $D_{i,t}$. □

Equation (8) is the accumulated update of the main text, with $\Phi_t$ acting on row $i$ as $D_{i,t}$. It is an identity along the realized trajectory rather than a linearization: for Adam, $D_{i,t}$ depends on the gradients through the second-moment estimate. We analyze population gradients; with minibatches the same identity holds with the realized minibatch gradients.

### A.2 The mean row is conserved under row-shared updates

Lemma 2 describes how each row moves. The raw-mean partition, however, compares every row with the average row $\bar{u} = \frac{1}{V}\sum_i u_i$, so we first ask how that average moves. Under cross-entropy alone it does not.

**Proposition 1** (Zero-sum row gradients and mean-row conservation)**.** *For every output matrix $U$, state $x$ and target distribution $q$,*

$$\sum_{i \in \mathcal{V}} \nabla_{u_i} \ell_{\mathrm{CE}} = \Big(\sum_{i \in \mathcal{V}} \big(p_i(x) - q_i\big)\Big) x = 0.$$

*The identity is preserved by averaging over examples, batches and sequence positions. Consequently, if the output matrix is untied and trained by a row-shared optimizer without weight decay, then $\bar{u}^{(K)} = \bar{u}^{(0)}$ for every $K$. With decoupled weight decay at rate $\lambda$, $\bar{u}^{(K)} = \prod_{t<K}(1-\eta_t\lambda)\,\bar{u}^{(0)}$: the mean row is rescaled but never rotated.*

*Proof.* The identity follows from Lemma 1, $\sum_i p_i(x) = 1$ and $\sum_i q_i = 1$, and linearity extends it to any average. Hence $\sum_i g_i^{(t)} = 0$ at every step, and by Lemma 2 with $D_{i,t} = D_t$, $V(\bar{u}^{(K)} - \bar{u}^{(0)}) = \sum_t \eta_t D_t \sum_i g_i^{(t)} = 0$. Decoupled weight decay multiplies every row, and hence the mean row, by $1 - \eta_\tau\lambda$ at step $\tau$, while the remaining update still sums to zero. □

**Remark 1** (What conservation implies for the partition)**.** For an untied output matrix, a common shift of every row leaves each softmax, and each backpropagated signal $\sum_i (p_i - q_i)u_i$, unchanged, so the mean row never enters the loss. Under a conserving optimizer it therefore stays at a direction fixed at initialization that training never uses, and the partition $\mathcal{O}_{\mathrm{raw}}$ is referenced to that random direction; any frequency alignment would have to come from correlations between the learned rows and it. The rows themselves keep moving, and their frequency-dependent motion is analyzed in Proposition 2. Empirically the reference direction indeed carries no frequency information: in Section G.6, every untied run whose output matrix is updated by momentum SGD places 40–47% of the rows in $\mathcal{O}_{\mathrm{raw}}$, with overlap $0.42$–$0.48$ against the size-matched frequent set, close to the chance levels $0.40$–$0.47$ set by those outlier rates.

Two ingredients can move the mean row away from its initial direction. *Row-dependent rescaling*, as in Adam, AdamW, RMSProp and signSGD, or per-row gradient clipping, makes $D_{i,t}$ differ across rows, so the zero sum is no longer preserved. *Weight tying* adds an input-path gradient whose row sum need not vanish (Section A.8). Weight decay, in contrast, only shrinks the initial mean row $\bar{u}^{(0)}$ toward the origin without changing its direction, and because $\mathcal{O}_{\mathrm{raw}}$ depends on $\bar{u}$ only through its direction, it leaves the partition referenced to that random direction. Adam and AdamW, which train our models, rescale each row differently, so the first ingredient is always present: Section A.3 computes how it moves the rows and the mean row of an untied output matrix, and Section A.8 adds the input path of tied models.

### A.3 Early phase: premise (P2), the frequency threshold and the movement of the mean row

Theorem 1 takes premises (P1) and (P2) as hypotheses. This subsection studies the early phase of training, while predictions are still close to uniform, where the accumulated update of Lemma 2 splits into a repulsion shared by all rows and an attraction scaled by token frequency. From this split, Proposition 2 derives (P2) and the frequency threshold that defines $F$, and Proposition 3 computes the movement of the mean row along $-e$ that (P1) assumes; Proposition 4 then shows how (P1) turns the threshold into the measured partition. Finally, Proposition 5 applies the same split to the unit-mean axis of Corollary 1, predicting its direction and the frequency ordering of the rows along it.

**Assumption 1** (Early phase)**.** Over steps $t < T$: (i) the output matrix is untied and trained by a row-diagonal optimizer, and the state distribution at each step is taken as given, so that only the output-path update is analyzed; (ii) predictions are near uniform, $\max_{i,x} |p_i^{(t)}(x) - 1/V| \le \varepsilon_p/V$ with $\varepsilon_p \ll 1$; (iii) the mean state keeps a fixed direction, $\bar{x}^{(t)} = X_t e$ with $\|e\|_2 = 1$ and $X_t > 0$.

Each clause serves one step of the derivation. Clause (i) lets Lemma 2 describe the whole update of each row: an untied matrix receives no input-path gradient, and the states, hence $\mu_i^{(t)}$ and $\bar{x}^{(t)}$, are taken as given. Clause (ii) makes the repulsion term $\mathbb{E}[p_i^{(t)}(x)x^{(t)}]$ of the gradient equal to $\bar{x}^{(t)}/V$ for every row up to a small remainder, so the gradient repels every row by the same vector, which each row's preconditioner then rescales; this shared repulsion is the common component of (P2). Clause (iii) keeps that vector along one direction $e$, so the repulsion accumulates over steps instead of partly cancelling, and it fixes the axis along which the frequency threshold and the movement of the mean row are measured.

Under (ii), $\mathbb{E}[p_i^{(t)}(x)x^{(t)}] = \bar{x}^{(t)}/V + \varrho_i^{(t)}$ with $\|\varrho_i^{(t)}\|_2 \le (\varepsilon_p/V)\,\mathbb{E}\|x^{(t)}\|_2$. Equation (8) over the first $T$ steps therefore splits as

$$\Delta u_i = c_i + s_i + \varrho_i,$$
$$c_i = -\frac{1}{V}\sum_{t<T} \eta_t X_t\, D_{i,t} e, \qquad s_i = \pi_i \sum_{t<T} \eta_t\, D_{i,t}\mu_i^{(t)}, \qquad \varrho_i = -\sum_{t<T} \eta_t\, D_{i,t}\varrho_i^{(t)}, \tag{9}$$

a shared repulsion $c_i$, which is the same for every row when the optimizer is row-shared, a token-specific attraction $s_i$ scaled by $\pi_i$, and a remainder $\varrho_i$ of relative order $\varepsilon_p$. Statements made *to leading order* below drop this remainder, so they hold up to corrections of relative order $\varepsilon_p$. With

$$M_j^{(i)} = \sum_{t<T} \eta_t\, \mu_j^{(t)\top} D_{i,t} e, \quad M_i = M_i^{(i)}, \quad S_i = \sum_{t<T} \eta_t\, \bar{x}^{(t)\top} D_{i,t} e = \sum_{t<T} \eta_t X_t\, e^\top D_{i,t} e > 0,$$

the projections on $e$ are $s_i^\top e = \pi_i M_i$ and $c_i^\top e = -S_i/V$, and $\bar{x}^{(t)} = \sum_j \pi_j \mu_j^{(t)}$ gives the identity

$$S_i = \sum_{j \in \mathcal{V}} \pi_j \, M_j^{(i)}. \tag{10}$$

Here $M_j^{(i)}$ is the pull of token $j$'s conditional mean along $e$, accumulated with row $i$'s preconditioner. Along $e$, row $i$ is therefore attracted by its own pull weighted by its frequency, $\pi_i M_i$, and repelled by $S_i/V$, where $S_i$ is the frequency-weighted average of the pulls of all tokens.

This split yields (P2) and the frequency threshold directly. The attraction $s_i$ carries the factor $\pi_i$ and the repulsion $c_i$ the factor $1/V$, so their ratio carries the factor $\pi_i V$, which is smaller the rarer the token. Sufficiently rare rows therefore cluster around one vector, the shared repulsion, spread by their attraction, their initialization and the row-to-row differences of the preconditioner, which is (P2). Projected on $e$, the split compares each row's attraction $\pi_i M_i$ with its repulsion $S_i/V$, and to leading order the rows in which the attraction wins are those that move toward $e$. Proposition 2 states both results with their error terms: item 1 bounds the root-mean-square deviation of the rows of a set $G$ from their mean, which gives (P2) with an explicit ratio $\bar{r}$, and item 2 gives the sign of each row's movement along $e$, which gives the frequency threshold.

**Proposition 2** (Premise (P2) and the frequency threshold, leading order)**.** *Adopt Assumption 1. For a set $G$ of rows, write $\langle\cdot\rangle_G$ for the average over $G$, let $c = \langle c_i\rangle_G$ be the average of the repulsions $c_i$ over $G$, assume $C = \|c\|_2 > 0$, and write each row at the end of the phase as $u_i^{(T)} = c + w_i$, with the row-specific remainder $w_i = u_i^{(0)} + s_i + \varrho_i + (c_i - c)$. Let*

$$\bar{\epsilon} = \big\langle \|w_i\|_2^2 \big\rangle_G^{1/2}, \qquad \bar{w} = \langle w_i\rangle_G.$$

1. ***(Premise (P2).)*** *If $\|\bar{w}\|_2 < C$, the rows of $G$ have the nonzero mean $\bar{u}_G = c + \bar{w}$, and their root-mean-square deviation from it, relative to $\|\bar{u}_G\|_2$, satisfies*

$$\bar{r} \le \frac{\bar{\epsilon}}{C - \|\bar{w}\|_2}, \qquad \bar{\epsilon} \le \bar{\epsilon}_0 + \bar{\epsilon}_s + \bar{\epsilon}_\varrho + \bar{\zeta} C,$$

   *where $\bar{\epsilon}_0$, $\bar{\epsilon}_s$ and $\bar{\epsilon}_\varrho$ are the root mean squares over $G$ of $\|u_i^{(0)}\|_2$, $\|s_i\|_2$ and $\|\varrho_i\|_2$, and $\bar{\zeta} C$ is that of $\|c_i - c\|_2$. For a row-shared optimizer $c_i = c$ for every row, so $\bar{\zeta} = 0$.*

2. ***(Frequency threshold.)*** *$\Delta u_i^\top e = \pi_i M_i - S_i/V + \varrho_i^\top e$. If the conditional means are comparable through each row's preconditioner, $M_j^{(i)} = M_i$ for all $j$, then Equation (10) gives $S_i = M_i > 0$ and*

$$\Delta u_i^\top e = M_i\Big(\pi_i - \frac{1}{V}\Big) + \varrho_i^\top e,$$

   *so to leading order rows with $\pi_i > 1/V$ move along $e$ and the rest along $-e$.*

*Proof. Item 1.* Equation (9) gives $u_i^{(T)} = u_i^{(0)} + c_i + s_i + \varrho_i = c + w_i$. Averaging over $G$ gives $\bar{u}_G = c + \bar{w}$, so $\|\bar{u}_G\|_2 \ge C - \|\bar{w}\|_2 > 0$. Since $u_i^{(T)} - \bar{u}_G = w_i - \bar{w}$ and these differences average to zero over $G$, $\langle \|u_i^{(T)} - \bar{u}_G\|_2^2\rangle_G = \bar{\epsilon}^2 - \|\bar{w}\|_2^2 \le \bar{\epsilon}^2$, and dividing by $\|\bar{u}_G\|_2^2$ bounds $\bar{r}$. The bound on $\bar{\epsilon}$ is the triangle inequality for the root mean square over $G$ (Minkowski's inequality), applied to the four terms of $w_i$. *Item 2.* Project Equation (9) on $e$; the comparable case follows from Equation (10) and $\sum_j \pi_j = 1$, and $M_i = S_i > 0$ because $S_i > 0$. □

Item 1 derives (P2) with an explicit ratio $\bar{r}$: the token-specific attraction, the initialization, the remainder and the row-to-row variation $\bar{\zeta}$ of the preconditioner, each measured in root mean square over $G$ against the accumulated shared repulsion $C$. Because each $s_i$ carries the factor $\pi_i$, $\bar{\epsilon}_s$ is small next to $C$ only on a set of rare rows, which is why Theorem 1 takes the set $G$ of (P2) inside $N$. The initialization enters $\bar{\epsilon}$ through its root mean square $\bar{\epsilon}_0$, but it enters $\bar{w}$ only through its average over $G$, which is small for a large set of independently initialized rows. When the initialization dominates the remainder and $\|\bar{w}\|_2$ is small next to both $C$ and $\bar{\epsilon}$, the identity in the proof gives $\bar{r} \approx \bar{\epsilon}_0/C$: the rows spread around the shared repulsion by their initialization. Item 2 gives the threshold that defines $F$: to leading order a row moves toward $e$ exactly when its attraction outweighs its repulsion, $\pi_i M_i > S_i/V$, which becomes $\pi_i > 1/V$ when the conditional means are comparable.

Both items describe individual rows, whereas the measured partition compares each row with the mean row, and Proposition 1 keeps the mean row fixed under row-shared updates. The partition therefore also requires the mean row to move, which is the first part of (P1); its second part, the row-wise condition of Proposition 4, requires that the rest of the mean row not reverse the side that this movement assigns to a row. Proposition 3 computes this movement along $-e$ by averaging the same early drift over the vocabulary, and shows that it vanishes when every row shares its preconditioner.

**Proposition 3** (Movement of the mean row, leading order)**.** *Adopt Assumption 1. To leading order in $\varepsilon_p$, the mean row moves along $-e$ by*

$$\kappa_e := -\big(\bar{u}^{(T)} - \bar{u}^{(0)}\big)^\top e = \frac{1}{V}\sum_{j\in\mathcal{V}} \pi_j \big(\overline{M}_j - M_j^{(j)}\big), \qquad \overline{M}_j = \frac{1}{V}\sum_{i\in\mathcal{V}} M_j^{(i)}.$$

*In particular $\kappa_e = 0$ for a row-shared optimizer, in agreement with Proposition 1, and $\kappa_e > 0$ exactly when the tokens' own preconditioners, weighted by frequency, shorten the pulls of their conditional means along $e$ more than the vocabulary-average preconditioner does. If every preconditioner acts on the relevant directions as a scalar, $D_{i,t} \approx d_{i,t} I$, then*

$$\kappa_e = \frac{1}{V}\sum_{j\in\mathcal{V}} \pi_j \sum_{t<T} \eta_t \big(\bar{d}_t - d_{j,t}\big) \mu_j^{(t)\top} e, \qquad \bar{d}_t = \frac{1}{V}\sum_{i\in\mathcal{V}} d_{i,t},$$

*so the mean row moves against the mean state when the frequent tokens, which carry most of the weight $\pi_j \mu_j^{(t)\top} e$, take smaller steps than the average row.*

*Proof.* Averaging Proposition 2(2) over rows gives $-(\bar{u}^{(T)} - \bar{u}^{(0)})^\top e = \frac{1}{V}\sum_i (S_i/V - \pi_i M_i) - \frac{1}{V}\sum_i \varrho_i^\top e$, and the last term, of relative order $\varepsilon_p$, is dropped. Substituting Equation (10) gives $\frac{1}{V}\sum_i S_i = \sum_j \pi_j \overline{M}_j$, while $\sum_i \pi_i M_i = \sum_j \pi_j M_j^{(j)}$, which gives the formula. For a row-shared optimizer $M_j^{(i)}$ does not depend on $i$, so $\overline{M}_j = M_j^{(j)}$. In the scalar case $M_j^{(i)} = \sum_t \eta_t d_{i,t}\, \mu_j^{(t)\top} e$. □

The proposition identifies what (P1) requires of the optimizer: a preconditioner that is smaller for frequent rows, in the frequency-weighted sense above. Adam's second-moment normalization is a natural source. Its per-coordinate steps have comparable size for every row, which is the sense in which the main text calls its step size largely independent of token frequency, so a frequent row, whose gradient is larger, is scaled down more; Section A.4 develops this heuristic, and Section G.6 measures the movement of the mean row directly. The proposition gives only the component $\kappa$ of the movement along $-e$. Proposition 2(2) and Proposition 3 thus give the movement of each row along $e$ and of the mean row along $-e$, but the partition $\mathcal{O}_{\text{raw}}$ is defined by the sign of $u_i^\top \bar{u}$, which also depends on each row's initialization and on the rest $\bar{u}_{\text{rest}}$ of the mean row, its initial value $\bar{u}^{(0)}$ and its movement $\Delta_\perp$ off $e$. Proposition 4 combines the two results and identifies the rows on which the partition coincides with $F$: those on which this rest does not reverse the sign set by the movement, the row-wise condition of (P1), and whose initialization does not outweigh their update along $e$.

**Proposition 4** (The partition is the frequency threshold set once the mean row moves)**.** *Write $\bar{u}^{(K)} - \bar{u}^{(0)} = -\kappa e + \Delta_\perp$ with $\Delta_\perp \perp e$, and let $\bar{u}_{\text{rest}} = \bar{u}^{(0)} + \Delta_\perp = \bar{u}^{(K)} + \kappa e$ be the rest of the mean row.* (P1) *holds on row $i$ if $\kappa > 0$ and*

$$\operatorname{sign}(v_i^\top e)\, v_i^\top \bar{u}_{\text{rest}} < \kappa\, |v_i^\top e|. \tag{11}$$

*Every row on which* (P1) *holds satisfies*

$$u_i^\top \bar{u}^{(K)} \le 0 \quad \Longleftrightarrow \quad v_i^\top e > 0,$$

*and every row with $\operatorname{sign}(v_i^\top e)\, v_i^\top \bar{u}_{\text{rest}} > \kappa |v_i^\top e|$ violates this equivalence, so Equation (11) is sharp up to ties. It holds in particular on every row with $|v_i^\top \bar{u}_{\text{rest}}| < \kappa |v_i^\top e|$, and, for $\kappa > 0$, on every row with $|v_i^\top e| > (\|\bar{u}^{(0)}\|_2 + \|\Delta_\perp\|_2)/\kappa$. If in addition the accumulated update of row $i$ dominates its initial projection, $|u_i^{(0)\top} e| < |\Delta u_i^\top e|$, then $v_i^\top e > 0$ if and only if $\Delta u_i^\top e > 0$, which by Proposition 2(2) is, to leading order, $\pi_i M_i > S_i/V$. Hence, for $\kappa > 0$, $\mathcal{O}_{\text{raw}}$ and $F = \{i : \pi_i M_i > S_i/V\}$ agree outside the band*

$$\mathcal{B} = \big\{ i : \operatorname{sign}(v_i^\top e)\, v_i^\top \bar{u}_{\text{rest}} \ge \kappa |v_i^\top e| \big\} \cup \big\{ i : |u_i^{(0)\top} e| \ge |\Delta u_i^\top e| \big\}.$$

*Proof.* Write $u_i = \|u_i\|_2 v_i$ and expand

$$u_i^\top \bar{u}^{(K)} = \|u_i\|_2 \big(-\kappa\, v_i^\top e + v_i^\top \bar{u}_{\text{rest}}\big).$$

If $v_i^\top e > 0$, the bracket is $v_i^\top \bar{u}_{\text{rest}} - \kappa |v_i^\top e|$, which is negative under Equation (11) and positive when $v_i^\top \bar{u}_{\text{rest}} > \kappa |v_i^\top e|$. If $v_i^\top e < 0$, the bracket is $\kappa |v_i^\top e| + v_i^\top \bar{u}_{\text{rest}}$, which is positive under Equation (11), where $-v_i^\top \bar{u}_{\text{rest}} < \kappa |v_i^\top e|$, and negative when $-v_i^\top \bar{u}_{\text{rest}} > \kappa |v_i^\top e|$. If $v_i^\top e = 0$, Equation (11) cannot hold. In each case covered by Equation (11) the bracket has the sign of $-v_i^\top e$, which gives the equivalence, and in each case of the reversed strict inequality it has the opposite sign. The two sufficient conditions follow from $\operatorname{sign}(v_i^\top e)\, v_i^\top \bar{u}_{\text{rest}} \le |v_i^\top \bar{u}_{\text{rest}}| \le \|\bar{u}_{\text{rest}}\|_2 \le \|\bar{u}^{(0)}\|_2 + \|\Delta_\perp\|_2$. Since $u_i^\top e = u_i^{(0)\top} e + \Delta u_i^\top e$, the sign of $v_i^\top e$ is that of $\Delta u_i^\top e$ whenever $|u_i^{(0)\top} e| < |\Delta u_i^\top e|$. □

The proposition reduces the partition to one vector, the movement of the mean row, which is directly measurable. Proposition 3 gives its component along $e$ to leading order, and Proposition 1 shows that it vanishes for momentum SGD on an untied output matrix. The condition is one-sided: it still holds when $\bar{u}_{\text{rest}}$ projects onto $v_i$ at least as strongly as the movement does, provided the two projections, $v_i^\top \bar{u}_{\text{rest}}$ and $-\kappa\, v_i^\top e$, have the same sign, so that $\bar{u}_{\text{rest}}$ reinforces the sign set by the movement, whereas the two-sided and uniform sufficient conditions fail on such rows. Section G.6 measures this movement and its alignment $\cos(\bar{u}^{(K)} - \bar{u}^{(0)}, -\bar{x})$: 0.83–0.85 in the untied Adam and AdamW runs and 0.69–0.70 in the tied ones, against a movement of order $10^{-6}$ of the average row length under momentum SGD.

Propositions 2 to 4 supply what Theorem 1 needs, but they do not describe the vector through which the main text reads the ICL signature. Corollary 1 writes $\Delta\mathcal{A} = (\Delta\widehat{x})^\top m_{\text{unit}}$ with the unit-mean axis $m_{\text{unit}} = \sum_i v_i / V$, which differs from the raw mean $\bar{u}$ because it weights every token equally. The corollary guarantees only that $m_{\text{unit}} \neq 0$; it says neither where the axis points nor how the tokens are ordered along it, as measured by the depth $\delta_i = -v_i^\top \widehat{m}$ along $\widehat{m} = m_{\text{unit}} / \|m_{\text{unit}}\|_2$ in Sections G.6 and G.7. Proposition 5 derives both from the same early-phase split.

**Proposition 5** (The unit-mean axis in the early phase, leading order)**.** *Adopt Assumption 1 and the notation of Proposition 2.*

1. ***(Axis sign.)*** *Every row of $G$ with $\|w_i\|_2 < -c^\top e$ satisfies $v_i^\top e < 0$. If $\big\|\sum_{i\notin G} v_i\big\|_2 < -\sum_{i\in G} v_i^\top e$, then $m_{\text{unit}}^\top e < 0$: the unit-mean axis points against the mean state.*

2. ***(Depth.)*** *Every row of $G$ with $\|w_i\|_2 < C$ and a negligible initial projection $u_i^{(0)\top} e$ has $v_i^\top e$ equal, up to a factor in $[(1+\|w_i\|_2/C)^{-1}, (1-\|w_i\|_2/C)^{-1}]$, to $(\pi_i M_i - S_i/V)/C$, so, up to that factor, the depth $\delta_i$ along an axis $\widehat{m} \approx -e$ increases with $\pi_i M_i - S_i/V$ among these rows.*

*Proof. Item 1.* Each $c_i^\top e = -S_i/V$ is negative, so $c^\top e < 0$, and $u_i^\top e = c^\top e + w_i^\top e \le c^\top e + \|w_i\|_2 < 0$ for the stated rows. If the rows outside $G$ satisfy the hypothesis, Cauchy–Schwarz gives $V m_{\text{unit}}^\top e = \sum_{i\in G} v_i^\top e + \big(\sum_{i\notin G} v_i\big)^\top e \le \sum_{i\in G} v_i^\top e + \big\|\sum_{i\notin G} v_i\big\|_2 < 0$. *Item 2.* $\|u_i\|_2 = \|c + w_i\|_2 \in [C - \|w_i\|_2,\, C + \|w_i\|_2]$, and with a negligible initial projection $u_i^\top e = \Delta u_i^\top e$, whose leading-order value is given by Proposition 2(2). □

Proposition 5 thus supplies the two properties of the axis that Corollary 1 leaves open. Item 1 gives its direction: every row of $G$ whose remainder is shorter than the component of the shared repulsion along $-e$ points against the mean state, and the axis follows when the rows of $G$ outweigh the rest. This depends on how many rows $G$ contains, not on their norms: $m_{\text{unit}}$ averages unit rows, so every token contributes with weight one, and a Zipf marginal makes the rare set $G$ large. The hypothesis of item 1, that the rows outside $G$ do not cancel this contribution, is an additional assumption, analogous to the non-cancellation condition of Corollary 1, rather than a consequence of Proposition 2. Item 2 gives the order of the rows along the axis: among the rows whose remainder is shorter than the shared repulsion, a row's depth grows with its attraction relative to its repulsion, $\pi_i M_i - S_i/V$, which is $M_i(\pi_i - 1/V)$ when the conditional means are comparable, so rows with similar $M_i$ are ordered by frequency. Section G.7 tests both items, together with Propositions 2 and 3, on Pythia-1B checkpoints, and Section G.6 reports the depth–frequency correlation for every optimizer.

**Remark 2** (Scope: the early phase)**.** Propositions 2, 3 and 5 describe the phase in which Assumption 1 holds and make no prediction beyond it. Equation (9) is accurate to leading order in the accumulated logit spread, and that spread is produced by the attractions $s_i$, which are large only for the *frequent* rows: a repulsion shared by all rows shifts every logit by the same amount and leaves the softmax unchanged. Through the softmax normalizer the spread perturbs $p_i(x)$ for every token, so the phase in which Assumption 1(ii) holds ends once the frequent rows have moved far enough to make the predictions non-uniform, however little the rare rows have moved by then. In Pythia-1B, predictions on the rare tokens stay close to uniform for about the first 64 steps, and every sign predicted by the three propositions appears within this window (Section G.7). Afterwards the frequency ordering and the axis norm $\|m_{\text{unit}}\|_2$ persist, whereas the anti-alignment of the axis with the mean state keeps its sign but weakens, from $-0.94$ at step 128 to $-0.02$ at step 143,000 (Table 12). Because the residual stream admits an approximate change of basis absorbed by the surrounding weights, axis directions are compared only within a checkpoint.

### A.4 Which optimizer realizes the premises

Propositions 1 and 3 say that the partition requires a preconditioner that differs across rows and is smaller for frequent rows. They do not say which optimizer delivers that together with (P2), and Section G.6 shows that per-coordinate normalization alone does not. The optimizers can be compared through $g_i$, the negative gradient of the batch loss with respect to row $i$, as in Definition 1. By Lemma 1, $g_i$ is the batch average of $(\mathbf{1}\{y = i\} - p_i(x))\, x$, which splits into two parts. The *background*, the batch average of $-p_i(x)x$, is present at every step; in the early phase it equals $-\bar{x}_{\text{batch}}/V$ for every row, where $\bar{x}_{\text{batch}}$ is the average final state over the batch, so all rows receive the same vector. The *attraction*, the batch average of $\mathbf{1}\{y = i\}x$, is nonzero only at steps whose batch contains token $i$ as a target and averages to $\pi_i\mu_i$, so for a rare token it is small and intermittent. How an optimizer turns these two parts into a step determines which premises it satisfies. The account below is a heuristic rather than a derivation.

*SGD* updates $u_i \leftarrow u_i + \eta\, g_i$, and momentum replaces $g_i$ by a running average of past gradients. Either way it applies one preconditioner to every row, so by Proposition 1 the mean row of an untied matrix stays at its initialization, and the raw-mean partition is referenced to a random direction (Remark 1). Its step is proportional to the raw gradient, which is small for a rare row, $\pi_i < 1/V$: the background has size $\|\bar{x}\|_2/V$, and the attraction averages $\pi_i\|\mu_i\|_2 < \|\mu_i\|_2/V$, so when $\|\mu_i\|_2$ and $\|\bar{x}\|_2$ are comparable the whole gradient is at most of order $\|\bar{x}\|_2/V$. The gradient of a frequent row $f$ is instead dominated by its attraction, of size about $\pi_f\|\mu_f\|_2$, which under the same comparability is larger by a factor of about $\pi_f V$. In the same steps, a rare row therefore covers only a fraction of order $1/(\pi_f V)$ of the distance a frequent row covers. The early phase ends once the frequent rows have moved far enough to make the predictions non-uniform (Remark 2), so by then the accumulated shared repulsion $C$ of the rare rows is smaller than the movement of the frequent rows by this factor and need not exceed the root mean square $\bar{\epsilon}_0$ of their initialization. The rare rows then keep nearly independent random directions, their deviation from their mean is large next to the mean itself, and the bound of Theorem 1 is vacuous for them.

*signSGD* updates $u_i \leftarrow u_i + \eta\,\text{sign}(g_i)$, with the sign taken separately for each of the $d$ coordinates of $g_i$. Write $B$ for the number of target positions in a batch, $x_b$ and $y_b$ for the state and the target at position $b$, $n_i$ for the number of positions with $y_b = i$, and $\bar{x}_{i,\text{batch}}$ for the average of $x_b$ over them. The batch version of Equation (5) replaces $\pi_i\mu_i$ by $(n_i/B)\,\bar{x}_{i,\text{batch}}$ and $\mathbb{E}[p_i(x)x]$ by its batch average, so coordinate $k$ of row $i$ moves by

$$\eta\,\text{sign}\Big(\frac{n_i}{B}\,\bar{x}_{i,\text{batch},k} - \frac{1}{B}\sum_b p_i(x_b)\,x_{b,k}\Big) \approx \eta\,\text{sign}\Big(\frac{n_i}{B}\,\bar{x}_{i,\text{batch},k} - \frac{\bar{x}_{\text{batch},k}}{V}\Big),$$

where the approximation holds in the early phase. The second term, the background, has the same sign for every row, that of $-\bar{x}_{\text{batch},k}$, so at every coordinate where it outweighs the attraction, all rows take the same step. When the states of token $i$ resemble the batch mean, the attraction outweighs the background only if $n_i > B/V$, the count expected under uniform frequencies. For most rows with $\pi_i$ well below $1/V$, $\mathbb{E}[n_i] = B\pi_i$ is well below $B/V$, so this happens at a minority of steps, and the rare rows move along one common sign vector, which predicts that they cluster tightly around their mean. In contrast, a frequent row has $n_i \gg B/V$ at most steps and follows the signs of its attraction, those of $\bar{x}_{i,\text{batch}} \approx \mu_i$. When the final states share a strong common direction, each $\mu_i$ is

that direction plus a smaller token-specific part, so its coordinate signs mostly agree with those of $\bar{x}$; the frequent rows then also move along nearly one sign vector, which predicts frequent rows that are not dispersed. The partition can still carry frequency information, because how often the attraction rather than the background sets the sign depends on how often the token occurs.

*Adam-family optimizers*, Adam and AdamW, keep running estimates of the first and second moments of each coordinate of $g_i$,

$$m_i^{(t)} = \beta_1 m_i^{(t-1)} + (1-\beta_1)\, g_i^{(t)}, \qquad v_i^{(t)} = \beta_2 v_i^{(t-1)} + (1-\beta_2)\left(g_i^{(t)}\right)^2,$$

with the square taken coordinatewise and $m_i^{(0)} = v_i^{(0)} = 0$, and update $u_i \leftarrow u_i + \eta\, \hat{m}_i/(\sqrt{\hat{v}_i} + \epsilon_{\mathrm{A}})$ coordinatewise, where $\hat{m}_i = m_i^{(t)}/(1-\beta_1^t)$ and $\hat{v}_i = v_i^{(t)}/(1-\beta_2^t)$ correct the bias of the zero initialization. AdamW also multiplies every row by $1-\eta\lambda$ at each step, which rescales all rows alike and does not affect the argument below. Since $\hat{m}_i$ is linear in the past gradients, its average over steps is the mean gradient, while with $\beta_2$ close to one $\hat{v}_i$ changes slowly and estimates the mean square. Writing $\bar{g}_k$ and $\sigma_k^2$ for the mean and variance of coordinate $k$ of a row's gradient, the average step at coordinate $k$ is therefore approximately $\eta\, \bar{g}_k/\sqrt{\bar{g}_k^2 + \sigma_k^2}$ when $\epsilon_{\mathrm{A}}$ is negligible. This step is unchanged when the gradient is rescaled, so a small gradient moves a row as fast as a large one; it approaches $\eta$ when the gradient is consistent across steps, $\sigma_k \ll |\bar{g}_k|$, and shrinks when the gradient fluctuates. The background is consistent across steps and the same for every row, so it moves every row, rare or frequent, by about $\eta$ per coordinate along a common direction. Unlike under SGD, the accumulated shared repulsion $C$ of the rare rows is therefore of the order of the movement of the frequent rows, which lets it outgrow their initialization and keeps their deviation $\bar{r}$ moderate. Unlike the sign, the running mean retains each row's attraction in proportion to how often the token occurs, which gives frequent rows token-specific movements and keeps the frequency dependence that the partition needs. The same normalization makes the preconditioner of Lemma 2 approximately a scalar $d_{i,t}$ of order $1/\sqrt{\bar{g}^2 + \sigma^2}$, inversely proportional to the size of the row's gradient, which grows with target frequency because the attraction does, so frequent rows receive the smallest $d_{i,t}$. With comparable conditional means, $\mu_j^{(t)\top} e \approx X_t$, the scalar form of Proposition 3 becomes $\kappa_e \approx V^{-1} \sum_{t<T} \eta_t X_t \big(\bar{d}_t - \sum_j \pi_j d_{j,t}\big)$, which is positive because the frequency-weighted average of $d_{j,t}$ puts most weight on the frequent rows, whose $d_{j,t}$ is smallest.

Section G.6 tests these predictions and confirms them. The reading of Adam above has its own range of validity: it requires $\hat{v}_i$ to average many gradients drawn from roughly the same distribution. The attraction of token $i$ enters only at the steps where the token occurs, about $B\pi_i$ times per step, so the window of about $1/(1-\beta_2)$ steps over which $\hat{v}_i$ averages must contain several occurrences, $1/(1-\beta_2) \gtrsim 1/(B\pi_i)$; otherwise $\hat{v}_i$ jumps at each occurrence and decays in between. For tokens so rare that $\sqrt{\hat{v}_i}$ falls to the scale of $\epsilon_{\mathrm{A}}$, the step becomes $\eta\, \hat{m}_i/\epsilon_{\mathrm{A}}$, proportional to the raw gradient as under SGD.

### A.5 Formal statement and proof of Theorem 1

We now assemble these pieces into the formal version of Theorem 1: premise (P2), derived in Proposition 2(1), yields spectral concentration, and Proposition 4 yields the partition. Write

$$\mathrm{EDim}(U) = \frac{\left(\sum_k \sigma_k^2(U)\right)^2}{\sum_k \sigma_k^4(U)}, \qquad A_U = \sum_{i \in \mathcal{V}} \|u_i\|_2^2.$$

**Definition 2** (Raw-mean outlier set)**.** For nonzero rows and a nonzero raw vocabulary mean $\bar{u} = \frac{1}{V}\sum_i u_i$, define $\mathcal{O}_{\mathrm{raw}}(U) = \{i \in \mathcal{V} : u_i^\top \bar{u} \le 0\}$, equivalently $\{i : \cos(u_i, \bar{u}) \le 0\}$.

**Theorem 2** (Formal version of Theorem 1)**.** *Let $U = U^{(K)}$ be the output matrix, with nonzero rows, after $K$ steps of a row-diagonal optimizer, let $e$ be a unit vector, and let $F_\Delta = \{i : \Delta u_i^\top e > 0\}$ and $N_\Delta = \mathcal{V} \setminus F_\Delta$.*

1. ***(Spectral concentration under** (P2)**.)** Let $G$ be a set of rows with nonzero mean row $\bar{u}_G = \frac{1}{|G|}\sum_{i\in G} u_i$, let $\bar{r} \ge 0$ be the root-mean-square deviation of these rows from $\bar{u}_G$ relative to its norm, $\bar{r}^2 = \frac{1}{|G|}\sum_{i\in G} \|u_i - \bar{u}_G\|_2^2/\|\bar{u}_G\|_2^2$, and let $\omega_{\mathrm{g}} = \sum_{i\in G}\|u_i\|_2^2/A_U$. Then*

$$\frac{\mathrm{EDim}(U)}{d} \le \frac{(1+\bar{r}^2)^2}{d\,\omega_{\mathrm{g}}^2}.$$

2. ***(Partition under* (P1).)** *Suppose the mean row moves along $-e$, $\kappa = -(\bar{u}^{(K)} - \bar{u}^{(0)})^\top e > 0$, and let $\mathcal{B}$ be the band of Proposition 4, the rows on which the row-wise condition of* (P1) *fails or the initial projection on $e$ outweighs the accumulated update. Then $\mathcal{O}_{\mathrm{raw}} \setminus \mathcal{B} = F_\Delta \setminus \mathcal{B}$. With the band fractions $\eta = |\mathcal{B} \cap N_\Delta|/|N_\Delta|$ and $\eta_F = |\mathcal{B} \cap F_\Delta|/|F_\Delta|$,*

$$\frac{(1-\eta_F)|F_\Delta|}{V} \le \frac{|\mathcal{O}_{\mathrm{raw}}|}{V} \le \frac{|F_\Delta| + \eta|N_\Delta|}{V}, \qquad \frac{|\mathcal{O}_{\mathrm{raw}} \cap F_\Delta|}{|\mathcal{O}_{\mathrm{raw}}|} \ge \frac{(1-\eta_F)|F_\Delta|}{|F_\Delta| + \eta|N_\Delta|}, \tag{12}$$

*whenever the sets in the denominators are nonempty, and $\mathcal{O}_{\mathrm{raw}} = F_\Delta$ when the band is empty.*

3. ***(Early phase.)** Under Assumption 1 with $T = K$, $F_\Delta$ equals $F = \{i : \pi_i M_i > S_i/V\}$ to leading order in $\varepsilon_p$, Proposition 2(1) bounds $\bar{r}$ on any $G \subseteq N$ with $\|\bar{w}\|_2 < C$ by $\bar{\epsilon}/(C - \|\bar{w}\|_2)$, and Proposition 3 supplies the component $\kappa = \kappa_e$ of* (P1) *along $e$.*

*Proof. Item 1.* Let $K_G = U_G U_G^\top$ be the Gram matrix of the rows of $G$. Retaining only the pairs in $G$,

$$\sum_k \sigma_k^4(U) = \|UU^\top\|_{\mathrm{F}}^2 = \sum_{i,j} (u_i^\top u_j)^2 \ge \|K_G\|_{\mathrm{F}}^2 \ge \Big(\frac{\mathbf{1}^\top K_G \mathbf{1}}{|G|}\Big)^2 = |G|^2 \|\bar{u}_G\|_2^4,$$

where the second inequality bounds the Frobenius norm of the positive semidefinite $K_G$ below by its largest eigenvalue, and that eigenvalue below by the Rayleigh quotient at the unit vector $\mathbf{1}/\sqrt{|G|}$. Since the deviations $u_i - \bar{u}_G$ sum to zero over $G$, $\sum_{i\in G} \|u_i\|_2^2 = |G|\,\|\bar{u}_G\|_2^2(1+\bar{r}^2)$, so $|G|\,\|\bar{u}_G\|_2^2 = \omega_{\mathrm{g}} A_U/(1+\bar{r}^2)$. With $\sum_k \sigma_k^2(U) = A_U$ we obtain $\mathrm{EDim}(U) \le A_U^2(1+\bar{r}^2)^2/(\omega_{\mathrm{g}} A_U)^2 = (1+\bar{r}^2)^2/\omega_{\mathrm{g}}^2$. *Item 2.* By Proposition 4, every $i \notin \mathcal{B}$ satisfies $i \in \mathcal{O}_{\mathrm{raw}} \Leftrightarrow v_i^\top e > 0 \Leftrightarrow \Delta u_i^\top e > 0 \Leftrightarrow i \in F_\Delta$. Hence $F_\Delta \setminus \mathcal{B} \subseteq \mathcal{O}_{\mathrm{raw}} \subseteq F_\Delta \cup (\mathcal{B} \cap N_\Delta)$, so $|\mathcal{O}_{\mathrm{raw}}| \ge (1-\eta_F)|F_\Delta|$, $|\mathcal{O}_{\mathrm{raw}}| \le |F_\Delta| + \eta|N_\Delta|$ and $|\mathcal{O}_{\mathrm{raw}} \cap F_\Delta| \ge (1-\eta_F)|F_\Delta|$, which give Equation (12). *Item 3.* By Proposition 2(2), $\Delta u_i^\top e = \pi_i M_i - S_i/V + \varrho_i^\top e$ with $\varrho_i$ of relative order $\varepsilon_p$; the remaining statements are Proposition 2(1) and Proposition 3. □

The spectral bound holds for every set $G$ with a nonzero mean row. It is informative beyond the universal bound $\mathrm{EDim}(U)/d \le 1$ when $(1+\bar{r}^2)^2 < d\,\omega_{\mathrm{g}}^2$, which (P2) in Theorem 1 states as $(1+\bar{r}^2)/\omega_{\mathrm{g}} \ll \sqrt{d}$, and it caps the effective dimension at $(1+\bar{r}^2)^2/\omega_{\mathrm{g}}^2$ independently of the width. Its two inputs are how tightly the rows of $G$ cluster around their mean, measured by $\bar{r}$, and their share of the squared row norms, and Proposition 2(1) derives $\bar{r}$ from the early drift rather than assuming it.

The two statements rest on different premises and can fail independently. The spectral bound needs only (P2), whereas the partition needs the mean row to move along $-e$ without the rest $\bar{u}_{\mathrm{rest}}$ of the mean row reversing the side of the rows, which is (P1). Whether the mean row moves depends on the optimizer: Proposition 1 rules it out for row-shared optimizers on an untied output matrix, and Proposition 3 attributes it to a preconditioner that differs across rows. signSGD in Section G.6 separates the two premises: it satisfies (P2) most tightly, with $\bar{r} = 0.26$ on the rare rows, but its mean row moves mostly off $-\bar{x}$ (cosine 0.225), so its rest $\bar{u}_{\mathrm{rest}}$ is 4.3 times as long as $\kappa$ and can reverse the side of many rows (Proposition 4), and its partition captures only 0.30 of the frequent set. When the conditional means are comparable, $F$ reduces to $\{i : \pi_i > 1/V\}$, and the threshold match of Table 2 compares $\mathcal{O}_{\mathrm{raw}}$ with this set directly; the other experiments use the size-matched frequent set described at the start of this appendix.

### A.6 Additional group-level geometric analysis

The main theorem concerns full-matrix effective dimension and the raw-mean partition. The supplementary experiments in Section G.10 additionally compare frequent and non-frequent rows using three group-level summaries. We first give exact metric identities and then extend premise (P2) of Theorem 1 from the set $G$ to bounds on the complete groups. The extension permits non-frequent rows outside $G$, accounting for their number and for the share of non-frequent row energy that $G$ covers. On the frequent side it requires only average dispersion.

The results in this subsection use $F$ only through its rows, so they hold for any split of the vocabulary into a group $F$ and its complement $N = \mathcal{V} \setminus F$ that contains the set $G$ of (P2): for the threshold set $F$ of Theorem 1, and for the size-matched frequent set of the experiments with $G$ taken inside its

complement. We write $n_F = |F|$ and $n_N = |N|$. For any nonempty group $Q$ of nonzero rows, let $n_Q = |Q|$, $r_i = \|u_i\|_2$, $v_i = u_i/r_i$, $\bar{u}_Q = \frac{1}{n_Q}\sum_{i\in Q} u_i$, and $\bar{v}_Q = \frac{1}{n_Q}\sum_{i\in Q} v_i$. First, define the *cohesion* of $Q$ as

$$\mathrm{Coh}(Q) = \frac{1}{n_Q^2}\sum_{\substack{i,j\in Q\\ i\neq j}} v_i^\top v_j. \tag{13}$$

Each inner product is the cosine similarity between two distinct normalized rows. Cohesion therefore measures their aggregate within-group directional agreement: it is large and positive when many rows point similarly, and small when their directions cancel or disperse.

Second, define the *compactness* of $Q$ as

$$\mathrm{Comp}(Q) = \frac{1}{n_Q}\sum_{i\in Q}\cos(u_i, \bar{u}_Q), \qquad \bar{u}_Q = \frac{1}{n_Q}\sum_{i\in Q} u_i. \tag{14}$$

Compactness averages the cosine similarity between every row and the group's raw centroid. It measures how tightly the rows cluster around that centroid and, unlike cohesion, allows unequal row norms to influence the reference direction $\bar{u}_Q$. We set $\mathrm{Comp}(Q) = 0$ when $\bar{u}_Q = 0$.

Finally, because the participation-rank effective dimension $\mathrm{EDim}(\cdot)$ is already defined in Section 4.1, define the *group-normalized effective dimension* as

$$\mathrm{GNEDim}(Q) = \frac{\mathrm{EDim}(U_Q)}{n_Q}. \tag{15}$$

This is the participation rank per group row. Larger values indicate that the rows collectively spread their energy across more singular directions relative to the group's cardinality; smaller values indicate greater directional redundancy and spectral concentration. Dividing by $n_Q$ permits comparison between the differently sized frequent and non-frequent groups.

**Lemma 3** (Exact group-geometry identities)**.** *For every group $Q$,*

$$\mathrm{Coh}(Q) = \|\bar{v}_Q\|_2^2 - \frac{1}{n_Q}, \tag{16}$$

$$\mathrm{Comp}(Q) = \chi_Q\sqrt{\mathrm{Coh}(Q) + \frac{1}{n_Q}}, \tag{17}$$

$$\mathrm{GNEDim}(Q) = \frac{(\sum_i r_i^2)^2}{n_Q\sum_{i,j} r_i^2 r_j^2 (v_i^\top v_j)^2}, \tag{18}$$

*where $\chi_Q = \cos(\bar{u}_Q, \bar{v}_Q)$ when both centroids are nonzero and $\chi_Q = 0$ otherwise, consistently with the convention $\mathrm{Comp}(Q) = 0$ when $\bar{u}_Q = 0$.*

*Proof.* Expanding $\|\bar{v}_Q\|_2^2$ yields $n_Q/n_Q^2$ from the diagonal and the ordered off-diagonal sum, proving equation 16. When both centroids are nonzero,

$$\mathrm{Comp}(Q) = \bar{v}_Q^\top\frac{\bar{u}_Q}{\|\bar{u}_Q\|_2} = \chi_Q\|\bar{v}_Q\|_2,$$

which, with equation 16, proves equation 17. If either centroid is zero, both sides of equation 17 are zero under the stated conventions. Finally, $\sum_k \sigma_k^2(U_Q) = \sum_i r_i^2$ and

$$\sum_k \sigma_k^4(U_Q) = \|U_Q U_Q^\top\|_{\mathrm{F}}^2 = \sum_{i,j} r_i^2 r_j^2 (v_i^\top v_j)^2,$$

giving equation 18. □

All three metrics are functions of the pairwise cosines $v_i^\top v_j$ within the group, taken in different ways. Cohesion averages them with equal weights. Compactness is the norm $\sqrt{\mathrm{Coh}(Q) + 1/n_Q}$ of the centroid of the normalized rows, rescaled by $\chi_Q$, the cosine between that centroid and the raw centroid, which falls below one when unequal row norms rotate the raw centroid. Group-normalized effective

dimension averages their squares with row-energy weights $r_i^2 r_j^2$, so it treats opposite directions as redundant and gives long rows more influence. The three metrics therefore capture different aspects of the same group geometry. Non-frequent rows that cluster around a common mean favor positive pairwise cosine, alignment with the group centroid, and concentration of group singular energy, subject to the terms for the non-frequent rows outside $G$ retained below. A dispersed frequent group has a smaller normalized centroid and distributes its energy more broadly, subject to row-norm imbalance.

**Theorem 3** (Full-group geometry under premise (P2))**.** *Assume $n_F, n_N \geq 2$. Let the set $G \subseteq N$ of (P2) contain $g \geq 2$ rows with nonzero mean row $\bar{u}_G$, let $\bar{r}_G$ be the root-mean-square deviation of these rows from $\bar{u}_G$ relative to its norm, as in Theorem 2(1), and let*

$$\rho_G = \frac{1}{g(g-1)} \sum_{\substack{i,j \in G \\ i \neq j}} v_i^\top v_j$$

*be the average cosine between distinct normalized rows of $G$. Write $n_Q = |Q|$,*

$$A_Q = \sum_{i \in Q} r_i^2, \qquad B_Q = \sum_{i \in Q} r_i^4, \qquad \nu_Q = \frac{n_Q B_Q}{A_Q^2},$$

*and define the share of non-frequent row energy that $G$ covers by*

$$\omega_N = \frac{\sum_{i \in G} r_i^2}{A_N}.$$

*Assume the frequent rows satisfy the aggregate dispersion bounds*

$$\frac{1}{n_F(n_F-1)} \sum_{\substack{i,j \in F \\ i \neq j}} v_i^\top v_j \leq \rho_1,$$
$$\frac{\sum_{\substack{i,j \in F \\ i \neq j}} r_i^2 r_j^2 (v_i^\top v_j)^2}{A_F^2 - B_F} \leq \rho_2^2.$$

*Here $\rho_1 \in [-1, 1]$ and $0 \leq \rho_2 \leq 1$. Then, with*

$$L_N = \frac{\rho_G g(g-1) - [n_N(n_N-1) - g(g-1)]}{n_N^2},$$
$$U_F = \frac{n_F - 1}{n_F} \rho_1,$$
$$D_N = \frac{n_N \omega_N^2}{(1 + \bar{r}_G^2)^2},$$
$$D_F = \nu_F + (n_F - \nu_F)\rho_2^2,$$

*the complete groups obey*

$$\mathrm{Coh}(N) \geq L_N, \qquad \mathrm{Coh}(F) \leq U_F, \tag{19}$$
$$\mathrm{Comp}(N) \geq \underline{\chi}_N \sqrt{\max\{0, L_N + 1/n_N\}}, \qquad \mathrm{Comp}(F) \leq \overline{\chi}_F \sqrt{U_F + 1/n_F}, \tag{20}$$
$$\mathrm{GNEDim}(N) \leq \frac{1}{D_N}, \qquad \mathrm{GNEDim}(F) \geq \frac{1}{D_F}, \tag{21}$$

*whenever $0 \leq \underline{\chi}_N \leq \chi_N$ and $\chi_F \leq \overline{\chi}_F$ with $0 \leq \overline{\chi}_F \leq 1$. Consequently,*

$$L_N > U_F, \quad \underline{\chi}_N \sqrt{\max\{0, L_N + 1/n_N\}} > \overline{\chi}_F \sqrt{U_F + 1/n_F}, \quad D_F < D_N$$

*respectively imply* $\mathrm{Coh}(N) > \mathrm{Coh}(F)$, $\mathrm{Comp}(N) > \mathrm{Comp}(F)$, *and* $\mathrm{GNEDim}(F) > \mathrm{GNEDim}(N)$ *for the full groups.*

*Proof of Theorem 3.* Among the $n_N(n_N-1)$ ordered distinct pairs in $N$, exactly $g(g-1)$ have both members in $G$. Their cosines sum to $\rho_G g(g-1)$; every other cosine is at least $-1$. Summing these

two contributions and dividing by $n_N^2$ gives $\mathrm{Coh}(N) \geq L_N$. The definition of $\rho_1$ similarly gives $\mathrm{Coh}(F) \leq U_F$.

By Lemma 3, $\mathrm{Comp}(Q) = \chi_Q \sqrt{\mathrm{Coh}(Q) + 1/n_Q}$. Applying the stated lower and upper bounds on $\chi_N$ and $\chi_F$, together with the cohesion bounds, proves equation 20. The maximum with zero is valid because $\mathrm{Coh}(N) + 1/n_N = \|\bar{v}_N\|_2^2 \geq 0$.

For group-normalized effective dimension, the argument of Theorem 2(1), applied to the pairs inside $G$ with the diagonal included, gives

$$\sum_{i,j \in N} r_i^2 r_j^2 (v_i^\top v_j)^2 \geq g^2 \|\bar{u}_G\|_2^4 = \frac{\left(\sum_{i \in G} r_i^2\right)^2}{(1 + \bar{r}_G^2)^2} = \frac{\omega_N^2 A_N^2}{(1 + \bar{r}_G^2)^2}.$$

Substitution into equation 18 yields $\mathrm{GNEDim}(N) \leq 1/D_N$. The bound $\rho_2$ on the norm-weighted average squared cosine in $F$ gives

$$\sum_{i,j \in F} r_i^2 r_j^2 (v_i^\top v_j)^2 \leq B_F + \rho_2^2 (A_F^2 - B_F),$$

and hence $\mathrm{GNEDim}(F) \geq 1/D_F$. Direct comparison of the corresponding bounds proves the three stated ordering conditions. □

The theorem extends (P2) from the set $G$ to the complete groups: whenever the three stated conditions hold, the non-frequent group is more cohesive and more compact than the frequent group and has a smaller group-normalized effective dimension, which are the orderings measured in Section G.10. Each bound pays for the non-frequent rows outside $G$ in one way: through their count in $L_N$, where their pairwise cosines are bounded only by $-1$; through the centroid factor $\chi_N$ in the compactness bound; and through the covered energy fraction $\omega_N$ in $D_N$. On the frequent side the theorem needs only small average cosine and average squared cosine, so no single unusually similar pair invalidates it. These assumptions are geometric, and the next lemma shows that marginal frequency alone cannot guarantee the frequent-side one.

**Lemma 4** (Marginal frequency alone does not determine the geometry)**.** *Fix any target-frequency vector $\pi \in \mathbb{R}^V$ with $\pi_i > 0$ and $\sum_i \pi_i = 1$. At the same initial readout $U = 0$, two data distributions with this identical $\pi$ can produce a rank-at-most-one matrix of output-side CE updates or a matrix of rank $\min(d, V-1)$, the largest rank these updates can have. Hence $\pi$ alone does not determine the rank of the output-side updates, and in particular cannot imply the frequent-row dispersion assumed in Theorem 3.*

*Proof.* At $U = 0$, every logit is zero and $p_i(x) = \frac{1}{V}$. By Lemma 1, the negative output-side gradient for row $i$ is

$$g_i = \pi_i \mu_i - \frac{1}{V} \bar{\mu}, \qquad \bar{\mu} = \mathbb{E}[x] = \sum_{j=1}^{V} \pi_j \mu_j.$$

First choose $\mu_i = m$ for every token, where $m \neq 0$. Then $g_i = (\pi_i - \frac{1}{V}) m$, so every update is collinear and the update matrix has rank at most one; in particular every frequent row, $\pi_i > \frac{1}{V}$, moves along $m$, so the frequent rows are not dispersed. The rows of the update matrix always sum to $\sum_i \pi_i \mu_i - \bar{\mu} = 0$, so its rank is at most $\min(d, V-1)$. For the second construction with $d \geq V$, choose orthonormal vectors $e_1, \ldots, e_V$ and set $\mu_i = e_i$. Then

$$g_i = \pi_i e_i - \frac{1}{V} \sum_{j=1}^{V} \pi_j e_j.$$

If $\sum_i c_i g_i = 0$, the coefficient of $e_j$ gives $\pi_j (c_j - \frac{1}{V} \sum_i c_i) = 0$. Since every $\pi_j > 0$, all $c_j$ are equal; the all-ones relation is the only linear dependence, so the row span has dimension $V - 1$. If $d < V$, choose an orthonormal basis $e_1, \ldots, e_d$ of $\mathbb{R}^d$ and set $\mu_i = e_i$ for $i \leq d$ and $\mu_i = 0$ for $i > d$. Then $\bar{\mu} = \sum_{j \leq d} \pi_j e_j$, and every row with $i > d$ equals $-\bar{\mu}/V$, so $\bar{\mu}$ lies in the row span. Hence so does $\pi_i e_i = g_i + \bar{\mu}/V$ for every $i \leq d$, and since every $\pi_i > 0$ the rows span $\mathbb{R}^d$, a rank of $d$. □

The lemma isolates the role of the conditional means $\mu_i$. At the first update, frequency sets only the size of each row's attraction, $\pi_i\mu_i$, while the conditional means set its direction: identical means make every update collinear, whereas means that span the state space make the updates span it too. Removing the topic structure from our synthetic corpus realizes the first construction: the target of a language-modeling token no longer depends on its prefix, so $\mu_i^{(t)} = \bar{x}^{(t)}$, the early drift becomes $(\pi_i - 1/V)\, D_{i,t}\bar{x}^{(t)}$, collinear up to the preconditioner, and the comparability condition of Proposition 2(2) holds exactly. This is the control of Table 2 that keeps the spectrum concentrated and places the outliers on the threshold set $\{i : \pi_i > 1/V\}$.

### A.7 Proof of the global alignment-axis corollary

The main-text corollary connects the non-frequent rows of premise (P2) to mean unembedding alignment. It controls the possible cancellation by the remaining non-frequent rows and by the frequent rows through their aggregate vector norms. Let $G \subseteq N$ be the set of (P2) in Theorem 1, let $E = N \setminus G$, and retain the frequent set $F$. The following refinement expresses their norms through observable group coherence. For $Q \in \{G, E, F\}$, set $q = |Q|$ and $s_Q = \|\sum_{i\in Q} v_i\|_2$. When $q > 1$, let $\rho_Q$ be its average off-diagonal cosine; set $s_Q = 0$ when $Q$ is empty. If $s_G > 0$, define $\widehat{m}_G = s_G^{-1}\sum_{i\in G} v_i$. Then

$$m_{\text{unit}}^\top \widehat{m}_G \geq B_{\text{coh}} := \frac{s_G - s_E - s_F}{V}, \qquad s_Q = \sqrt{q[1 + (q-1)\rho_Q]} \quad (q > 1). \tag{22}$$

Thus $B_{\text{coh}} > 0$ is the directly testable coherence form of the condition in Corollary 1.

*Proof of Corollary 1.* For any group $Q$ of $q > 1$ normalized rows,

$$\left\|\sum_{i\in Q} v_i\right\|_2^2 = \sum_{i,j\in Q} v_i^\top v_j = q + q(q-1)\rho_Q,$$

which proves $s_Q = \sqrt{q[1 + (q-1)\rho_Q]}$; the cases $q = 0$ and $q = 1$ follow directly. Since $\sum_{i\in G} v_i^\top \widehat{m}_G = s_G$, Cauchy–Schwarz applied to the sums over $E$ and $F$ gives

$$m_{\text{unit}}^\top \widehat{m}_G = \frac{1}{V}\Big(s_G + \sum_{i\in E} v_i^\top \widehat{m}_G + \sum_{i\in F} v_i^\top \widehat{m}_G\Big) \geq \frac{s_G - s_E - s_F}{V},$$

which proves equation 22. If $B_{\text{coh}} > 0$, then $m_{\text{unit}}$ has positive projection on $\widehat{m}_G$ and therefore cannot be zero. More generally, the condition of Corollary 1 and the reverse triangle inequality give $\|V m_{\text{unit}}\|_2 \geq s_G - s_E - s_F > 0$. Finally,

$$\mathcal{A}(x) = \frac{1}{V}\sum_i \widehat{x}^\top v_i = \widehat{x}^\top m_{\text{unit}};$$

where $\widehat{x} = x/\|x\|_2$. Subtracting the paired ZSL value proves the alignment identity. □

The three terms in $B_{\text{coh}}$ respectively measure the coherent non-frequent majority and the maximum aggregate cancellation allowed by the remaining non-frequent rows $E$ and the frequent rows $F$. A positive difference certifies a nonzero normalized vocabulary mean along the learned coherent direction. The paired ICL–ZSL difference still depends on the separate prompt-time factor $(\Delta\widehat{x})^\top m_{\text{unit}}$, as tested in Section G.3.

### A.8 Weight tying

The standard synthetic model shares its input and output embeddings, whereas Lemma 2 and Assumption 1 treat an untied output matrix. With tying, the gradient reaching row $i$ also contains the gradient with respect to the embedding of token $i$; write $g_i^{\text{in},(t)}$ for its negative. A row-diagonal optimizer applies one preconditioner to the sum of the two paths, so by linearity

$$\Delta u_i = \Delta u_i^{\text{out}} + \Delta u_i^{\text{in}}, \qquad \Delta u_i^{\text{in}} = \sum_{t<K} \eta_t\, D_{i,t}\, g_i^{\text{in},(t)},$$

where $\Delta u_i^{\text{out}}$ is Equation (8) evaluated along the tied run. The input path therefore enters the two premises of Theorem 1 additively.

**Proposition 6** (Weight tying and the premises of Theorem 1). *1. **(Premise** (P1)**.)** The mean row moves by $\bar{u}^{(K)} - \bar{u}^{(0)} = \frac{1}{V}\sum_i \Delta u_i^{\text{out}} + \frac{1}{V}\sum_i \Delta u_i^{\text{in}}$. For a row-shared optimizer the first term vanishes, so the mean row moves exactly by the mean input-path update.*

*2. **(Premise** (P2)**.)** Suppose that without the input path the rows of $G$, $u_i^{(0)} + \Delta u_i^{\text{out}}$, have a nonzero mean $\bar{a}$ with $C = \|\bar{a}\|_2$ and root-mean-square deviation $\bar{r}C$ from it, and that their input-path updates $\Delta u_i^{\text{in}}$ have mean $c_{\text{in}}$ and root-mean-square deviation $\bar{\xi}C$ from it. Then the tied rows of $G$ have mean $\bar{u}_G = \bar{a} + c_{\text{in}}$, and if $\bar{u}_G \neq 0$ they satisfy (P2) with $\bar{r}' \leq (\bar{r} + \bar{\xi})\, C/\|\bar{a} + c_{\text{in}}\|_2$, which reduces to $\bar{r}' \leq \bar{r} + \bar{\xi}$ when $c_{\text{in}} = 0$.*

*Proof.* Item 1 averages $\Delta u_i = \Delta u_i^{\text{out}} + \Delta u_i^{\text{in}}$ over rows. The output-path gradients still sum to zero at every step (Proposition 1), so under a row-shared preconditioner $\sum_i \Delta u_i^{\text{out}} = 0$. For item 2, averaging $u_i = (u_i^{(0)} + \Delta u_i^{\text{out}}) + \Delta u_i^{\text{in}}$ over $G$ gives $\bar{u}_G = \bar{a} + c_{\text{in}}$, and the triangle inequality for the root mean square over $G$, applied to $u_i - \bar{u}_G = (u_i^{(0)} + \Delta u_i^{\text{out}} - \bar{a}) + (\Delta u_i^{\text{in}} - c_{\text{in}})$, bounds the root-mean-square deviation of the tied rows by $(\bar{r} + \bar{\xi})C$. □

Tying can therefore move the mean row even under a row-shared optimizer, but only along the mean input-path update, and it preserves (P2) when the input-path updates of the rows of $G$ are nearly common. In Section G.6 the first effect is small and unaligned: tied momentum SGD moves the mean row by 0.3–4% of the average row length, with a cosine between $-0.13$ and $+0.18$ to $-\bar{x}$, so tying does not substitute for a row-dependent preconditioner. Under Adam, tying lowers the alignment of the movement of the mean row with $-\bar{x}$ from 0.83–0.85 to 0.69–0.70. The tied model of Section G.5 keeps the partition of its untied counterpart at a lower effective dimension, and in Section G.6 tying lowers $\bar{r}$ on the rare rows from 1.84–1.85 to 1.54 under Adam and AdamW. By Proposition 6(2), a largely common input-path update would produce this: it adds $c_{\text{in}}$ to the mean of the rows of $G$, and when $c_{\text{in}}$ points along $\bar{a}$, the bound on $\bar{r}'$ falls below $\bar{r}$ exactly when $\|c_{\text{in}}\|_2 > \bar{\xi}C/\bar{r}$. We have not measured the input-path update directly.

The input path can move the rows without changing any prediction, because a component common to all rows shifts every logit by the same amount. To make this precise, write an output-side reference matrix as $W = [w_1, \ldots, w_V]^\top$ and the additional tied-path contribution as $\Delta W = [\Delta w_1, \ldots, \Delta w_V]^\top$, so the tied rows are $u_i = w_i + \Delta w_i$. For one training trajectory, $\Delta w_i$ can be the cumulative input-side update. When comparing separately trained tied and untied models, $\Delta W = U_{\text{tied}} - U_{\text{untied}}$ also includes changes induced in the rest of the trajectory.

**Proposition 7** (Exact and approximate softmax stability). *For a fixed state $x$, let $p^W(x)$ and $p^U(x)$ be the softmax distributions under $W$ and $U = W + \Delta W$. They are identical if and only if there is a scalar $c_x$ such that*

$$\Delta w_i^\top x = c_x \qquad \text{for every } i.$$

*In particular, a common row perturbation $\Delta w_i = c$ leaves softmax unchanged for every $x$. More generally, if $\max_i \|\Delta w_i\|_2 \leq \delta$, then*

$$\max_i |\log p_i^U(x) - \log p_i^W(x)| \leq 2\delta\|x\|_2.$$

*If the tied input path also changes the hidden state to $x' = x + \delta_x$ and $\max_i \|w_i\|_2 \leq R$, then*

$$\max_i |\log p_i^U(x') - \log p_i^W(x)| \leq 2\big[\delta\|x\|_2 + (R + \delta)\|\delta_x\|_2\big].$$

*Proof.* Softmax ratios determine the distribution. Equality of all ratios is equivalent to $(\Delta w_i - \Delta w_j)^\top x = 0$ for every $i, j$, which is exactly the common-scalar condition. For the approximate statement, set $\lambda_i = \Delta w_i^\top x$ and define $\text{LSE}(a) = \log \sum_j e^{a_j}$. Then

$$\log p_i^U(x) - \log p_i^W(x) = \lambda_i - \big[\text{LSE}(z + \lambda) - \text{LSE}(z)\big].$$

The bracket lies between $\min_j \lambda_j$ and $\max_j \lambda_j$, so its absolute value after subtracting $\lambda_i$ is at most $2\|\lambda\|_\infty \leq 2\delta\|x\|_2$. If the state also changes, each logit perturbation is

$$(w_i + \Delta w_i)^\top (x + \delta_x) - w_i^\top x = \Delta w_i^\top x + w_i^\top \delta_x + \Delta w_i^\top \delta_x,$$

whose magnitude is at most $\delta\|x\|_2 + (R + \delta)\|\delta_x\|_2$. Applying the same log-sum-exp bound proves the final inequality. □

This proposition separates two effects that weight tying otherwise conflates. A common component of the tied-path update can substantially move the stored rows while cancelling exactly from the output distribution; row-differential components change logits, and changes to the hidden state add a second source of output variation. The tied/untied controls therefore test whether the frequency geometry survives the full training perturbation, not whether the two parameterizations must produce identical rows or predictions.

### A.9 Exact identity for directional retention

The additional geometric analysis in Section G.9 measures how far each learned row rotates away from its initialization or reference checkpoint. It is not a direct consequence of Theorem 1. Because cosine retention is invariant to positive radial rescaling, only the component of the accumulated update transverse to the reference row can reduce it. The following identity isolates that component without presupposing a frequency-group ordering.

**Lemma 5** (Exact directional-retention formula)**.** *Let $u_i^{(0)} \neq 0$, let $\Delta_i = u_i^{(K)} - u_i^{(0)}$, and define*

$$s_i = 1 + \frac{\langle u_i^{(0)}, \Delta_i \rangle}{\|u_i^{(0)}\|_2^2}, \qquad b_i = \frac{\|P_i^{\perp} \Delta_i\|_2}{\|u_i^{(0)}\|_2},$$

*where*

$$P_i^{\perp} = I - \frac{u_i^{(0)} (u_i^{(0)})^{\top}}{\|u_i^{(0)}\|_2^2}.$$

*If $s_i > 0$, then*

$$\mathrm{Ret}_i := \cos(u_i^{(K)}, u_i^{(0)}) = \frac{s_i}{\sqrt{s_i^2 + b_i^2}} = \frac{1}{\sqrt{1 + \vartheta_i^2}}, \qquad \vartheta_i = \frac{b_i}{s_i}.$$

*Proof.* Orthogonally decompose the final row as

$$u_i^{(K)} = s_i u_i^{(0)} + P_i^{\perp} \Delta_i.$$

The inner product with $u_i^{(0)}$ is $s_i \|u_i^{(0)}\|_2^2$, while the squared norm is $\|u_i^{(0)}\|_2^2 (s_i^2 + b_i^2)$. Substitution into the cosine gives the stated identity. □

By Lemma 2, an untied output row has $\Delta_i = \mathcal{A}_i - \mathcal{R}_i$ with accumulated attraction and repulsion

$$\mathcal{A}_i = \sum_{t<K} \eta_t D_{i,t}\, \pi_i \mu_i^{(t)}, \qquad \mathcal{R}_i = \sum_{t<K} \eta_t D_{i,t}\, \mathbb{E}[p_i^{(t)}(x) x^{(t)}];$$

in a tied model $\mathcal{R}_i$ also absorbs the accumulated input-path update, with the opposite sign.

Substitution into Lemma 5 gives the exact CE decomposition

$$\vartheta_i = \frac{\|P_i^{\perp}(\mathcal{A}_i - \mathcal{R}_i)\|_2}{s_i \|u_i^{(0)}\|_2}, \qquad \mathrm{Ret}_i = \frac{1}{\sqrt{1 + \vartheta_i^2}}.$$

Section G.9 finds that frequent rows retain more of their initial direction than non-frequent rows, in both the synthetic model and Pythia-1B; by this identity, that ordering is exactly an ordering of the normalized net transverse updates. Frequency scales $\mathcal{A}_i$, but it does not by itself determine the direction of $\mathcal{A}_i - \mathcal{R}_i$ relative to the initialization. In particular, an ordering of frequent and non-frequent retention would require an additional assumption about the joint geometry of initialization, conditional target directions, and repulsion. Rather than encode the observed ordering in such an assumption, we use the identity above to interpret the empirical training trajectories in Section G.9.

## B Implementation and evaluation details

### B.1 Language models

We use the Hugging Face causal-language-model implementations listed in Table 3. Hidden states are taken after the model's final normalization at the last prompt token; logits are computed with the model's own LM head. Llama3-70B is loaded in 4-bit NF4 for memory efficiency; other models use bfloat16. All reported geometry is computed in float32 after extraction.

Table 3: Evaluated language models and architectural sizes used by the head analyses.

| Model | Layers | Hidden size | Attention heads (all layers) |
|---|---|---|---|
| Llama3.2-1B | 16 | 2048 | 512 |
| Llama3.2-3B | 28 | 3072 | 672 |
| Llama3-8B | 32 | 4096 | 1024 |
| Llama3.1-8B | 32 | 4096 | 1024 |
| Llama3-70B | 80 | 8192 | 5120 |
| GLM4-9B | 40 | 4096 | 1280 |
| GLM4-32B | 61 | 5120 | 3904 |

### B.2 Datasets and prompt pools

We use SST-2, SNLI, TREC, RTE, MR, SUBJ, and MNLI. Demonstrations are sampled from each task's training pool and queries from its validation/test pool according to the dataset loader. The cached evaluation counts are 436, 494, 250, 138, 500, 500, and 500, respectively. We do not use CB: its validation split yields only 28 evaluation prompts, too few to keep the prompts that select induction heads apart from those used for held-out evaluation. The same saved demonstration-index matrix is reused across experiments for a fixed model, dataset, and shot count. This pairing is important: ZSL, natural-label ICL, symbol-label ICL, and ablated ICL differ only in the intended intervention.

Unless a panel explicitly displays individual datasets, its LLM curve first averages prompts within a dataset and then averages the seven dataset values. This prevents the larger MR/SUBJ/MNLI pools from dominating RTE or TREC. Synthetic results average held-out prompts directly because there is one synthetic task distribution.

## C Prompt templates and label types

### C.1 Natural-label templates

Table 4: Prompt templates and natural single-token labels. A period and space separate demonstration records; the query ends immediately after the final delimiter.

| Dataset | Record template | Labels |
|---|---|---|
| SST-2 | `{sentence} Sentiment: {label}` | negative / positive |
| MR | `{sentence} Sentiment: {label}` | negative / positive |
| SUBJ | `Text: {sentence} Type: {label}` | objective / subjective |
| TREC | `Question: {sentence} Type: {label}` | abbreviation / entity / description / person / number / location |
| RTE | `{premise} The question is: {hypothesis} True or false? The answer is: {label}` | true / false |
| SNLI | `{premise} The question is: {hypothesis}? True or maybe or false? The answer is: {label}` | true / maybe / false |
| MNLI | same NLI format as SNLI | true / maybe / false |

A ten-shot prompt concatenates ten records followed by the query record without its label. ZSL is the same query record without the demonstration prefix. Tokenizer-specific leading spaces are included when locating label-token IDs, and runtime checks require every candidate label to map to exactly one vocabulary row.

### C.2 Symbol remapping

For each model we use one shared ordered list of six one-token symbol strings. A task with $C$ classes uses the first $C$ strings. The class identity is unchanged: class $j$ in every demonstration and in evaluation is mapped to symbol token $j$. Queries, demonstration examples, demonstration order, and saved demonstration indices are identical between the natural and symbol conditions. The candidate-label unembeddings are the only intended change.

## D Synthetic model and data generation

In this section, we provide further details of the synthetic model's training-data generation, training setup, and evaluation.

Table 5: Natural-to-symbol label mappings used in the remapping experiments. The leading whitespace included in each symbol string for tokenization is omitted in the display.

| Dataset | Label mapping |
|---|---|
| SST-2 | negative / positive → ∀ / ← |
| MR | negative / positive → ∀ / ← |
| SUBJ | objective / subjective → ∀ / ← |
| TREC | abbreviation / entity / description / person / number / location → ∀ / ← / ↔ / ↓ / ★ / ※ |
| RTE | true / false → ∀ / ← |
| SNLI | true / maybe / false → ∀ / ← / ↔ |
| MNLI | true / maybe / false → ∀ / ← / ↔ |

### D.1 Vocabulary and data generator

The non-ICL language-modeling sequences use a vocabulary of 30 major tokens and 3,990 minor tokens. Their empirical next-token counts define the synthetic frequency ranking used in the geometry experiments, directly estimating the target marginal $\pi_i$ of the dominant language-modeling component of the training corpus. The major tokens are

```
~ ` ! $ % ^ & * ( ) - _ = + [ { ] } " \ ; ' , < . > / ? @ #
```

and the minor tokens are named `T0`, ..., `T3989`. The tokens `:` and `|` delimit fields and demonstrations and are used only in ICL-format sequences. The following is the first 20-token excerpt of one generated length-128 language-modeling sequence:

```
/ / / / T1035 / / / / / / / / / T3638 T1285 / T803 /
```

The repeated major token reflects its high rank under the sampled topic, while the minor tokens are lower-ranked members drawn from the same topic-specific full-vocabulary distribution. The construction and training corpus are defined in Section 4.2; here we provide concrete instances sufficient to reproduce the serialization.

The default model uses learned absolute positional embeddings added to token embeddings. Each Transformer block is pre-normalized with RMSNorm ($\epsilon = 10^{-5}$): one RMSNorm precedes the causal self-attention sublayer and a second precedes the MLP sublayer, followed by a final RMSNorm before unembedding. The MLP uses a GELU activation between bias-free linear maps. Input and output embeddings are tied in the main configuration. We optimize for 10,000 steps with batch size 64 using AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, weight decay 0.01, and a fixed learning rate of $3 \times 10^{-4}$ without scheduling. Token embeddings are initialized from $\mathcal{N}(0, 0.02^2)$, positional embeddings are initialized to zero, and the maximum sequence length is 128.

### D.2 ICL prompt generation

ICL prompts draw their labels from a shared pool of 30 tokens, `T0`–`T29`. Each demonstration picks one of the prompt's two clusters uniformly at random and draws its input token, with replacement, from a fixed set of 20 minor tokens of that cluster; the query repeats one of the demonstrated inputs. One generated 10-shot prompt is

```
T3808 : T25 | T30 : T25 | T668 : T12 | T835 : T12 |
T3963 : T12 | T3430 : T12 | T3675 : T12 | T3264 : T25 |
T2696 : T12 | T681 : T12 | T3430 :
```

All unordered cluster pairs are shuffled with seed 2026; the first 20 are held out for evaluation and the next 100 generate training ICL strings. From the 20 held-out pairs, we generate 1,000 10-shot prompts for evaluation.

For the shot sweep, we first replace each generated query by the input token from its first demonstration and use that demonstration's label as the answer. We then construct the $k$-shot condition by retaining the first $k$ complete demonstrations from the same 10-shot prompt while keeping this query–answer pair fixed. Because every condition includes the first demonstration, the query has an in-context precedent at every evaluated shot count $k = 1, \ldots, 10$; only the number of additional demonstrations changes.

### D.3 Natural and symbol labels

The natural synthetic condition uses the two randomly sampled label tokens already present in a prompt. The symbol condition replaces only the demonstration-label slots by the fixed pair `T3988`/`T3989`; the cluster assignment, demonstrations, order, and query are unchanged. These tokens are valid vocabulary rows but lie outside the prompt's sampled natural pair.

### D.4 Training perturbations

We evaluate four orthogonal controls:

- **Zipf, no topic:** keeps the Zipf marginal but removes topic-conditioned sampling by pooling the tokens of all language-modeling sequences and redistributing them at random, so that each sequence is an i.i.d. sample from the marginal; the ICL training strings and held-out prompts are unchanged.
- **Topic + uniform:** retains latent topics but samples tokens uniformly within them.
- **IID uniform:** removes Zipf frequency skew and topic-conditioned sampling.
- **Untied:** uses separate input and output embedding matrices.

The ICL accuracy of Tables 1 and 2 is the full-vocabulary top-1 accuracy on the second half of the 1,000 held-out 10-shot prompts, the 500 prompts that are not used to rank induction heads; the IID-uniform model is evaluated on its own held-out prompts, which its generator draws with a different seed.

These interventions distinguish frequency, conditional structure, and tying. The accumulated-update identity in Section A.1 is stated for the output path; tying adds the input-path gradient discussed in Section A.8, which the untied control removes.

## E Induction-head selection and interventions

### E.1 Induction-head scoring and prompt split

For a prompt $q$ with final query-delimiter position $t_q$, let $\mathcal{J}_q$ denote all demonstration label-token positions. For head $h = (l, k)$, its dataset-specific IH score is

$$s_h = \frac{1}{|Q|} \sum_{q \in Q} \sum_{j \in \mathcal{J}_q} \mathrm{Attn}_q^{(h)}[t_q, j]. \tag{23}$$

Heads are ranked by $s_h$. For the synthetic model, the 1,000 held-out prompts are split in half: the first 500 rank heads, and the second 500 are shared by the shot and ablation sweeps. At every ablation dose, top-ranked heads are compared with ten random sets of equal cardinality. This makes synthetic head identification and causal evaluation disjoint. For the LLMs, heads are ranked on the first 50 queries of each dataset, each with its own randomly drawn demonstrations. The ablation and attribution metrics are computed on the first half of each dataset's prompts, which contains these 50 queries, and IH steering is evaluated on the second half. The LLM ablation and attribution results should therefore be read as causal intervention on the operationally selected dataset-specific heads, not as out-of-sample head-ranking generalization.

### E.2 Final-normalization scaling for DLA

DLA maps a residual-stream component through the same final-normalization scale as the prompt endpoint while keeping that scale fixed during attribution. Let $h_{n,p}^{L_{\mathrm{lay}}} \in \mathbb{R}^d$ be the pre-normalization final state for prompt $p$, let $w_{\mathrm{final}}$ be the learned final-normalization weight, and define

$$r_p = \sqrt{\frac{1}{d} \sum_{j=1}^{d} (h_{n,p,j}^{L_{\mathrm{lay}}})^2}, \qquad \mathrm{Norm}_p(v) = \frac{w_{\mathrm{final}} \odot v}{r_p}.$$

For the paired ICL–ZSL head-output difference, the implementation uses the ICL endpoint to compute $r_p$ and applies this same map to every selected head. Because $\mathrm{Norm}_p$ is fixed and linear during attribution,

$$\sum_{(l,k)\in\mathcal{I}} \mathrm{Norm}_p(\Delta a^l_{n,k,p}) = \mathrm{Norm}_p\left(\sum_{(l,k)\in\mathcal{I}} \Delta a^l_{n,k,p}\right).$$

Projecting this normalized component through $U$ gives its vocabulary-logit DLA. This operation attributes components under the endpoint scale; it does not independently rerun the nonlinear normalization on each head output.

### E.3 Head ablation

For a selected head, its output contribution at each token is set to zero before the residual addition. The model then completes the remaining layers normally. Doses from $1\%$ through $10\%$ use nested prefixes of the IH ranking. Random head sets have the same size and are independently drawn for ten repeats at every dose.

### E.4 Total-variation distance and ZSL recovery

For candidate-label probabilities we append the remaining vocabulary as one bucket,

$$p_{\bar{L}} = 1 - \sum_{\ell\in L} p_\ell,$$

and compute total variation on the resulting $|L|+1$ distribution $P = (p_L, p_{\bar{L}})$ of Section 4.4. For two such distributions $P$ and $Q$,

$$D_{\mathrm{TV}}(P,Q) = \frac{1}{2}\sum_{c\in L\cup\{\bar{L}\}} |P_c - Q_c|.$$

If $P_{\mathrm{abl}}$ is the ablated distribution, its recovery toward ZSL relative to the original ICL–ZSL gap is

$$R_{\mathrm{ZSL}} = 1 - \frac{D_{\mathrm{TV}}(P_{\mathrm{abl}}, P_{\mathrm{ZSL}})}{D_{\mathrm{TV}}(P_{\mathrm{ICL}}, P_{\mathrm{ZSL}})}.$$

Prompts whose ICL–ZSL denominator is below $10^{-6}$ are excluded from this ratio but retained for raw accuracy and alignment.

## F Supplementary Materials for Subsection 4.1

### F.1 Replication of Figure 2 for Other Models

Figures 10–15 repeat the paired ICL–ZSL alignment experiment from Figure 2 on Llama3.2-1B/3B, Llama3-8B/70B, and GLM4-9B/32B. The Llama models consistently reproduce the positive vocabulary-wide tendency, and GLM4-32B retains it more weakly. GLM4-9B instead reverses the aggregate direction, although its task panels remain heterogeneous. The frequency-organized geometry is therefore more consistent across architectures than the sign of its interaction with each model's ICL-induced state movement, as summarized jointly in Table 8.

### F.2 Robustness of the Vocabulary-Wide Alignment Shift to the Number of Demonstrations

To test whether the vocabulary-wide signature strengthens with the amount of in-context evidence, we sweep from one to ten demonstrations while keeping the query and evaluation construction fixed. The structured ICL prompts are compared with matched random-prefix ZSL controls. For each shot count, the line bands bootstrap prompts with replacement for 1,000 replicates and show the 2.5th and 97.5th percentiles.

Figures 16–22 report Llama3.1-8B and the six additional LLMs, each alongside the synthetic model. As demonstrations are added, mean unembedding alignment and the ICL unembedding win rate

generally increase, whereas matched random prefixes do not reproduce the same separation. The shot dependence therefore supports the interpretation that the vocabulary-wide shift is induced by structured ICL rather than prompt length alone.

### F.3 Token-Frequency Estimation for Figure 3 and Its Replications

In the theory, token frequency is the true next-token target marginal $\pi_v = \Pr_{\mathcal{D}_{\text{train}}}(y = v)$, which is also the coefficient of the target-attraction term in Equation (5). Because the evaluated LLMs' original pretraining mixtures are unavailable, the frequency analyses estimate this ranking independently for each tokenizer using the training split of WikiText-103 (Merity et al., 2017). We discard empty documents, tokenize the remaining text in batches of 128 with the model's native tokenizer, and remove padding through the attention mask. We also remove the tokenizer-specific initial special-token prefix: one token for the Llama models, two for GLM4-9B, and none for GLM4-32B. Every remaining token ID contributes one count $c_v$ to its corresponding vocabulary row.

Tokens are ranked by $c_v$ to define the top-frequency sets. For log-frequency analyses, tokens observed in WikiText-103 use $f_v = \log c_v$. A vocabulary row with zero observed occurrences is assigned the smallest positive count observed for that tokenizer before taking the logarithm, avoiding an undefined value while applying the most conservative frequency floor. We compute an SVD $U = Q\Sigma V^\top$ of the unembedding, assign each row its leading coordinate $u_v^\top \psi_1$, and report the Pearson correlation between this coordinate and $f_v$. Because the sign of a singular vector is arbitrary, $\psi_1$ is oriented so that this correlation is positive. Normalizing counts into empirical probabilities would subtract the same constant $\log \sum_w c_w$ from every $f_v$ and therefore would not change the Pearson correlation.

### F.4 Replication of the Leading-Direction Result in Figure 3

Figures 23–28 replicate the right panel of Figure 3, comparing empirical log token frequency with the leading unembedding singular coordinate. Across both Llama and GLM families, higher-frequency tokens occupy systematically different coordinates along this direction. The slopes and dispersion vary with model scale, but the recurring association supports the claim that the dominant unembedding geometry is frequency organized rather than peculiar to Llama3.1-8B.

### F.5 Replication of the Bipartition and Spectrum in Figure 3

Figures 29–34 replicate the left and center panels of Figure 3 for the remaining models. The frequency-based mean partition remains visible across architectures, while the spectra concentrate substantially before the ambient dimension is exhausted. Together with the leading-direction results above, these replications establish that both the row-group separation and low-dimensional concentration are shared properties of the evaluated unembeddings.

### F.6 Robustness Checks for the Mean-Unembedding Geometry

Figure 3 uses the learned unembedding in its original parameterization: its left panel partitions rows by their projection onto the raw row mean, its center panel uses the singular spectrum of $U$, and its right panel relates the leading singular coordinate to token probability. We conduct two controls to determine whether these observations reflect a stable frequency–geometry structure or instead arise from a removable common offset or heterogeneous row norms.

**Does the frequency-bearing singular direction survive centering?** The first singular direction of an uncentered matrix can partly reflect its common row mean. To test whether the association in Figure 3c is reducible to that mean, we subtract the vocabulary-row centroid from every $u_i$, recompute the leading singular direction of the centered matrix, project every valid-frequency token onto it, and correlate the resulting coordinates with log token frequency. Because a singular vector has arbitrary sign, Table 6 reports the absolute Pearson correlation. The correlation remains substantial in every model ($|r| = 0.495$–$0.683$), showing that frequency remains encoded along the dominant direction after the common offset is removed.

**Is the mean direction an artifact of row norms?** Mean alignment uses cosine similarity and hence unit-normalized token unembeddings, whereas the raw bipartition begins from their ordinary mean. To test whether a few large-norm token unembeddings rotate these two constructions toward

unrelated directions, we compare the mean before and after individual normalization. Table 7 gives cosines of 0.9995–1.0000 across models. Thus normalization changes magnitude but essentially not direction, ruling out heterogeneous token-unembedding norms as the source of the observed mean axis. Taken together, the two controls show that the frequency signal survives centering and normalization; task-relative contrasts in Section 4.3 are additionally gauge invariant by construction.

# G Supplementary Materials for subsection 4.2

## G.1 Robustness of the Synthetic Unembedding Geometry to Model Width and Depth

The main synthetic model uses dimension 256 and six Transformer layers. Because the frequency-linked partition or spectral concentration could depend on this particular capacity, we retrain the model across dimensions $\{128, 256, 512\}$ and depths from three through nine layers. The data generator, training objective, and frequency-group definition remain fixed. For every trained model, we measure normalized effective dimension, the fraction of rows in the raw-mean outlier group, and the overlap between that group and an equal-size set of the most frequent tokens.

Figure 35 shows that the qualitative geometry persists throughout the grid. Outlier–frequency overlap remains between 0.825 and 0.962, whereas the corresponding random-set overlap is only the outlier fraction, 0.016–0.032. Normalized effective dimension remains low (0.025–0.080), showing spectral concentration across widths and depths despite capacity-dependent variation in its magnitude. The frequency-linked partition and low-dimensional structure are therefore not artifacts of the default width-256, six-layer architecture.

## G.2 Replication of Table 1 for Other Models

The main comparison combines three properties of the learned unembedding, ICL accuracy, and two properties of its paired ICL–ZSL readout. To test whether this combination is restricted to Llama3.1-8B, we compute the same quantities for every additional model. Geometry statistics use the full vocabulary, with the frequent set in each model size-matched to its raw-mean outlier set. LLM accuracy, ICL win rate, and $\Delta A$ first average prompts within each task and then average the seven task values. Accuracy uses the natural-label condition and, as in Section 4.4, the held-out second half of each task's evaluation queries; alignment quantities match Figure 2. In the synthetic model, accuracy, win rate and $\Delta A$ all use the second half of the 1,000 held-out prompts, the 500 prompts that are not used to rank the heads.

Table 8 shows that, across all additional models, the raw-mean partition isolates only a small fraction of vocabulary rows, and that group overlaps with a size-matched frequent set far more strongly than its vocabulary prevalence would predict. The spectra are also concentrated relative to the ambient dimension, although the degree varies with architecture and scale. Every model exhibits functional ICL accuracy, but accuracy alone does not determine the direction of the vocabulary-wide readout. The Llama family reproduces the positive alignment shift and high rowwise win rate seen in the synthetic model and Llama3.1-8B. GLM4-32B shows the same readout direction more weakly, whereas GLM4-9B is the clear exception despite its high task accuracy and retained frequency-linked bipartition. Thus the unembedding geometry generalizes more uniformly than the sign and magnitude of its interaction with a particular ICL-induced state-direction change.

## G.3 Empirical Verification of Corollary 1

Corollary 1 asks whether a coherent non-frequent majority outweighs the aggregate cancellation available to the remaining token unembeddings. We normalize every token unembedding, take $F$ to be the top $|\mathcal{O}_{\text{raw}}|$ tokens by empirical frequency, and set $N = \mathcal{V} \setminus F$. For each exception rate $\phi$, we use the normalized centroid of $N$ as a selection axis, assign the lowest-projecting $\phi|N|$ rows to $E$, and let $G = N \setminus E$. We compute each group's average off-diagonal cosine $\rho_Q$ and evaluate $B_{\text{coh}} = (s_G - s_E - s_F)/V$, where $s_Q = \sqrt{|Q|[1 + (|Q| - 1)\rho_Q]}$. This is the corollary's aggregate endpoint condition, which retains the cancellation exhibited within $E$ and $F$.

Table 9 shows that the coherent term dominates both cancellation terms by a wide margin. The resulting bound is positive and close to the observed projection in the synthetic model and both Llama3.2 models. Sweeping $\phi$ from 0 to 0.05 leaves the bound positive in all three systems. This verifies the corollary's aggregate endpoint condition; it does not reconstruct the accumulated attraction and repulsion moments along training.

Figures 36–43 repeat the coherence-bound sweep for the synthetic model and all seven LLMs. The synthetic model gives the largest certificate: $B_{\mathrm{coh}}$ decreases only from $0.5019$ at $\phi = 0$ to $0.4853$ at $\phi = 0.05$, while its observed projection remains approximately $0.5032$. Removing the lowest-projecting rows raises the coherence of the retained set $G$, but the growing norm of the exception set makes the lower bound progressively more conservative. Across the LLMs, $B_{\mathrm{coh}}$ ranges from $0.1119$ to $0.3751$ at $\phi = 0.005$ and remains positive ($0.1092$–$0.3613$) at $\phi = 0.05$. Thus all eight learned unembeddings satisfy the corollary's aggregate condition throughout the sweep. This establishes formation of a nonzero mean-unembedding axis; the sign of a task's prompt-time $\Delta A$ remains a separate property of its ICL-induced state movement.

### G.4 Additional Training-Dynamics Diagnostics for Figure 4

The main training sweep uses effective dimension and outlier–frequency overlap to locate when frequency organization emerges. Because both quantities compress the row distribution into a spectral or thresholded summary, we additionally ask whether the same transition is visible directly in pairwise row geometry. For every checkpoint and frequency group, we compute cohesion from within-group normalized-row similarities and compactness from dispersion around the group centroid. We repeat this procedure for the synthetic model and Pythia-1B, using an independent vertical scale for each system so that the temporal ordering rather than cross-model magnitude is compared.

Figure 44 shows that frequent and non-frequent cohesion and compactness begin to separate as training proceeds, alongside the transition in Figure 4. Thus the reported dynamics are not specific to an outlier threshold or participation-rank statistic: they are also present in the normalized pairwise relations and centroid dispersion from which the group geometry is built. The agreement across complementary metrics strengthens the interpretation of a genuine row-set reorganization during training.

### G.5 Additional Controls for Table 2

Table 2 varies the data distribution, but the synthetic model also shares its input and output embedding weights by default. Weight tying could in principle create row geometry through the input pathway rather than through the output-side cross-entropy forces analyzed in Equation (5). We therefore train an otherwise identical model with independent embedding and unembedding matrices. Figure 45 shows that tied and untied models are nearly indistinguishable on the principal frequency-geometry diagnostics. The main difference is raw effective rank, which increases from $8.102$ with tying to $14.706$ without it. Even the latter remains very small relative to the model dimension $d = 256$ ($\mathrm{EDim}(U)/d = 0.057$), retaining clear spectral concentration. Weight tying therefore modulates the magnitude but is not required for the effect in this controlled setting. Section A.8 shows how the input path enters premises (P1) and (P2); it bounds the effect of tying rather than certifying these two trained models.

### G.6 Which Optimizer Satisfies the Premises of Theorem 1

Theorem 1 rests on two premises: (P1), that the optimizer moves the mean row along $-\bar{x}$ and that the rest of the mean row, its initial value and its movement off $-\bar{x}$, does not reverse the side of the partition that this movement assigns to a row, and (P2), that the non-frequent rows cluster around their mean tightly enough, measured by their root-mean-square deviation $\bar{r}$ from it, for the bound $(1 + \bar{r}^2)^2/(d\omega_{\mathrm{g}}^2)$ of Theorem 1 to fall well below one. Neither is a property of cross-entropy: Proposition 1 shows that (P1) fails identically for an untied matrix trained by gradient descent or momentum SGD, since the row gradients sum to zero, and (P2) concerns how the optimizer transports the shared repulsion into the rows. Section A.4 predicts which optimizers satisfy which premise: momentum SGD satisfies neither, signSGD satisfies (P2) most tightly but does not disperse the frequent rows, and Adam-family optimizers satisfy both. What is still missing is a measurement

of both premises across optimizers, together with a check that the geometry of Section 4.2 appears when they hold. This subsection supplies both and tests these predictions.

**Runs** We retrain the synthetic model of Section D with every non-optimizer factor held fixed: the same Zipf–topic corpus and ICL training strings, the same width-256 six-layer architecture, batch size 64, and 10,000 steps. Within a seed, all runs share the initialization and the batch order, so two runs differ only in the update rule. We compare AdamW in the configuration of the main experiments (learning rate $3 \times 10^{-4}$, weight decay $0.01$), Adam with weight decay $0$ so that the comparison with SGD is not confounded by decay, momentum SGD ($\beta = 0.9$, no weight decay; SGD below) at learning rates $0.03$, $0.1$ and $0.3$, since the main configuration fixes no SGD rate, so that a failure under SGD cannot be attributed to one poorly chosen rate, signSGD with $u \leftarrow u - \eta\,\mathrm{sign}(\nabla u)$ at $\eta = 10^{-4}$, and a hybrid that applies Adam to every parameter except the output matrix and momentum SGD to the output matrix, which separates Adam acting on $U$ from Adam shaping the hidden states. AdamW, Adam and SGD are run with tied and untied output matrices, and the untied Adam run and the untied SGD run at learning rate $0.1$ are repeated with seeds 2026, 7 and 13.

**Measuring** (P1) We measure the relative movement of the mean row, $\|\bar{u}^{(K)} - \bar{u}^{(0)}\|_2$ divided by the average length $\frac{1}{V}\sum_i \|u_i\|_2$ of the rows, and its alignment $\cos(\bar{u}^{(K)} - \bar{u}^{(0)}, -\bar{x})$ with the mean state $\bar{x}$, computed on held-out synthetic language-modeling sequences drawn from the training generator after the training strings. A positive cosine means that the mean row moves along $-\bar{x}$, and the smaller the cosine, the longer its movement off $-\bar{x}$ relative to its movement along $-\bar{x}$. We do not measure the row-wise condition of (P1) separately: by Proposition 4, when the cosine is positive it fails exactly on the rows that point toward $\bar{x}$ but lie outside $\mathcal{O}_{\mathrm{raw}}$ or point away from $\bar{x}$ but lie inside it, so it is tested through the partition itself.

**Measuring** (P2) We take $G$ to be the rare rows, $\pi_i < 1/V$ under the training target marginal, and measure their root-mean-square deviation $\bar{r}$ from their mean row, relative to the norm of that mean, and their share $\omega_{\mathrm{g}}$ of the squared row norms. At the final step we compare the bound $(1+\bar{r}^2)^2/(d\omega_{\mathrm{g}}^2)$ of Theorem 1 with the measured $\mathrm{EDim}(U)/d$. At step 100 we test instead the mechanism by which Proposition 2(1) derives (P2). The proposition writes each rare row as the average shared repulsion $c$, of norm $C$, plus a remainder $w_i$ that collects its initialization, its attraction and the row-to-row differences of the repulsion, and bounds $\bar{r} \le \bar{\epsilon}/(C - \|\bar{w}\|_2)$, where $\bar{\epsilon}$ is the root mean square of $\|w_i\|_2$ and $\bar{w}$ their average. When the updates of the rare rows are nearly common, the remainders reduce to the initializations, which average out over the many rows of $G$, and then $\bar{r} \approx \bar{\epsilon}_0/C$, with $\bar{\epsilon}_0$ the root mean square of $\|u_i^{(0)}\|_2$ over $G$. We compare $\bar{r}$ with this initialization-to-update ratio, estimated as $\|U_G^{(0)}\|_{\mathrm{F}}/\|U_G - U_G^{(0)}\|_{\mathrm{F}}$ over the rows of $G$, which equals $\bar{\epsilon}_0/C$ when the updates are exactly common.

**Geometry and checks** We measure the geometry that Theorem 1 predicts from the two premises, normalized effective dimension and the raw-mean outlier rate with its overlap with the size-matched frequent set, and the two properties of the unit-mean axis described by Corollary 1 and Proposition 5, its norm $\|m_{\mathrm{unit}}\|$ and the Spearman correlation between the depth $\delta_i = -v_i^\top \widehat{m}$ and log target frequency. Unlike the label depth of Section 4.3, the cosine of a label's contrast direction $u_\ell - \bar{u}_{\bar{L}}$ with $m_{\mathrm{unit}}$, $\delta_i$ is computed from each token's own normalized row and grows toward the frequent side.Because Section A.4 predicts that signSGD, unlike the Adam family, also moves the remaining rows, $\pi_i \ge 1/V$, along nearly one common direction, we also measure their residual coherence, the average cosine between distinct frequent rows after their normalized directions are projected off the unit-mean axis, so that the component that every row has along that axis is not counted. Held-out loss and ICL accuracy on the $1{,}000$ held-out prompts check that every run trains and performs ICL, and $\Delta\mathcal{A}$ with its label and orthogonal parts gives the ICL signature read along the axis. At every checkpoint, the final states used for these measurements reproduce the model's logits through the output matrix.

**Overview** Table 10 lists the premise measurements and training checks of every run, and Table 11 the resulting geometry and alignment shift. The two premises sort the optimizers into three of the four possible cases: Proposition 1 predicts that (P1) fails for every optimizer that rescales all rows of an untied output matrix alike, and Section A.4 predicts that (P2) orders them by how each turns the shared repulsion into a step. The Adam family satisfies both premises, signSGD only (P2), and momentum SGD and the hybrid neither; none of the optimizers we test satisfies (P1) alone.

**Adam family: both premises** Both premises hold under the Adam family. For (P1), Adam moves the mean row by a relative 0.455–0.457 across the three seeds, with cosine 0.829–0.840 to $-\bar{x}$, and AdamW by 0.457, with cosine 0.845; tying lowers this alignment to 0.69–0.70. For (P2), the rare rows deviate from their mean by $\bar{r} = 1.84$–$1.85$ in all three seeds of Adam and in AdamW, which bounds $\mathrm{EDim}(U)/d$ by 0.082–0.083 against a measured 0.058–0.060; tying lowers $\bar{r}$ to 1.54, with a bound of 0.048–0.049 against 0.030–0.031. At step 100 the rare rows deviate by $\bar{r} = 1.23$–$1.25$, close to their initialization-to-update ratio $\bar{\epsilon}_0/C \approx 1.17$–$1.18$, as Proposition 2(1) predicts when their updates are nearly common and their spread comes from the initialization.

With both premises, the Adam family forms the frequency-separated low-rank partition: across the three seeds Adam places 2.1–2.2% of the rows in $\mathcal{O}_{\mathrm{raw}}$, with overlap 0.905–0.933 against the size-matched frequent set and normalized effective dimension 0.058–0.060, and AdamW gives 2.2%, 0.910 and 0.058. Its frequent rows stay dispersed, with residual coherence 0.01, and the untied runs form a unit-mean axis with $\|m_{\mathrm{unit}}\| = 0.45$, along which depth correlates with log frequency at 0.90–0.91. $\Delta\mathcal{A}$ is $+0.031/+0.024/+0.009$ across the three seeds of Adam, with label parts $-0.018/-0.016/-0.013$ and orthogonal parts $+0.050/+0.040/+0.023$, and under AdamW $\Delta\mathcal{A} = +0.014$, with label part $-0.019$ and orthogonal part $+0.034$; both reproduce the decomposition of Figure 6.

**signSGD: (P2) only** signSGD satisfies (P2) but not (P1). It moves the mean row furthest, by a relative 0.980, but with cosine only 0.225 to $-\bar{x}$, so it fails (P1) through the direction rather than the size of the movement: its movement off $-\bar{x}$ is 4.3 times as long as its movement along $-\bar{x}$.

Part of this misalignment is a moving target rather than a misdirected movement. At step $t$ the shared repulsion pushes the rows against the mean state $\bar{x}^{(t)}$ of that step, whereas the cosine above is taken with the final mean state $\bar{x}^{(K)}$. Under untied Adam and AdamW the two nearly coincide, with $\cos(\bar{x}^{(t)}, \bar{x}^{(K)}) \geq 0.80$ from step 10 onwards, so every push already points close to $-\bar{x}^{(K)}$. Under signSGD the mean state rotates throughout training, with $\cos(\bar{x}^{(t)}, \bar{x}^{(K)}) = 0.00$ at step 100 and 0.42 at step 5,000, so the early pushes point in directions nearly orthogonal to $\bar{x}^{(K)}$: from step 50 to step 400 the accumulated movement of the mean row has cosine 0.93–0.98 with the current $-\bar{x}^{(t)}$ but within 0.01 of zero with $-\bar{x}^{(K)}$. The final movement sums these pushes, so its cosine with $-\bar{x}^{(K)}$ stays low even where each push followed the mean state of its step. Testing (P1) against the final mean state is therefore accurate only when that state is stable, which holds for untied Adam and AdamW but not for signSGD.

signSGD satisfies (P2) most tightly: the rare rows deviate from their mean by $\bar{r} = 0.26$, which bounds $\mathrm{EDim}(U)/d$ by 0.0045 against a measured 0.0040, and at step 100 by $\bar{r} = 2.56$ against an initialization-to-update ratio of 2.52. With (P2) alone, it reaches the lowest normalized effective dimension of any run, 0.004, and the strongest unit-mean axis, with $\|m_{\mathrm{unit}}\| = 0.93$, but its frequent rows also follow nearly one sign vector and stay coherent (residual coherence 0.25). Its partition captures only 0.30 of the frequent set. Relative to its own chance level of 0.0075, however, this overlap is as enriched as Adam's (0.91 against 0.022), so the partition is much smaller rather than less informative about frequency. Its label part ($-0.271$) and orthogonal part ($+0.259$) are each the largest in magnitude of any run, but they nearly cancel, leaving a small negative $\Delta\mathcal{A}$ ($-0.013$).

**Momentum SGD and the hybrid: neither premise** Momentum SGD and the hybrid satisfy neither premise. Every untied run whose output matrix is updated by momentum SGD, including the hybrid, applies one preconditioner to all rows, and its mean row moves by a relative amount of order $10^{-6}$, the level of single-precision rounding error. This matches Proposition 1: the row gradients sum to zero, so such an optimizer without weight decay leaves the mean row exactly at its initialization, and the residual movement is numerical noise whose direction carries no information.

Weight tying adds an input path whose row sum need not vanish (Section A.8), but it does not substitute for a row-dependent preconditioner: tied momentum SGD moves the mean row by a relative 0.003–0.04, but not along $-\bar{x}$, with cosine between $-0.13$ and 0.18 across its three learning rates against 0.69–0.70 for tied Adam and AdamW.

Because the step of momentum SGD is proportional to the small gradient of a rare row, the rare rows deviate from their mean by $\bar{r} = 20$–$30$ in every such run, tied or untied and including the hybrid, so the bound on $\mathrm{EDim}(U)/d$ exceeds the trivial value one by three orders of magnitude. Without either premise, the partition of the untied runs carries no frequency information: the untied momentum

SGD runs and the hybrid place 40–47% of the rows in $\mathcal{O}_{\text{raw}}$ with overlap 0.42–0.48, close to the chance level set by their own outlier rate, at every learning rate tested. Because the hybrid differs from Adam only in the update applied to $U$, this is attributable to that update rather than to differences in the hidden states. Normalized effective dimension ranges from 0.06 to 0.38 with the learning rate of momentum SGD, and at the largest rate the untied run matches untied Adam without (P2) (0.061 against 0.058–0.060). The frequency ordering of the tokens along the unit-mean axis is the one property that survives, with depth–frequency correlation 0.67–0.87, albeit on an axis an order of magnitude weaker, with $\|m_{\text{unit}}\| = 0.03$–$0.08$.

Training quality does not explain the split, since momentum SGD reaches a lower held-out loss than untied Adam (1.65–1.68 against 2.15–2.21), and ICL accuracy is 0.48–0.53 in every run except tied SGD at the largest learning rate (0.38). $\Delta\mathcal{A}$ is $+0.003/-0.004/+0.001$ across the three seeds of momentum SGD at learning rate 0.1, with inconsistent sign even after normalizing by $\|m_{\text{unit}}\|$.

**Summary** These measurements support Theorem 1 in the form in which we state it. Its premises are satisfiable, they are not satisfied by cross-entropy alone, and the geometry tracks them: both premises, under the Adam family, give the frequency-separated low-rank partition, (P2) alone, under signSGD, gives spectral concentration but a partition that captures only 0.30 of the frequent set, and without either premise, under momentum SGD and the hybrid, no frequency-aligned partition appears, although momentum SGD at the largest learning rate still reaches a low effective dimension, so (P2) is a sufficient route to spectral concentration, not the only one.

Because the theorem states sufficient conditions, a run that violates a premise cannot contradict it. Proposition 1, in contrast, makes a prediction that these runs could have refuted: an untied run of momentum SGD keeps its mean row at its random initialization, so its partition should stay at chance whatever the learning rate. No run refutes this prediction; the partition stays close to chance at every learning rate tested. Comparable ICL accuracy across all runs but one further separates the geometry from the behavior it reads, which the main text uses when interpreting $\Delta\mathcal{A}$ as a readout channel rather than a prerequisite for in-context learning.

### G.7 When the Mean Axis Forms During Pretraining

The unit-mean axis $m_{\text{unit}}$ of Corollary 1, along which the main text measures $\Delta\mathcal{A}$, is not tracked during training by Figure 4 or Section G.4, which follow spectral concentration, the frequency partition and the cohesion of the frequency groups, and Section G.6 measures it only at the end of training. We ask when this axis acquires its direction and its frequency ordering, both of which the early-phase analysis of Section A.3 predicts. Write $\bar{x}$ for the mean final state and $e = \bar{x}/\|\bar{x}\|_2$ for its direction. While predictions are close to uniform, the rare rows and the raw mean row $\bar{u}$ move along $-\bar{x}$ (Propositions 2 and 3), and each row's projection $u_i^\top e$ grows with its frequency (Proposition 2(2)), which sets the frequency threshold of the raw-mean partition. The unit-mean axis, which averages the normalized rows $v_i$ instead, points along $-\bar{x}$, and the depth $\delta_i = -v_i^\top \widehat{m}$ of a row along it grows with its frequency (Proposition 5), which orders the tokens along this axis. These are predictions of signs and orderings, not of magnitudes, and they concern the phase in which Assumption 1 holds. We test them on a pretrained language model.

We measure 28 Pythia-1B-deduped checkpoints from step 0 to step 143,000 (Biderman et al., 2023), using 256 WikiText-103 validation windows of 128 tokens and discarding the first four positions of each window. As the frequency $\pi_i$ of token $i$ we use its unigram frequency in WikiText-103, and we call a token rare if $\pi_i < 1/V$. Pythia's output matrix is untied and trained by an Adam-family optimizer, as Assumption 1(i) requires. To locate the phase in which Assumption 1(ii) holds, we record $\bar{p}_R V$, the probability that the model assigns to a rare token, averaged over all rare tokens and positions and multiplied by $V$; it equals one when predictions are uniform, $p_i(x) = 1/V$, and falls below one as probability moves to the frequent tokens, so a value well below one marks the end of the phase that Assumption 1 describes. Because every prediction is stated along $\bar{x}$, we record how strongly the final states share its direction through the norm $\|\overline{\widehat{x}}\|_2$ of the average unit-normalized final state, which is one when all states point the same way and near zero when they are isotropic. The remaining quantities test the propositions in turn. Proposition 2 predicts that the rare rows move along $-\bar{x}$ by a repulsion they share, while their initializations and attractions differ from row to row, so we average them and measure the cosine between $-\bar{x}$ and the movement $\Delta\bar{u}_R$ of the mean rare row since initialization. Proposition 3 predicts that the raw mean row moves along $-\bar{x}$ as well, the movement

that (P1) assumes; once this movement outweighs the initial mean row, $\bar{u}$ itself points against $\bar{x}$, which we measure by $\cos(\bar{u}, \bar{x})$. For the frequency ordering of Proposition 2(2) we measure the projection–frequency correlation, the Spearman correlation between the projection $u_i^\top e$ and $\log \pi_i$, which is positive when more frequent tokens project further along $\bar{x}$. For Proposition 5 we measure $\cos(\widehat{m}, \bar{x})$, whose sign item 1 predicts, and the depth–frequency correlation, the Spearman correlation between the depth $\delta_i$ and $\log \pi_i$, which is positive when more frequent tokens point further against the axis, as item 2 predicts. As in Section G.6, we also record the axis norm $\|m_{\text{unit}}\|$, which Corollary 1 predicts to be nonzero once a coherent majority of non-frequent rows forms; we compare it with its random level, about $1/\sqrt{V}$ for the average of $V$ independent random unit rows. We measure the same endpoint quantities for the seven pretrained LLMs of the main text, and at these endpoints and the final Pythia checkpoint we also measure the Spearman correlation between the row norm $\|u_i\|_2$ and $\log \pi_i$, which tells whether frequency is encoded in the directions of the rows or in their norms. All cosines are taken between vectors of the same checkpoint, because the residual stream can change basis during training, with the change absorbed by the weights that read and write it, so a direction at one checkpoint need not correspond to the same direction at another. The one exception is $\Delta\bar{u}_R$, which spans checkpoints because the prediction of Proposition 2 concerns the accumulated parameter update of Equation (9). Every checkpoint verifies that the state we use reproduces its logits, with relative error at most $2 \times 10^{-6}$.

Table 12 lists all checkpoints. At initialization every quantity is at its random level, with $\cos(\widehat{m}, \bar{x}) = 0.000$, depth–frequency correlation $0.004$ and $\|m_{\text{unit}}\| = 0.0044 \approx 1/\sqrt{V}$. Predictions on the rare tokens stay close to uniform through step 64, with $\bar{p}_R V \geq 0.79$, and move away from it afterwards, to $0.48$ at step 128 and $0.24$ at step 1,000, so the phase that Assumption 1 describes ends between steps 64 and 128, the two checkpoints that bracket this drop.

Every predicted sign appears by step 64, within this phase, and the geometry forms quickly. The final states share a strong common direction ($\|\widehat{\bar{x}}\|_2 = 0.93$–$0.94$ at steps 16–32), so $e$ is well defined. The mean rare row moves along $-\bar{x}$ from the first update, with $\cos(\Delta\bar{u}_R, -\bar{x}) = 0.84$ at step 2, and after a transient dip to $0.58$ at step 16 this alignment reaches $0.92$ at step 64 and $0.95$ at step 128. The rare rows make up $89\%$ of the vocabulary, and under a preconditioner that differs across rows the frequent rows do not cancel their movement, so the raw mean row follows them and turns against $\bar{x}$, with $\cos(\bar{u}, \bar{x}) = -0.87$ at step 64 and $-0.94$ at step 128, just as Proposition 3 predicts. The unit-mean axis follows: $\cos(\widehat{m}, \bar{x})$ is negative from step 4 on and falls from $-0.11$ at step 16 to $-0.44$ at step 32, $-0.87$ at step 64 and $-0.94$ at step 128; Proposition 5(1) predicts only the sign of this cosine, and the measured anti-alignment is nearly maximal. The frequency orderings form in the same interval: the projection–frequency correlation rises from $0.07$ at step 32 to $0.26$ at step 64 and $0.50$ at step 128, and the depth–frequency correlation from $0.05$ to $0.25$ and $0.48$, as Proposition 2(2) and Proposition 5(2) predict. The axis also leaves its random level here, with $\|m_{\text{unit}}\|$ rising from $0.0044$ to $0.014$ at step 64 and $0.052$ at step 128, and it keeps growing after the phase, to $0.11$ at step 256 and $0.23$ at step 1,000, about 50 times its random level.

After the early phase the analysis makes no further prediction, so we report how each signature evolves. The two frequency orderings persist: from step 256 on the depth–frequency correlation stays between $0.50$ and $0.58$ and the projection–frequency correlation between $0.67$ and $0.73$, close to their values at step 256, so later training preserves the ordering set up in the early phase rather than undoing or amplifying it. The axis norm levels off: from step 1,000 on $\|m_{\text{unit}}\|$ stays between $0.21$ and $0.25$, so the axis that the main text reads is in place after less than $1\%$ of training. The direction of the axis relative to $\bar{x}$ changes instead: the anti-alignment that Proposition 5(1) predicts for the early phase keeps its sign at every checkpoint but weakens, with $\cos(\widehat{m}, \bar{x}) = -0.50$ at step 2,000, $-0.21$ at step 48,000 and $-0.02$ at the final checkpoint, while the depth–frequency correlation along the axis holds steady. Table 13 shows that the seven pretrained models end in the same configuration. Their axes are nearly orthogonal to $\bar{x}$, $|\cos(\widehat{m}, \bar{x})| \leq 0.20$, although the final states still share a clear common direction, $\|\widehat{\bar{x}}\|_2 = 0.41$–$0.63$; their depth–frequency correlation, $0.55$–$0.69$, matches Pythia's; and the Spearman correlation between the row norm $\|u_i\|_2$ and $\log \pi_i$ lies between $-0.19$ and $+0.03$, so frequency is encoded in the directions of the rows, which the depth and the cosine-based $\Delta\mathcal{A}$ read, rather than in their norms.

These measurements agree with the early-phase account. While predictions on the rare tokens are still close to uniform, the rare rows and the raw mean row move along $-\bar{x}$, the unit-mean axis forms along $-\bar{x}$, and the rows are ordered by frequency both along $e$ and along the axis, as Propositions 2, 3 and 5

predict. The frequency ordering then persists, at the level reached by step 256, for the remaining 99% of training, and the axis norm, which begins to grow in the early phase, levels off by step 1,000. The later weakening of the anti-alignment lies outside the early phase, where the analysis makes no prediction, and we report it as an observation.

### G.8 Relation to the Token-Frequency Direction

Corollary 1 and the experiments that follow read the vocabulary through the unit-mean axis $\widehat{m}$, along which tokens are ordered by frequency. Stolfo et al. (2024) identify a frequency direction in the output distribution. They define it in logit space from corpus statistics alone, as $v_{\text{freq}} = \log p_{\text{freq}} - \text{mean}(\log p_{\text{freq}}) \in \mathbb{R}^V$, and call a logit change frequency-mediated to the extent that it lies along $v_{\text{freq}}$. The two directions live in different spaces: $\widehat{m}$ in the space of final states that the unembedding reads, and $v_{\text{freq}}$ in the space of logits. We therefore ask to what extent moving a state along $\widehat{m}$ changes the logits along $v_{\text{freq}}$, that is, whether the two directions correspond.

Moving a state along $-\widehat{m}$ changes the logits by $-U\widehat{m}$. Its cosine with $v_{\text{freq}}$, after centering, equals the Pearson correlation between $-u_i^\top \widehat{m}$ and $\log \pi_i$ over tokens, which we compute over a random half of the tokens seen in WikiText-103. Because $\log \pi_i$ is not exactly linear in the rows, no direction attains a cosine of one; as a baseline for the attainable value, we also report the cosine that the least-squares direction for $\log \pi_i$, fitted on the other half of the tokens, reaches on the same tokens.

Table 14 reports both cosines. The logit change along $-\widehat{m}$ has cosine 0.54–0.67 with $v_{\text{freq}}$, against 0.84–0.91 for the baseline, so the coordinate along the axis explains 29–45% of the variance of log frequency. Moving along the axis therefore shifts the logits substantially in frequency order, so $\Delta\mathcal{A}$ modulates the unigram prior, but the axis is not the frequency direction. Its interest here is that cross-entropy training produces it (Section A) and that ICL moves states along it (Section 4.3).

### G.9 Additional Geometric Analysis of Directional Retention

As an additional geometric analysis, we use directional retention to ask whether the learned net updates rotate the two frequency groups by different amounts. This diagnostic complements the CE account but is not asserted to follow directly from token frequency or from Theorem 1. At each checkpoint, we measure the cosine between every token unembedding and its initialization, define the frequent set as the top $|\mathcal{O}_{\text{raw}}|$ tokens by frequency, and compare it with the full non-frequent complement in both the synthetic model and Pythia-1B.

Figure 46 shows the same ordering in the two systems: frequent rows retain more of their initial direction, whereas non-frequent rows undergo greater angular change. The comparison is within each model and checkpoint, so it does not assume that the synthetic and language-model parameterizations have identical scales. The exact identity in Lemma 5 shows that this ordering is entirely determined by the net transverse component of each accumulated update. Because CE alone does not order those transverse components without an additional relation between initialization and the conditional target directions, we treat retention as an empirical training-dynamic diagnostic rather than a guaranteed consequence of token frequency.

### G.10 Additional Group-Level Geometric Analysis

The main text uses full-matrix effective rank and the raw-mean partition because they correspond directly to the observed global geometry. As an additional geometric analysis, we compare the two frequency groups with finer summaries; these measurements are not presented as direct consequences of CE dynamics. For every model, we define the frequent set by the top $|\mathcal{O}_{\text{raw}}|$ corpus-frequency ranks and use the full vocabulary complement as the non-frequent group. Cohesion aggregates cosine similarity over distinct normalized row pairs, compactness averages cosine to the raw group centroid, and $\text{GNEDim}(Q) = \text{EDim}(U_Q)/|Q|$ measures spectral spread per group row. Theorem 3 conditionally extends premise (P2) from the set $G$ of non-frequent rows to the complete groups by accounting for the non-frequent rows outside it, the row energy it covers, and average frequent-side dispersion. When its aggregate bounds are separated, it gives greater non-frequent cohesion and compactness but greater frequent group-normalized effective dimension.

Figures 47–53 compare cohesion and compactness between the synthetic model and all seven LLMs. In every model, frequent rows are less cohesive and less compact than the non-frequent complement; equally sized random subsets do not reproduce the separation. Figures 54–60 show that frequent rows also have higher group-normalized effective dimension. These are the full-group orderings characterized conditionally by Theorem 3. The theorem allows a small non-frequent tail outside $G$ and requires only average frequent dispersion, while still making explicit that the required geometric margins are not implied by token frequency alone. The figures verify the orderings themselves; they do not separately certify every conservative bound in the theorem.

# H Supplementary Materials for Subsection 4.3

This section supports Section 4.3. It relates the alignment axis to the frequency axis, verifies the decomposition of Figure 6 in the synthetic model, replicates Figures 5 and 6 on the other LLMs, reports the outlier, frequency, and depth status of every label, and tests which parts of the prompt produce the orthogonal part. Prompt templates and symbol mappings are given in Section C.

## H.1 The Alignment Axis and the Frequency Axis

Let $\psi_1$ be the leading right singular vector of $U$, oriented so that the coordinate $u_v^\top \psi_1$ of Figure 3 correlates positively with log token frequency. The cosine between $\psi_1$ and $m_{\text{unit}}$ is $-0.997$ for Llama3.1-8B, $-0.999$ for Llama3.2-1B, $-1.000$ for Llama3.2-3B, $-0.996$ for Llama3-8B, $-0.990$ for Llama3-70B, $-0.995$ for GLM4-9B, and $-0.990$ for GLM4-32B. The leading principal component of the row-centered unembedding, oriented the same way, gives cosines from $-0.96$ to $-0.81$. In every model, $m_{\text{unit}}$ therefore points toward the low-frequency end of the frequency axis, and a positive $\Delta\mathcal{A}$ moves the final state away from the frequent tokens.

## H.2 Synthetic-Model Verification of the Alignment Decomposition in Figure 6

Figure 61 repeats the decomposition from Figure 6 in the synthetic model. On the 500 prompts of Table 1, with natural labels, the label part $\Delta\mathcal{A}_{\text{label}}$ is $-0.011$ and the orthogonal part $\Delta\mathcal{A}_{\perp}$ is $+0.065$, yielding the $\Delta\mathcal{A} = +0.054$ of Table 1. Under symbol labels, the label part is near zero ($-0.0003$) and the orthogonal part is $+0.032$, yielding $+0.032$; each value is rounded separately. Thus the positive mean-alignment shift is again supplied primarily by directional change orthogonal to the active task-label subspace, and this conclusion survives replacing the literal label unembeddings. Its agreement with Llama3.1-8B shows that the label-robust, vocabulary-wide component is reproduced in the controlled setting rather than being specific to a pretrained LLM.

## H.3 Replication of Figure 5

Figures 62–67 repeat Figure 5 for the six other LLMs. Across the 49 model–task pairs with natural labels, the label part is negative in all 49 and the orthogonal part is positive in 46; the three exceptions are GLM4-9B on SST-2, RTE, and MR. The total shift is positive in 40 pairs: in every task of Llama3.2-1B, Llama3.2-3B, Llama3-70B, and GLM4-32B; in all but TREC for Llama3.1-8B and Llama3-8B, where the label part of the six TREC labels offsets the orthogonal part; and in no task of GLM4-9B. In every model the right panel places all natural labels against $m_{\text{unit}}$ and near the frequent end, although how many of them fall inside $\mathcal{O}_{\text{raw}}$ depends on the model; Sections H.5 and H.6 summarize this status.

## H.4 Replication of the Alignment Decomposition in Figure 6

Figures 68–73 repeat Figure 6 for the six additional LLMs; the two parts sum exactly to each total. Section H.3 reports the natural-label pairs, so we focus here on the symbol labels. After remapping, the label part shrinks to at most 0.003 in magnitude, against up to 0.030 with natural labels, and loses a consistent sign: it is negative in 38 of the 49 model–task pairs and positive in 11. The orthogonal part stays positive in 42 pairs; the seven exceptions, GLM4-9B on SST-2, MR, and SUBJ and Llama3.2-1B and Llama3.2-3B on SNLI and MNLI, all lie within 0.006 of zero. The total shift is positive in 41 pairs, against 40 with natural labels. Averaged over tasks, the orthogonal part is positive under both label types in six of the seven models; GLM4-9B is the exception, with an orthogonal part

close to zero under both, so its label part keeps the total shift non-positive. Remapping to non-outlier symbols thus removes the negative label part in every model while leaving the orthogonal part largely intact.

### H.5 Outlier and Frequency Status of the Labels

Table 15 lists, for every model and task, how many natural and symbol labels lie in $\mathcal{O}_{\text{raw}}$ and in the size-matched frequent set $F_k$. Membership varies across models because the two sets differ in size: they hold 6.1–10.4% of the vocabulary in Llama3.1-8B, Llama3-8B, and Llama3-70B but only 0.5–2.2% in the other four models. The natural labels nevertheless rank near the frequent end in every model. Ranked by $\cos(u_i, m_{\text{unit}})$ from the most negative, their median lies in the top 1.4–3.0% of the vocabulary, and their median frequency rank lies in the top 2.1%; where the two sets are small, the labels sit just outside them. No symbol label lies in $\mathcal{O}_{\text{raw}}$ or $F_k$ in any model, and their median ranks lie at 30–62% along $m_{\text{unit}}$ and 34–39% by frequency. $\Delta\mathcal{A}_{\text{label}}$ follows this position rather than set membership: with natural labels it is negative in all 49 pairs, including models where few of them fall inside $\mathcal{O}_{\text{raw}}$, and with symbol labels, which lie near the middle of both rankings, it stays near zero and takes either sign (Section H.4).

### H.6 Label Depth

Table 16 reports, for every model, task, and label type, the label depth $\cos(u_\ell - \bar{u}_{\bar{L}}, m_{\text{unit}})$ of Section 4.3 averaged over the task's labels, together with its range over all 20 labels. Every natural label points against $m_{\text{unit}}$ in all seven LLMs: its depth is negative for all 140 labels, from $-0.47$ to $-0.03$, and the task averages range from $-0.38$ to $-0.07$. Symbol depths scatter around zero, from $-0.15$ to 0.12, with task averages between $-0.06$ and 0.01. These depths match $\Delta\mathcal{A}_{\text{label}}$ in Section H.4. With natural labels, whose depths are clearly negative, $\Delta\mathcal{A}_{\text{label}}$ is negative in all 49 model–task pairs. With symbol labels, whose depths lie near zero, $\Delta\mathcal{A}_{\text{label}}$ takes either sign, negative in 38 pairs and positive in 11, and never exceeds 0.003 in magnitude.

### H.7 Prompt Controls for the Orthogonal Part

To locate the parts of the prompt that produce the orthogonal part, we rerun the ICL queries under three controls that each keep part of the demonstrations.

Each control keeps the query of a paired ICL prompt and, where demonstrations appear, its ten demonstration examples; only the text before the query changes. We use the first half of each task's evaluation queries, on which Section 4.4 ablates the IHs and estimates the IH steering vectors, so that the ICL prompts are the ones analyzed there. With the record templates of Table 4, the three controls are:

- **Options:** the word "Options:" followed by the task's labels separated by commas, then the query; for SST-2, `Options: negative, positive. {sentence} Sentiment:`.
- **Label format:** one record per label with its input removed, in a random order drawn for each query, then the query; for SST-2, `Sentiment: positive. Sentiment: negative. {sentence} Sentiment:`. These records show the answer slot and every label but no input and no input–label pairing.
- **Random labels:** the ten ICL demonstrations, with each label replaced by one drawn uniformly and independently from the task's label set (Min et al., 2022).

Each control and the 10-shot ICL prompt are compared with ZSL on the same query, and $\Delta\mathcal{A}$, $\Delta\mathcal{A}_{\text{label}}$, and $\Delta\mathcal{A}_{\perp}$ follow Equation (7) with the natural-label subspace.

Figure 74 reports the three parts for every LLM. In the six LLMs whose ICL orthogonal part is positive, the share of that part reproduced by a control is 0.04–0.29 for the options, 0.48–0.87 for label-format records, and 1.02–1.20 for random labels, and this ordering holds in every model; in Llama3.1-8B the shares are 0.04, 0.48, and 1.07. The orthogonal part therefore needs demonstration records but not a correct input–label mapping, in line with the insensitivity of ICL accuracy to random labels (Min et al., 2022). The label part is negative under every control in all seven LLMs, and in

every model its magnitude grows from the options to label-format records to random labels, where it matches ICL. It therefore tracks how often the label tokens appear in the context, not whether they are paired with the right inputs: any prompt that shows the labels moves the state toward them and, given where the labels sit on the alignment axis (Section 4.3), lowers alignment. The two parts thus have different sources, the label part the label tokens themselves and the orthogonal part the demonstration records. We do not report shares for GLM4-9B, whose ICL orthogonal part is close to zero (Section H.4) and cannot serve as a denominator; in this model, listing the options alone yields a larger orthogonal part ($+0.005$) than ICL.

# I Supplementary Materials for Subsection 4.4

## I.1 Synthetic IH Significance for Label Probability

To test whether the IH-mediated increase in correct-label probability observed in Section 4.4 also arises in the synthetic model, we compare two quantities for every attention head. Its IH score is the mean attention from the final query delimiter to demonstration-label positions, while its correct-label DLA is the prompt-averaged contribution of its paired ICL–ZSL output change to the correct-label logit. The two head-level scores are positively correlated in Figure 75 (Pearson $r = 0.312$, $p = 0.00197$), and their largest values are concentrated in the first layer. Thus heads that retrieve demonstration labels also tend to write the output component that promotes the correct label.

Figure 76 further averages the IH scores and the entropy of each head's final-position attention distribution within every layer. The first layer has both the highest mean IH score and the lowest mean attention entropy. Its heads therefore place comparatively concentrated attention on the demonstration labels, consistent with the head-level DLA relationship and with a localized synthetic IH mechanism.

## I.2 Replication of Figure 7 for Other LLMs

Figures 77–82 repeat the 1–10% targeted-IH ablation sweep from Figure 7 on six additional LLMs. Llama3.2-1B, Llama3.2-3B, Llama3-8B, and GLM4-9B additionally include ten equal-size random-head ablations at every dose as controls. Across all six models, increasing the targeted-IH dose sharply reduces ICL accuracy and moves the output distribution toward its paired ZSL distribution; the ablated mean unembedding alignment also moves away from its ICL value and toward the ZSL reference. Across the four models with random ablation as a contrast, targeted IH removal produces substantially larger changes in all three quantities than the random-head mean, which remains within a comparatively narrow $\pm 2$-standard-deviation band. The replication therefore supports a concentrated causal role for IHs rather than a generic consequence of removing the same number of attention heads.

## I.3 Shot-Accuracy Sanity Checks

The shot sweep interprets increasing alignment as movement induced by increasingly informative demonstrations. That interpretation requires the prompts to exhibit functional ICL rather than merely a geometric response to longer context. We therefore evaluate the same one-to-ten-shot structured prompts using exact label accuracy and plot the paired ZSL accuracy as a fixed reference. At each shot, accuracy is computed from the same prompt pool used for the geometry sweep; unlike the alignment statistic, it depends only on whether the correct candidate label has maximal probability.

Figures 83–89 show that adding demonstrations strengthens the contextual task information available to the model and improves prediction over the ZSL reference across the evaluated sweep. Thus the shot-dependent geometric response in Section F.2 accompanies an increase in ICL functionality.

## I.4 Direct Attribution of IH Outputs to the Correct-Label Shift

The steering experiment in Section 4.4 measures the total effect of the IH output, since every later layer responds to the added vector. Direct attribution instead asks how much of the paired ICL–ZSL change in the correct-label log probability the IH outputs write into the final state themselves. For the selected set $\mathcal{I}$, define $\Delta h_{\mathcal{I}} = \sum_{(l,k)\in\mathcal{I}} \mathrm{Norm}(\Delta a^{l}_{n,k})$, where Norm applies the final-normalization weight and the frozen prompt-specific RMS scale (Section E.2). Softmax gives

$\log p_y(x) = u_y^\top x - \log \sum_v \exp(u_v^\top x)$, hence $\Delta \log p_y \approx u_y^\top \Delta x$ when the log-partition change is ignored. For the top $10\%$ IHs, DLA gives their correct-label contribution $d^{\mathcal{I}} = u_y^\top \Delta h_{\mathcal{I}}$. We measure agreement using $\mathrm{RIE}(\mathcal{I}) = \|d^{\mathcal{I}} - \Delta \log p_y\|_2 / \|\Delta \log p_y\|_2$, which is zero for equality and one when attribution explains none of the change.

Figures 90–96 report this relative identity error on every dataset for all seven LLMs, and Figures 97–103 show the underlying prompt-level relation between the selected IHs' DLA difference and the observed correct-label log-probability difference. For Llama3.1-8B, the selected IHs achieve equal-dataset mean RIE $0.406$, versus $0.887$ for matched random heads. The selected-IH attribution lies closer to the identity relation than equally sized random-head attributions, and the positive prompt-level association recurs across tasks despite differences in scale and dispersion. Although other heads, MLPs, and softmax normalization can explain the remainder, only $10\%$ of heads directly account for a large share of the correct-label shift. Section I.6 validates this attribution by ablation.

### I.5 Replication of Figure 8 and Figure 9 for Other LLMs

Figures 104–109 repeat Figure 8 on the six other LLMs. In every LLM, IH steering moves $\mathcal{A}$ in the direction of the ICL shift, by $0.40$–$0.85$ of its size, while random head sets move it much less. Except for GLM4-9B, the steered shift also keeps the sign pattern of Figure 6, with a positive orthogonal part and a negative label part, and IH steering recovers $0.50$–$0.64$ of the ICL distribution, against $-0.09$–$0.08$ for random heads.

GLM4-9B is the only LLM whose ICL shift is negative ($\Delta\mathcal{A} = -0.0099$). Its label part ($-0.0101$) is comparable to that of the other models, but its orthogonal part is close to zero ($+0.0001$), so nothing offsets the label part. IH steering reproduces this shift: $\Delta\mathcal{A} = -0.0085$, carried by the label part ($-0.0075$) with a small orthogonal part ($-0.0010$). The difference from the other LLMs therefore lies in the orthogonal part of the movement, and steering reconstructs the shift together with its sign. The low probability recovery of GLM4-9B ($0.151$ against $0.003$ for random heads) has a separate cause. The per-prompt ratio in $R_{\mathrm{ICL}}$ is unstable when $D_{\mathrm{TV}}(P_{\mathrm{ZSL}}, P_{\mathrm{ICL}})$ is close to zero, and on TREC a few such prompts lower the dataset mean to $-3.18$, although the TREC median is $0.49$. The other six datasets give $0.49$–$0.86$, and averaging dataset medians instead of means gives $0.718$. For the other LLMs, the two aggregations differ by at most $0.07$ (for Llama3.1-8B, $0.573$ versus $0.550$).

Figure 110 repeats Figure 9 on the four other LLMs for which all baselines were run. Every method is scored by full-vocabulary top-1 accuracy on the same held-out prompts and averaged over the seven tasks. For function and task vectors we report the best injection layer on these prompts, which favors both baselines; I2CL and state vectors use their public implementations with our prompt templates and label tokens, and state vectors are injected into the first half of the layers. IH steering is the most accurate training-free method on every model, $0.10$–$0.20$ above the best of the other three. Function and state vectors score at most $0.11$ on SUBJ for every model, which lowers their averages. Compared with I2CL, IH steering is more accurate on Llama3.2-3B ($0.600$ versus $0.568$), Llama3-8B ($0.646$ versus $0.631$), and GLM4-9B ($0.697$ versus $0.691$), and slightly less accurate on Llama3.2-1B ($0.496$ versus $0.500$).

### I.6 Additional Causal Diagnostics

The identity-error analysis of Section I.4 asks whether the ICL–ZSL DLA difference of selected IHs explains the corresponding correct-label probability difference. Because that comparison is attributional, we validate it causally by ablating those heads, measuring the resulting probability loss, and comparing that loss with the DLA of the removed outputs. We also test whether the selected IH set is distinguishable from arbitrary equal-size head sets by repeatedly sampling matched random sets and comparing the selected-head statistic with its empirical permutation distribution.

Figure 111 reports this attribution–ablation correspondence across tasks for all seven LLMs. Larger DLA contributions predict larger probability losses after removal, providing causal support for the correct-label attribution in Section I.4. Figure 112 further shows in all seven LLMs and the synthetic model that the selected IHs' aggregate ICL–ZSL attribution lies outside the corresponding random-head variation. Together, the controls support both the causal validity of DLA and the specificity of the selected IH set.

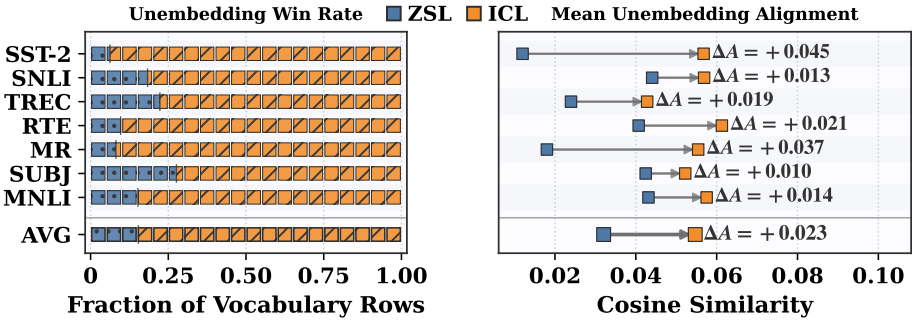



Figure 10: Paired ICL–ZSL unembedding alignment results for Llama3.2-1B.

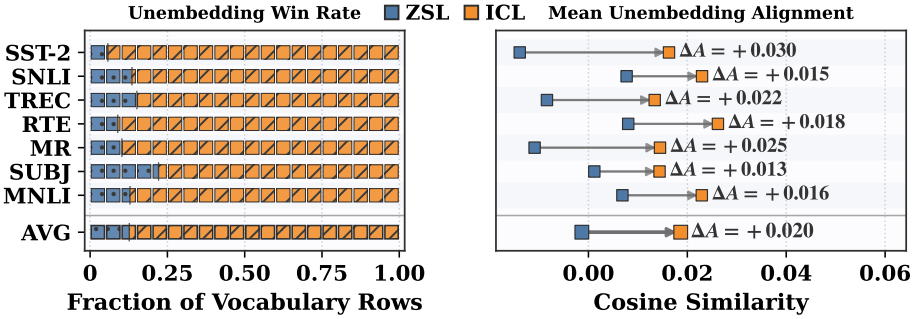



Figure 11: Paired ICL–ZSL unembedding alignment results for Llama3.2-3B.

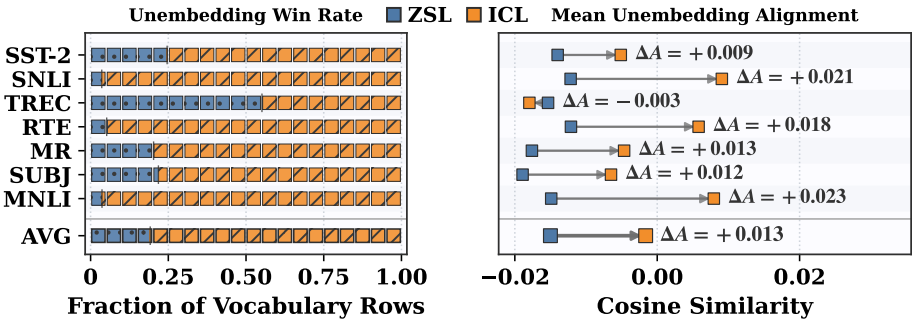



Figure 12: Paired ICL–ZSL unembedding alignment results for Llama3-8B.

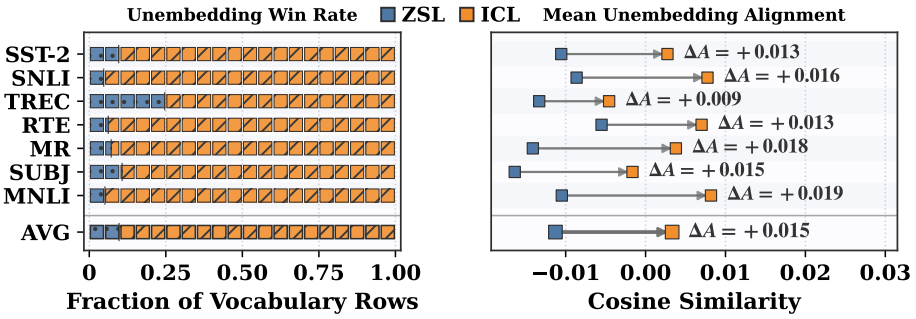



Figure 13: Paired ICL–ZSL unembedding alignment results for Llama3-70B.

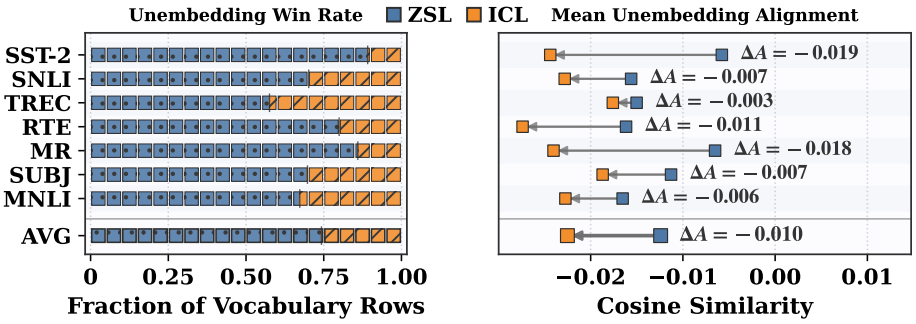



Figure 14: Paired ICL–ZSL unembedding alignment results for GLM4-9B.

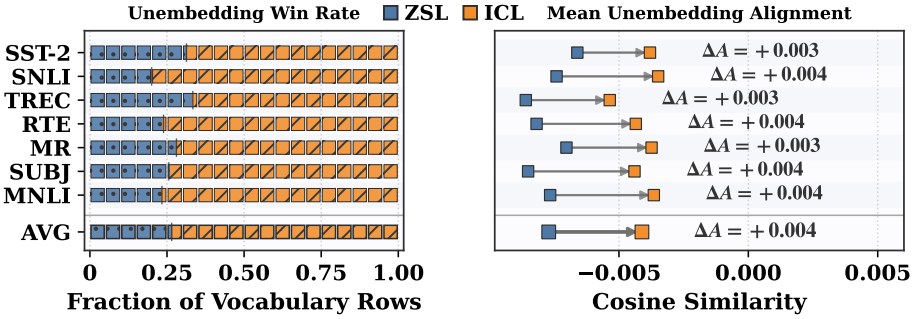



Figure 15: Paired ICL–ZSL unembedding alignment results for GLM4-32B.

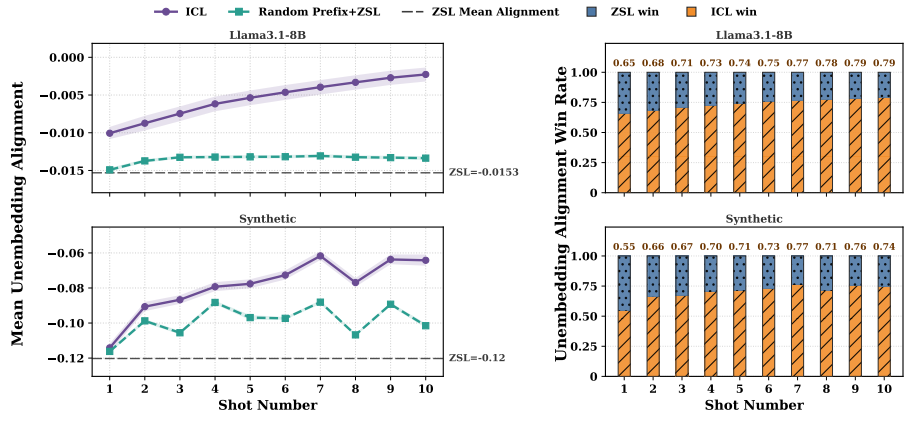



Figure 16: One-to-ten-shot unembedding alignment and win-rate sweep for Llama3.1-8B and the synthetic model.

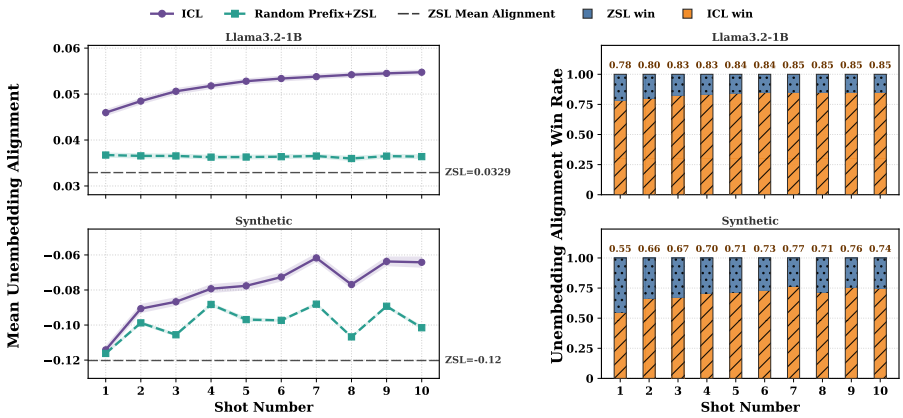



Figure 17: One-to-ten-shot unembedding alignment and win-rate sweep for Llama3.2-1B and the synthetic model.

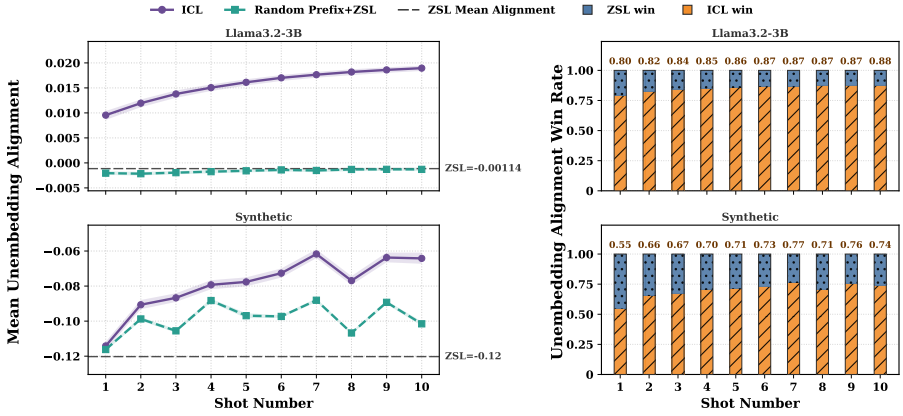



Figure 18: One-to-ten-shot unembedding alignment and win-rate sweep for Llama3.2-3B and the synthetic model.

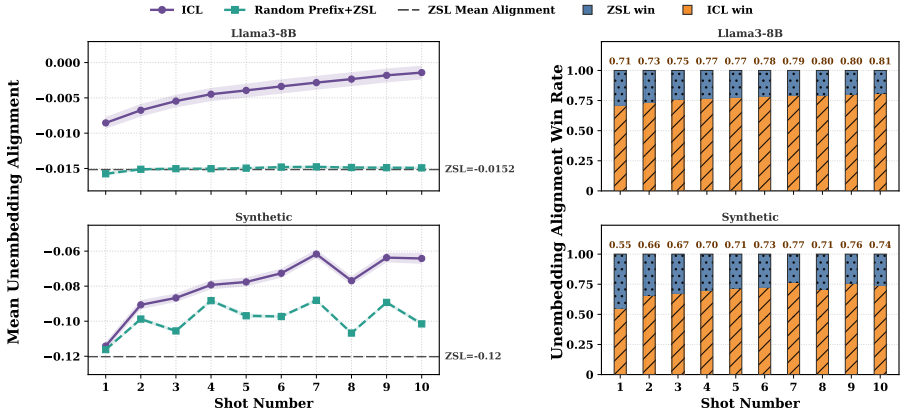



Figure 19: One-to-ten-shot unembedding alignment and win-rate sweep for Llama3-8B and the synthetic model.

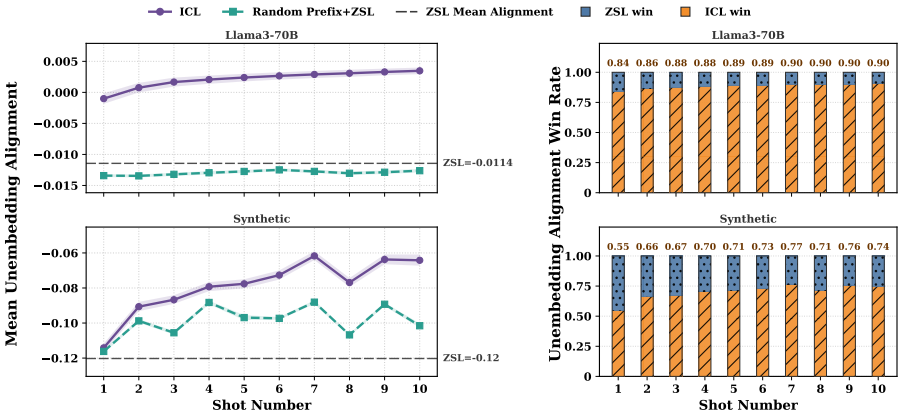



Figure 20: One-to-ten-shot unembedding alignment and win-rate sweep for Llama3-70B and the synthetic model.

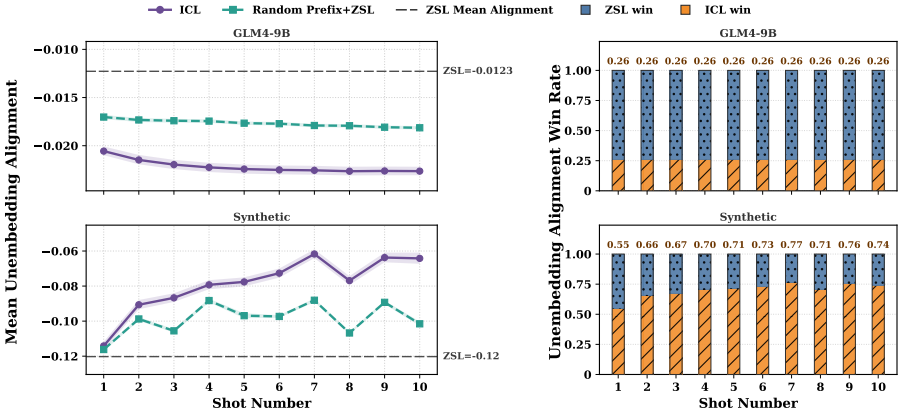



Figure 21: One-to-ten-shot unembedding alignment and win-rate sweep for GLM4-9B and the synthetic model.

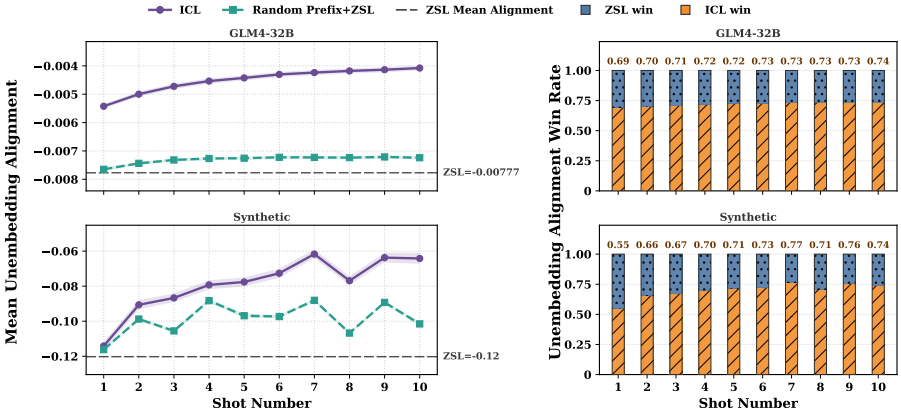



Figure 22: One-to-ten-shot unembedding alignment and win-rate sweep for GLM4-32B and the synthetic model.

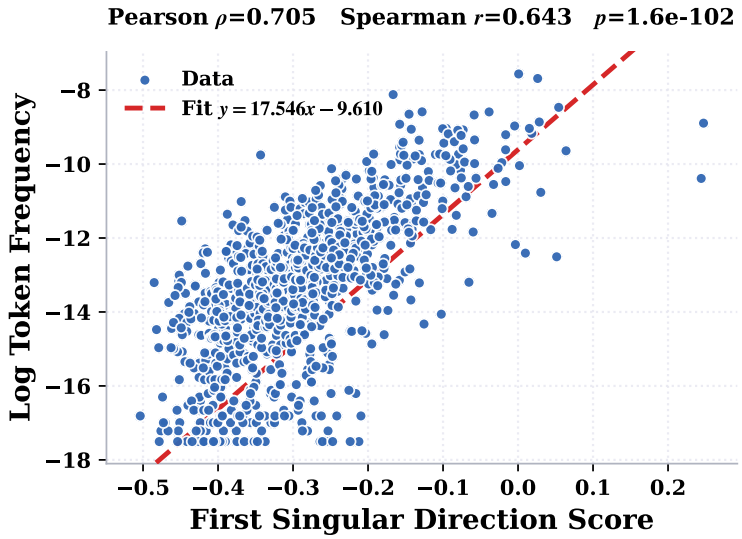



Figure 23: Empirical log token frequency versus the leading unembedding singular-direction coordinate for Llama3.2-1B.

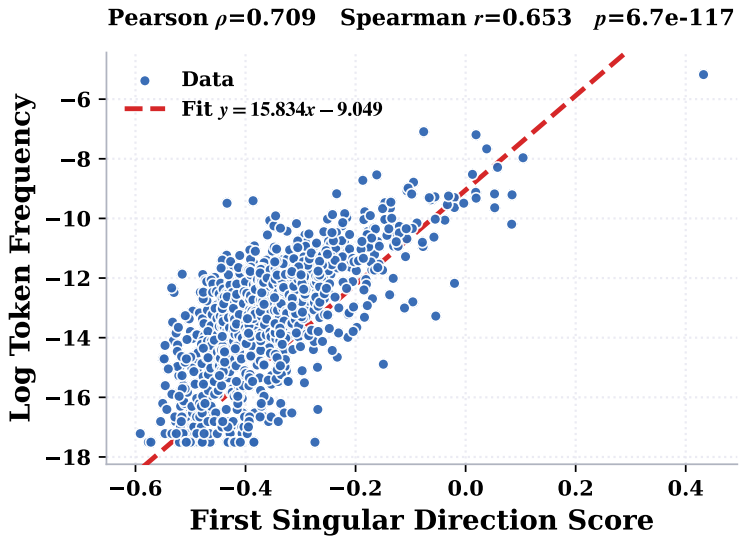



Figure 24: Empirical log token frequency versus the leading unembedding singular-direction coordinate for Llama3.2-3B.

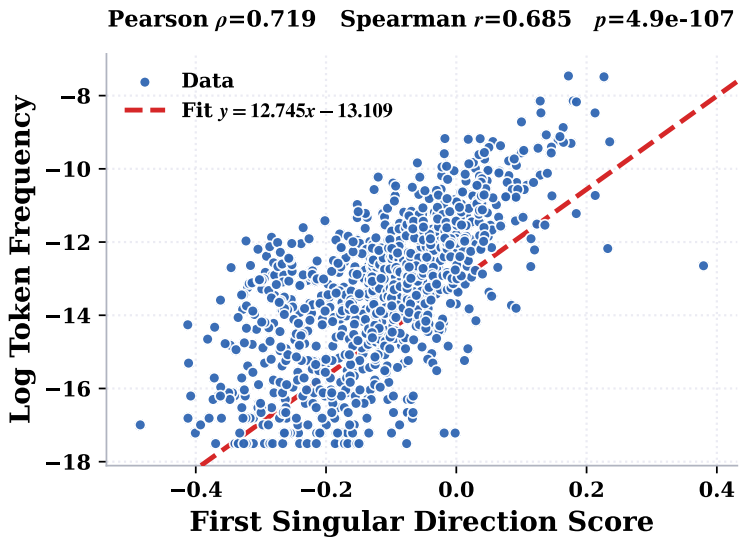



Figure 25: Empirical log token frequency versus the leading unembedding singular-direction coordinate for Llama3-8B.

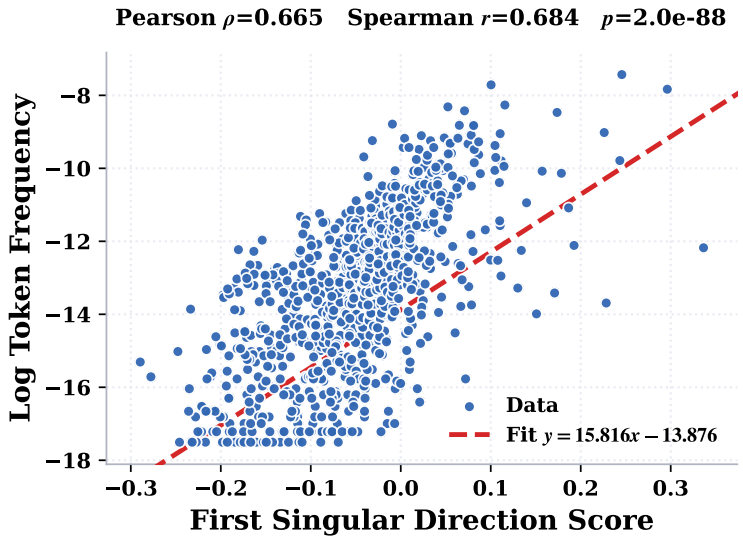



Figure 26: Empirical log token frequency versus the leading unembedding singular-direction coordinate for Llama3-70B.

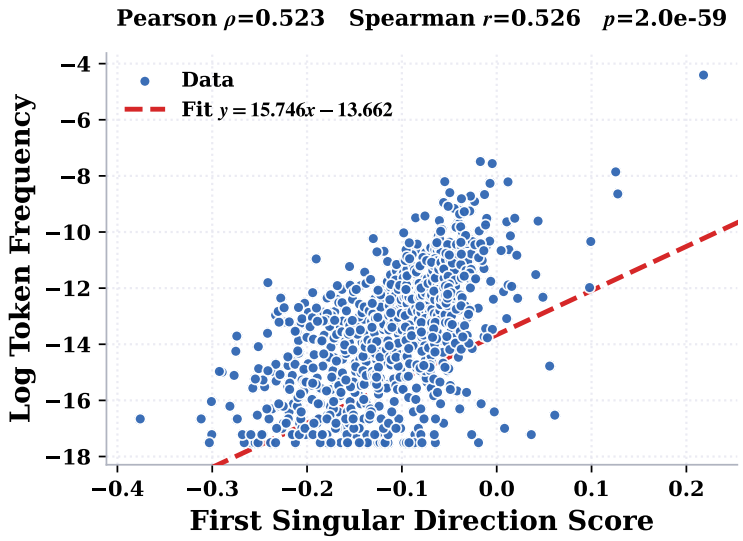



Figure 27: Empirical log token frequency versus the leading unembedding singular-direction coordinate for GLM4-9B.

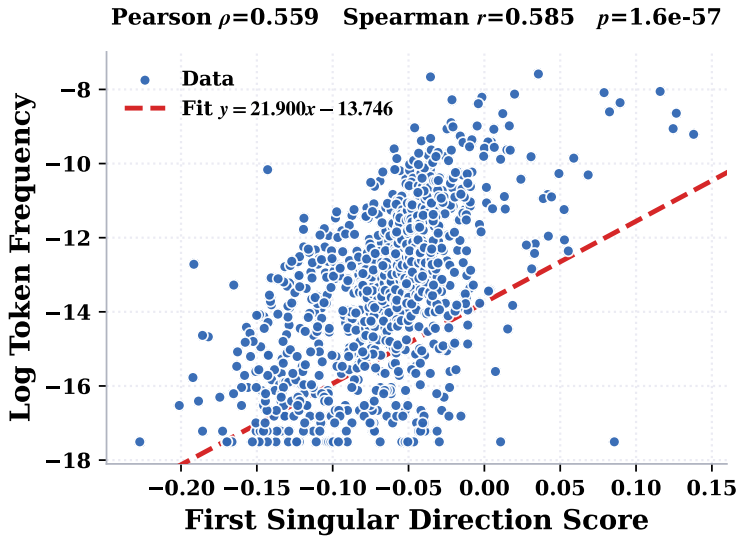



Figure 28: Empirical log token frequency versus the leading unembedding singular-direction coordinate for GLM4-32B.

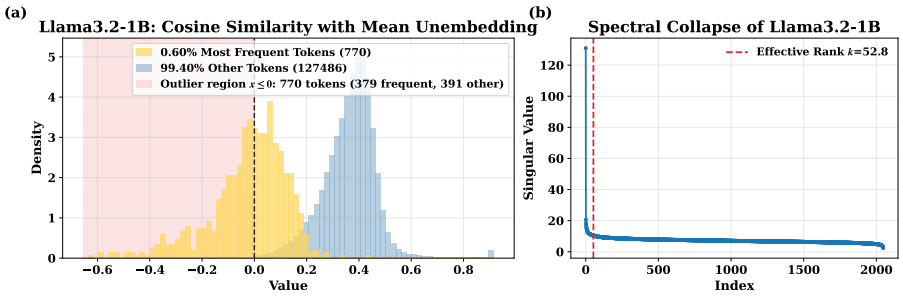



Figure 29: Frequency-linked unembedding bipartition and singular spectrum for Llama3.2-1B.

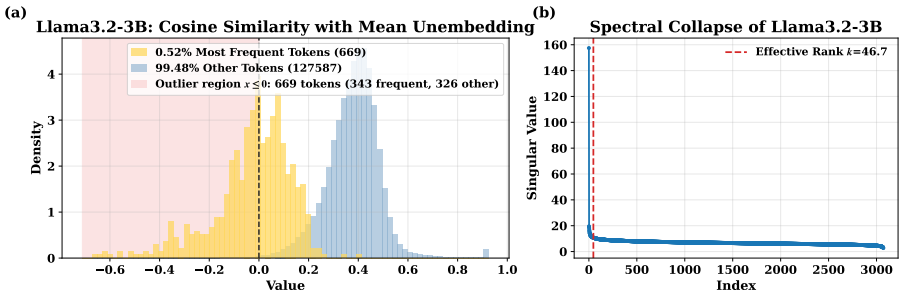



Figure 30: Frequency-linked unembedding bipartition and singular spectrum for Llama3.2-3B.

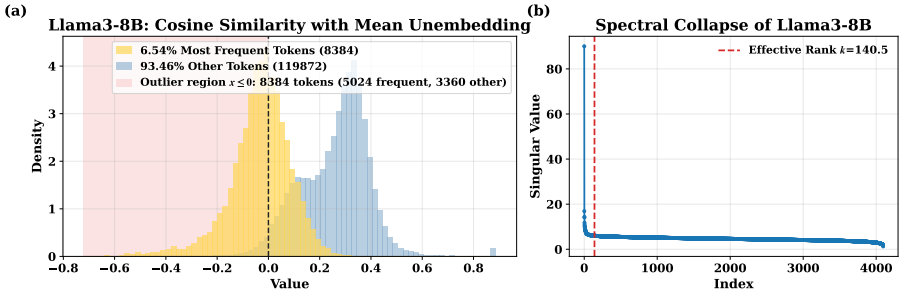



Figure 31: Frequency-linked unembedding bipartition and singular spectrum for Llama3-8B.

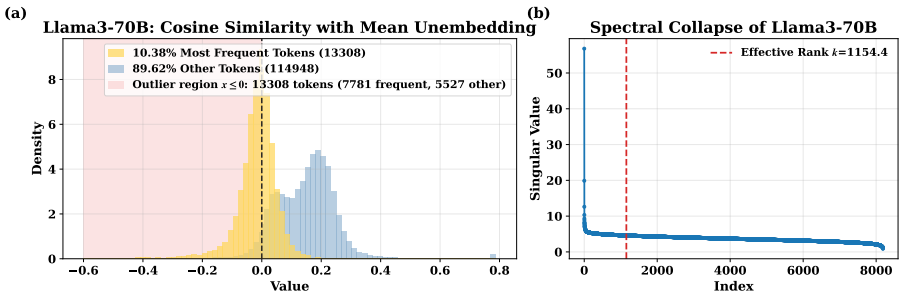



Figure 32: Frequency-linked unembedding bipartition and singular spectrum for Llama3-70B.

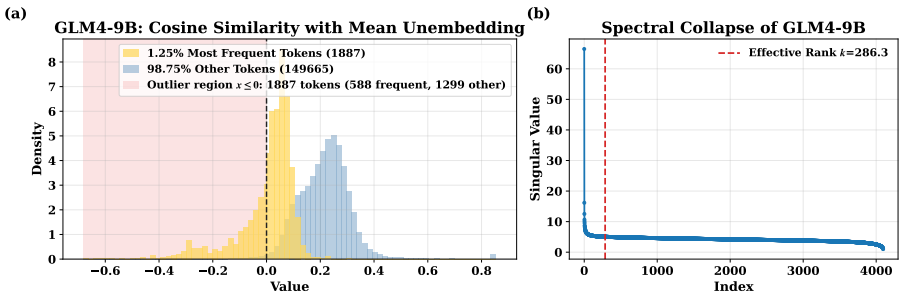



Figure 33: Frequency-linked unembedding bipartition and singular spectrum for GLM4-9B.

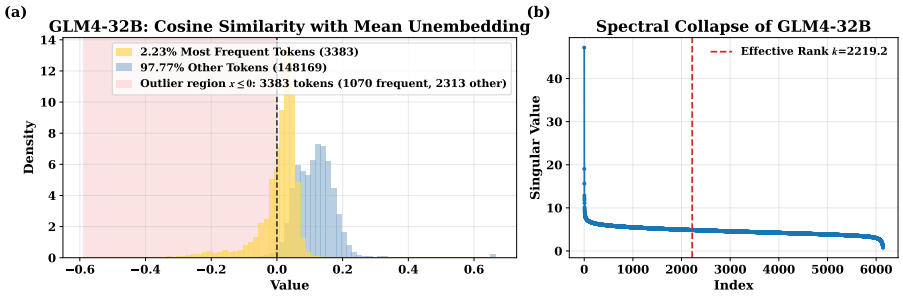



Figure 34: Frequency-linked unembedding bipartition and singular spectrum for GLM4-32B.

Table 6: Absolute Pearson correlation between centered PC1 coordinates and log token frequency (PC1 sign is arbitrary).

| Model | Centered PC1–Log Frequency $\lvert r \rvert$ |
|---|---|
| Llama3.2-1B | 0.6694 |
| Llama3.2-3B | 0.6826 |
| Llama3-8B | 0.6691 |
| Llama3.1-8B | 0.6699 |
| Llama3-70B | 0.5920 |
| GLM4-9B | 0.5469 |
| GLM4-32B | 0.4945 |

Table 7: Cosine similarity between raw and normalized mean-unembedding directions.

| Model | Raw–Normalized Mean Cosine |
|---|---|
| Llama3.2-1B | 1.0000 |
| Llama3.2-3B | 1.0000 |
| Llama3-8B | 0.9999 |
| Llama3.1-8B | 0.9999 |
| Llama3-70B | 0.9998 |
| GLM4-9B | 0.9999 |
| GLM4-32B | 0.9995 |

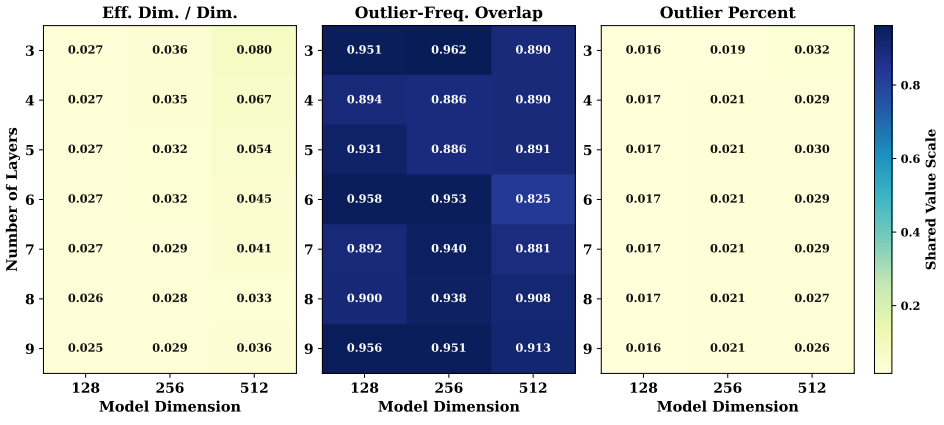



Figure 35: Architecture robustness of the synthetic-model unembedding geometry. Columns vary model dimension and rows vary Transformer depth. The panels report spectral concentration, raw-mean outlier prevalence, and overlap of the outlier group with an equal-size frequent-token set.

Table 8: Replication of Table 1 across model families and scales. Frequent sets are size-matched to raw-mean outlier sets, and $\Delta A = A(x_{\mathrm{ICL}}) - A(x_{\mathrm{ZSL}})$.

| Model | Eff. rank / dim. | Outlier rate | Outlier–freq. overlap | ICL acc. | ICL win rate | $\Delta A$ |
|---|---|---|---|---|---|---|
| Synthetic | 0.032 | 0.021 | 0.953 | 0.518 | 0.734 | +0.054 |
| Llama3.2-1B | 0.026 | 0.006 | 0.492 | 0.549 | 0.846 | +0.023 |
| Llama3.2-3B | 0.015 | 0.005 | 0.513 | 0.668 | 0.873 | +0.020 |
| Llama3-8B | 0.034 | 0.065 | 0.599 | 0.787 | 0.808 | +0.013 |
| Llama3-70B | 0.141 | 0.104 | 0.585 | 0.821 | 0.903 | +0.015 |
| GLM4-9B | 0.070 | 0.012 | 0.312 | 0.851 | 0.257 | -0.010 |
| GLM4-32B | 0.361 | 0.022 | 0.316 | 0.903 | 0.735 | +0.004 |

Table 9: Aggregate-coherence audit of Corollary 1 at exception rate $\phi = 0.005$. The bound $B_{\mathrm{coh}}$ is compared with the observed projection $m_{\mathrm{unit}}^{\top}\widehat{m}_G$.

| System | $\rho_G$ | $\rho_E$ | $\rho_F$ | $B_{\mathrm{coh}}$ | Observed |
|---|---|---|---|---|---|
| Synthetic | 0.2694 | 0.0099 | 0.0233 | 0.5007 | 0.5032 |
| Llama3.2-1B | 0.1333 | 0.0426 | 0.0413 | 0.3588 | 0.3607 |
| Llama3.2-3B | 0.1450 | 0.0266 | 0.0299 | 0.3751 | 0.3766 |

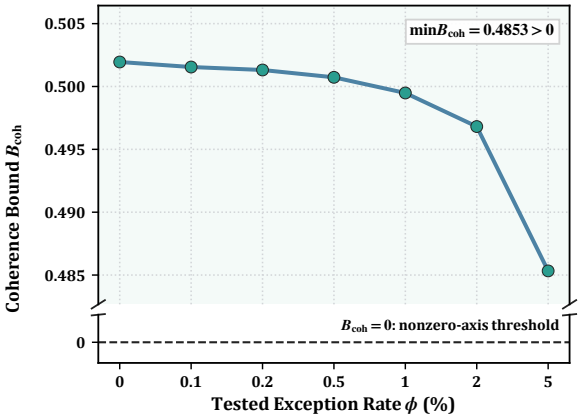



Figure 36: Aggregate-coherence certificate in the synthetic model. The bound remains far above zero throughout the exception-rate sweep.

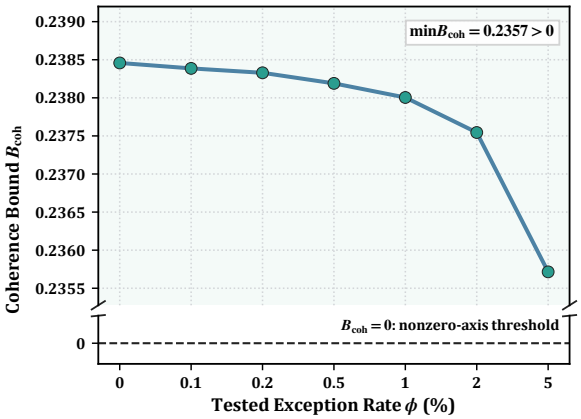



Figure 37: Aggregate-coherence certificate in Llama3.1-8B. The non-frequent majority outweighs exception- and frequent-group cancellation throughout the exception-rate sweep.

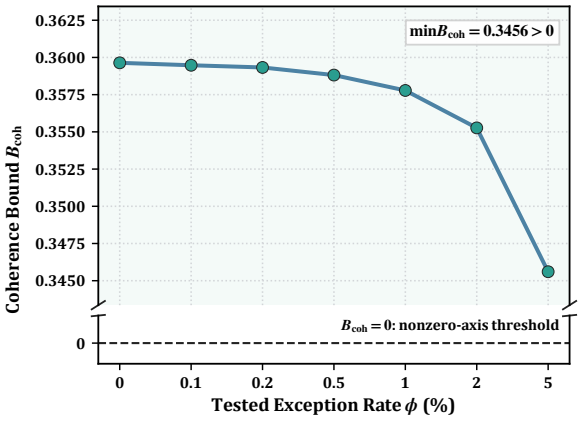



Figure 38: Aggregate-coherence certificate in Llama3.2-1B.

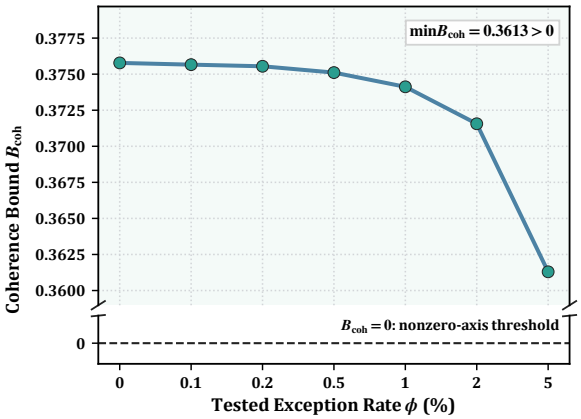



Figure 39: Aggregate-coherence certificate in Llama3.2-3B.

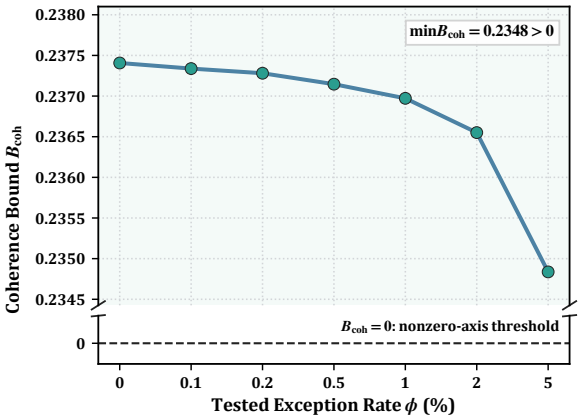



Figure 40: Aggregate-coherence certificate in Llama3-8B.

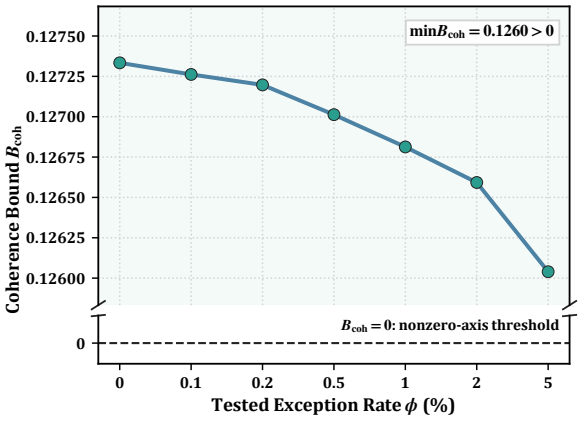



Figure 41: Aggregate-coherence certificate in Llama3-70B.

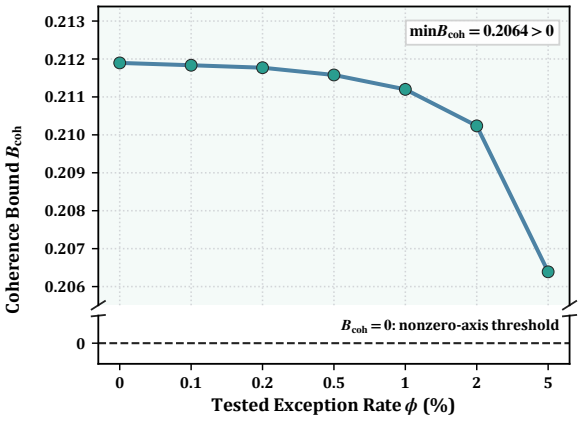



Figure 42: Aggregate-coherence certificate in GLM4-9B.

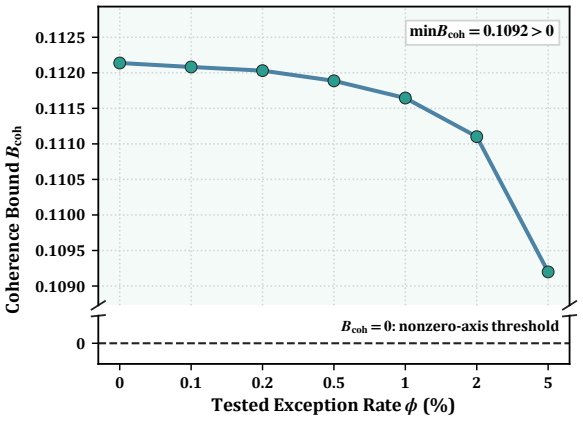
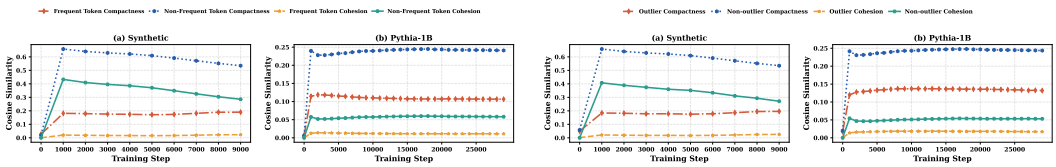



Figure 43: Aggregate-coherence certificate in GLM4-32B.




(a) Frequent/non-frequent cohesion and compactness. (b) Raw-mean group geometry.

Figure 44: Additional geometric training dynamics in the synthetic model and Pythia-1B. Each system uses its own vertical scale so that within-system trajectories remain visible.

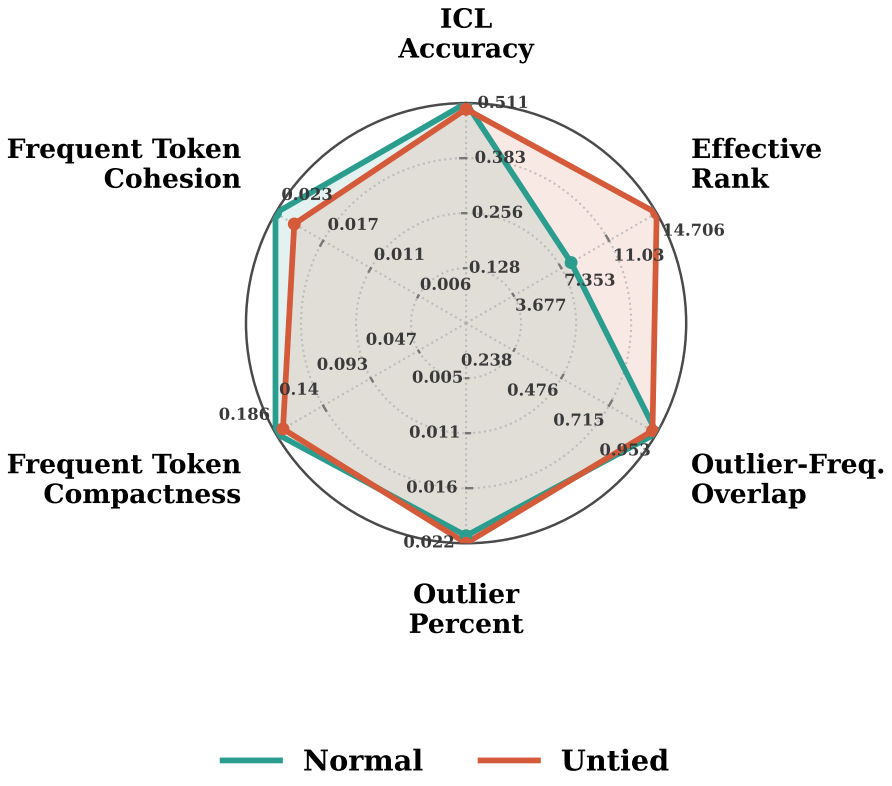



Figure 45: Synthetic comparison of tied and untied output embeddings. ICL accuracy here is computed on all 1,000 held-out prompts.

Table 10: Premises of Theorem 1 and training checks in the optimizer ablation on the synthetic 256-dimensional, 6-layer transformer at the final step (10k), one row per run; Table 11 gives the geometry of the same runs. *Tie*: weight tying; *Opt.*: optimizer, where SGD is momentum SGD with $\beta = 0.9$ and Adam/SGD($U$) applies Adam to every weight except the output matrix $U$, which uses momentum SGD. Untied Adam and untied SGD at lr 0.1 were run with seeds 2026, 7 and 13; all other configurations are single-seed. *Drift*: $\|\bar{u}_{10k} - \bar{u}_0\|/\overline{\|u\|}$, the relative movement of the mean row; $\cos(\Delta\bar{u}, -\bar{x})$: cosine between this movement and the negative mean final state, omitted (–) for the untied runs whose output matrix is updated by momentum SGD, where the movement is rounding error; $\bar{r}$: root-mean-square deviation of the rare rows from their mean row, relative to its norm; *Bnd*: the bound $(1+\bar{r}^2)^2/(d\omega_g^2)$ on EDim/$d$ of Theorem 1, with $G$ the rare rows and $\omega_g$ their share of the squared row norms, to be compared with EDim/$d$ in Table 11; *Loss*: evaluation cross-entropy; *Acc.*: ICL accuracy. $\bar{r}$ and Bnd are measured on reruns of the same configurations, whose EDim/$d$ matches that of the original run to within 4%.

| | | | | Premise (P1) | | Premise (P2) | | Checks | |
|---|---|---|---|---|---|---|---|---|---|
| Tie | Opt. | LR | Seed | Drift | $\cos(\Delta\bar{u}, -\bar{x})$ | $\bar{r}$ | Bnd | Loss | Acc. |
| tied | Adam | 0.0003 | 2026 | 0.526 | +0.702 | 1.54 | 0.0485 | 1.782 | 0.533 |
| tied | AdamW | 0.0003 | 2026 | 0.528 | +0.685 | 1.54 | 0.0479 | 1.775 | 0.531 |
| tied | SGD | 0.03 | 2026 | 0.003 | −0.132 | 28.70 | $3.6\times10^{3}$ | 1.654 | 0.529 |
| tied | SGD | 0.1 | 2026 | 0.005 | +0.179 | 26.96 | $3.1\times10^{3}$ | 1.658 | 0.516 |
| tied | SGD | 0.3 | 2026 | 0.041 | +0.049 | 21.15 | $5.8\times10^{3}$ | 1.660 | 0.382 |
| untied | Adam | 0.0003 | 7 | 0.455 | +0.836 | 1.85 | 0.0829 | 2.174 | 0.489 |
| untied | Adam | 0.0003 | 13 | 0.457 | +0.829 | 1.84 | 0.0817 | 2.148 | 0.483 |
| untied | Adam | 0.0003 | 2026 | 0.455 | +0.840 | 1.85 | 0.0829 | 2.211 | 0.507 |
| untied | Adam/SGD($U$) | 0.03 | 2026 | $1.5\times10^{-6}$ | – | 28.84 | $4.0\times10^{3}$ | 2.016 | 0.524 |
| untied | Adam/SGD($U$) | 0.1 | 2026 | $4.7\times10^{-6}$ | – | 27.06 | $3.3\times10^{3}$ | 2.001 | 0.503 |
| untied | Adam/SGD($U$) | 0.3 | 2026 | $6.7\times10^{-6}$ | – | 27.64 | $3.9\times10^{3}$ | 1.845 | 0.502 |
| untied | AdamW | 0.0003 | 2026 | 0.457 | +0.845 | 1.84 | 0.0817 | 2.205 | 0.476 |
| untied | SGD | 0.03 | 2026 | $1.8\times10^{-6}$ | – | 29.71 | $4.1\times10^{3}$ | 1.656 | 0.506 |
| untied | SGD | 0.1 | 7 | $5.0\times10^{-6}$ | – | 28.00 | $3.5\times10^{3}$ | 1.659 | 0.486 |
| untied | SGD | 0.1 | 13 | $5.7\times10^{-6}$ | – | 29.07 | $4.1\times10^{3}$ | 1.657 | 0.492 |
| untied | SGD | 0.1 | 2026 | $5.2\times10^{-6}$ | – | 27.89 | $3.5\times10^{3}$ | 1.659 | 0.511 |
| untied | SGD | 0.3 | 2026 | $4.3\times10^{-6}$ | – | 19.57 | $1.6\times10^{3}$ | 1.675 | 0.505 |
| untied | signSGD | 0.0001 | 2026 | 0.980 | +0.225 | 0.26 | 0.0045 | 1.725 | 0.513 |

Table 11: Geometry and alignment shift in the optimizer ablation, for the runs of Table 10. *Tie*: weight tying; *Opt.*: optimizer, where SGD is momentum SGD with $\beta = 0.9$ and Adam/SGD($U$) applies Adam to every weight except the output matrix $U$, which uses momentum SGD. Untied Adam and untied SGD at lr 0.1 were run with seeds 2026, 7 and 13; all other configurations are single-seed. *EDim/d*: participation ratio of $U$ divided by $d$; $O$: raw outlier fraction, i.e. rows with $\cos(u_i, \bar{u}) \leq 0$; *Ov.*: fraction of the $|O|$ most frequent tokens inside $O$ (chance $= O$); $\|m_{\text{unit}}\|$: norm of the mean unit row; $\rho_{\text{dep}}$: Spearman correlation between mean-axis depth $\delta_i = -v_i^\top \hat{m}$ and log target frequency $\log \pi_i$; $Res_{\text{freq}}$: mean off-diagonal cosine of frequent unit rows after removing their $\hat{m}$ component; $\Delta\mathcal{A}$, $\Delta\mathcal{A}_{\text{lab}}$, $\Delta\mathcal{A}_{\perp}$: mean-alignment shift and its label and orthogonal components.

| | | | | Geometry | | | | | | Alignment shift | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Tie | Opt. | LR | Seed | EDim/$d$ | $O$ | Ov. | $\|m_{\text{unit}}\|$ | $\rho_{\text{dep}}$ | $\text{Res}_{\text{freq}}$ | $\Delta\mathcal{A}$ | $\Delta\mathcal{A}_{\text{lab}}$ | $\Delta\mathcal{A}_{\perp}$ |
| tied | Adam | 0.0003 | 2026 | 0.0307 | 0.0201 | 0.938 | 0.510 | 0.933 | 0.014 | +0.0459 | −0.0129 | +0.0588 |
| tied | AdamW | 0.0003 | 2026 | 0.0301 | 0.0201 | 0.963 | 0.511 | 0.934 | 0.014 | +0.0387 | −0.0134 | +0.0521 |
| tied | SGD | 0.03 | 2026 | 0.3692 | 0.3723 | 0.476 | 0.031 | 0.675 | 0.037 | −0.0027 | −0.0009 | −0.0018 |
| tied | SGD | 0.1 | 2026 | 0.1451 | 0.3862 | 0.531 | 0.045 | 0.695 | 0.030 | +0.0022 | −0.0012 | +0.0035 |
| tied | SGD | 0.3 | 2026 | 0.0672 | 0.4667 | 0.378 | 0.076 | 0.810 | 0.050 | −0.0043 | −0.0003 | −0.0039 |
| untied | Adam | 0.0003 | 7 | 0.0593 | 0.0209 | 0.905 | 0.453 | 0.909 | 0.012 | +0.0239 | −0.0158 | +0.0397 |
| untied | Adam | 0.0003 | 13 | 0.0580 | 0.0221 | 0.933 | 0.453 | 0.909 | 0.012 | +0.0093 | −0.0132 | +0.0225 |
| untied | Adam | 0.0003 | 2026 | 0.0596 | 0.0221 | 0.910 | 0.453 | 0.905 | 0.012 | +0.0313 | −0.0183 | +0.0496 |
| untied | Adam/SGD($U$) | 0.03 | 2026 | 0.4003 | 0.3964 | 0.443 | 0.032 | 0.679 | 0.027 | +0.0015 | −0.0010 | +0.0024 |
| untied | Adam/SGD($U$) | 0.1 | 2026 | 0.2017 | 0.4113 | 0.479 | 0.050 | 0.718 | 0.024 | +0.0002 | −0.0013 | +0.0015 |
| untied | Adam/SGD($U$) | 0.3 | 2026 | 0.1152 | 0.4707 | 0.447 | 0.078 | 0.866 | 0.037 | +0.0032 | −0.0008 | +0.0040 |
| untied | AdamW | 0.0003 | 2026 | 0.0583 | 0.0221 | 0.910 | 0.454 | 0.907 | 0.012 | +0.0143 | −0.0192 | +0.0335 |
| untied | SGD | 0.03 | 2026 | 0.3829 | 0.3969 | 0.438 | 0.030 | 0.702 | 0.038 | −0.0006 | −0.0006 | −0.0000 |
| untied | SGD | 0.1 | 7 | 0.1637 | 0.4125 | 0.465 | 0.044 | 0.731 | 0.033 | −0.0043 | +0.0001 | −0.0044 |
| untied | SGD | 0.1 | 13 | 0.1456 | 0.4431 | 0.423 | 0.042 | 0.674 | 0.030 | +0.0007 | −0.0004 | +0.0011 |
| untied | SGD | 0.1 | 2026 | 0.1511 | 0.4100 | 0.470 | 0.043 | 0.679 | 0.033 | +0.0028 | −0.0004 | +0.0032 |
| untied | SGD | 0.3 | 2026 | 0.0606 | 0.4324 | 0.460 | 0.068 | 0.694 | 0.030 | −0.0051 | +0.0026 | −0.0076 |
| untied | signSGD | 0.0001 | 2026 | 0.0040 | 0.0075 | 0.300 | 0.928 | 0.684 | 0.252 | −0.0127 | −0.2714 | +0.2588 |

Table 12: Pythia-1B across all 28 released checkpoints; step 0 is the initialization. $\pi_i$: unigram frequency of token $i$ in WikiText-103, with rare tokens $\pi_i < 1/V$; $\bar{p}_R V$: probability that the model assigns to a rare token, averaged over the rare tokens and positions and multiplied by $V$ (one for uniform predictions); $\|\bar{\widehat{x}}\|_2$: norm of the average unit-normalized final state; $\cos(\bar{u}, \bar{x})$ and $\cos(\hat{m}, \bar{x})$: cosines of the raw mean row and of the unit-mean axis with the mean final state; $\cos(\Delta\bar{u}_R, -\bar{x})$: cosine between the movement of the mean rare row since initialization and $-\bar{x}$ (– before the rare rows first move); $\rho(u_i^\top e, \log \pi_i)$ and $\rho(\delta_i, \log \pi_i)$: the projection–frequency and depth–frequency correlations, i.e. Spearman correlations of the projection on $e = \bar{x}/\|\bar{x}\|_2$ and of the mean-axis depth $\delta_i = -v_i^\top \hat{m}$ with $\log \pi_i$; $\|m_{\text{unit}}\|$: norm of the mean unit unembedding row.

| Step | $\bar{p}_R V$ | $\|\bar{\widehat{x}}\|_2$ | $\cos(\bar{u}, \bar{x})$ | $\cos(\Delta\bar{u}_R, -\bar{x})$ | $\cos(\hat{m}, \bar{x})$ | $\rho(u_i^\top e, \log \pi_i)$ | $\rho(\delta_i, \log \pi_i)$ | $\|m_{\text{unit}}\|$ |
|---|---|---|---|---|---|---|---|---|
| 0 | 1.00 | 0.650 | +0.000 | – | +0.000 | −0.001 | +0.004 | 0.0044 |
| 1 | 1.00 | 0.650 | +0.000 | – | +0.000 | −0.001 | +0.004 | 0.0044 |
| 2 | 1.00 | 0.650 | +0.000 | +0.840 | +0.000 | −0.001 | +0.004 | 0.0044 |
| 4 | 1.00 | 0.651 | −0.001 | +0.825 | −0.001 | +0.000 | +0.004 | 0.0044 |
| 8 | 0.99 | 0.754 | −0.033 | +0.694 | −0.033 | +0.011 | +0.004 | 0.0044 |
| 16 | 0.92 | 0.930 | −0.106 | +0.584 | −0.106 | +0.028 | +0.007 | 0.0045 |
| 32 | 0.87 | 0.939 | −0.441 | +0.760 | −0.441 | +0.072 | +0.049 | 0.0054 |
| 64 | 0.79 | 0.811 | −0.871 | +0.920 | −0.871 | +0.261 | +0.251 | 0.0139 |
| 128 | 0.48 | 0.785 | −0.939 | +0.948 | −0.939 | +0.499 | +0.476 | 0.0523 |
| 256 | 0.38 | 0.605 | −0.758 | +0.779 | −0.758 | +0.678 | +0.551 | 0.1141 |
| 512 | 0.30 | 0.549 | −0.638 | +0.660 | −0.635 | +0.734 | +0.501 | 0.1888 |
| 1 000 | 0.24 | 0.440 | −0.678 | +0.695 | −0.671 | +0.719 | +0.544 | 0.2321 |
| 2 000 | 0.23 | 0.430 | −0.506 | +0.523 | −0.497 | +0.670 | +0.576 | 0.2166 |
| 3 000 | 0.23 | 0.447 | −0.391 | +0.409 | −0.385 | +0.683 | +0.578 | 0.2147 |
| 4 000 | 0.22 | 0.471 | −0.360 | +0.375 | −0.355 | +0.688 | +0.578 | 0.2159 |
| 6 000 | 0.22 | 0.503 | −0.327 | +0.339 | −0.323 | +0.690 | +0.576 | 0.2196 |
| 8 000 | 0.22 | 0.518 | −0.314 | +0.324 | −0.311 | +0.682 | +0.575 | 0.2248 |
| 12 000 | 0.22 | 0.513 | −0.294 | +0.304 | −0.291 | +0.684 | +0.571 | 0.2287 |
| 16 000 | 0.21 | 0.529 | −0.276 | +0.285 | −0.275 | +0.684 | +0.568 | 0.2319 |
| 24 000 | 0.21 | 0.527 | −0.248 | +0.259 | −0.247 | +0.680 | +0.558 | 0.2308 |
| 32 000 | 0.21 | 0.529 | −0.237 | +0.248 | −0.236 | +0.679 | +0.554 | 0.2296 |
| 48 000 | 0.21 | 0.520 | −0.214 | +0.228 | −0.214 | +0.683 | +0.560 | 0.2245 |
| 64 000 | 0.21 | 0.525 | −0.180 | +0.196 | −0.180 | +0.684 | +0.571 | 0.2213 |
| 80 000 | 0.21 | 0.535 | −0.135 | +0.152 | −0.134 | +0.683 | +0.577 | 0.2251 |
| 96 000 | 0.20 | 0.570 | −0.078 | +0.094 | −0.075 | +0.678 | +0.578 | 0.2334 |
| 112 000 | 0.21 | 0.601 | −0.034 | +0.050 | −0.030 | +0.677 | +0.577 | 0.2414 |
| 128 000 | 0.21 | 0.616 | −0.022 | +0.038 | −0.018 | +0.678 | +0.575 | 0.2473 |
| 143 000 | 0.20 | 0.625 | −0.022 | +0.037 | −0.018 | +0.674 | +0.573 | 0.2525 |

Table 13: Endpoint output geometry of seven open-weight models and the final Pythia-1B checkpoint. $\|m_{\text{unit}}\|$ is the norm of the mean unit unembedding row; $\|\bar{\widehat{x}}\|_2$ is the norm of the average unit-normalized final state; $\cos(\hat{m}, \bar{x})$ compares the unit-mean axis with the mean final state; $\rho$ columns are Spearman correlations of mean-axis depth $\delta_i = -v_i^\top \hat{m}$ and of row norm with $\log \pi_i$, where $\pi_i$ is the unigram frequency of token $i$ in WikiText-103.

| Model | $\|m_{\text{unit}}\|$ | $\|\bar{\widehat{x}}\|_2$ | $\cos(\hat{m}, \bar{x})$ | $\rho(\delta_i, \log \pi_i)$ | $\rho(\|u_i\|, \log \pi_i)$ |
|---|---|---|---|---|---|
| Llama3.2-1B | 0.361 | 0.633 | +0.107 | 0.662 | −0.136 |
| Llama3.2-3B | 0.377 | 0.534 | −0.029 | 0.673 | −0.099 |
| Llama3-8B | 0.242 | 0.430 | −0.105 | 0.686 | −0.130 |
| Llama3.1-8B | 0.243 | 0.412 | −0.129 | 0.681 | −0.132 |
| Llama3-70B | 0.132 | 0.418 | −0.123 | 0.648 | +0.025 |
| GLM4-9B | 0.214 | 0.429 | −0.195 | 0.585 | −0.189 |
| GLM4-32B | 0.114 | 0.629 | −0.111 | 0.546 | −0.026 |
| Pythia-1B (step 143000) | 0.253 | 0.625 | −0.018 | 0.573 | −0.172 |

Table 14: Cosine between the centered logit change $-U\hat{m}$ along the unit-mean axis and the frequency direction $v_{\text{freq}}$ of Stolfo et al. (2024), i.e. the Pearson correlation between $-u_i^\top \hat{m}$ and log unigram frequency, computed over a random half of the tokens seen in WikiText-103. Baseline: the same cosine for the least-squares direction for log frequency, fitted on the other half of the tokens, which approximates the largest cosine a single direction attains.

| | Cosine with $v_{\text{freq}}$ | |
|---|---|---|
| Model | $-U\hat{m}$ | Baseline |
| Llama3.2-1B | 0.661 | 0.911 |
| Llama3.2-3B | 0.673 | 0.913 |
| Llama3-8B | 0.666 | 0.903 |
| Llama3.1-8B | 0.667 | 0.902 |
| Llama3-70B | 0.628 | 0.896 |
| GLM4-9B | 0.560 | 0.840 |
| GLM4-32B | 0.536 | 0.867 |

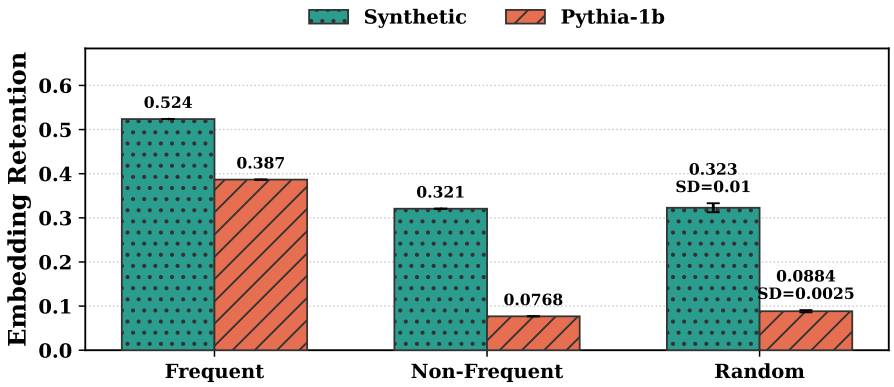



Figure 46: Directional retention during training in the synthetic model and Pythia-1B.

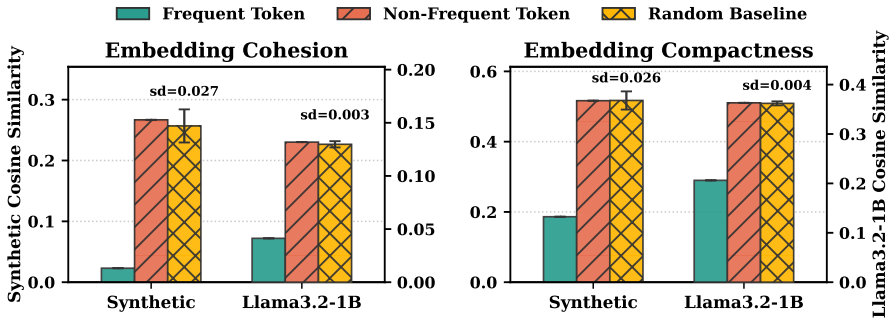



Figure 47: Frequent/non-frequent cohesion and compactness in the synthetic model and Llama3.2-1B.

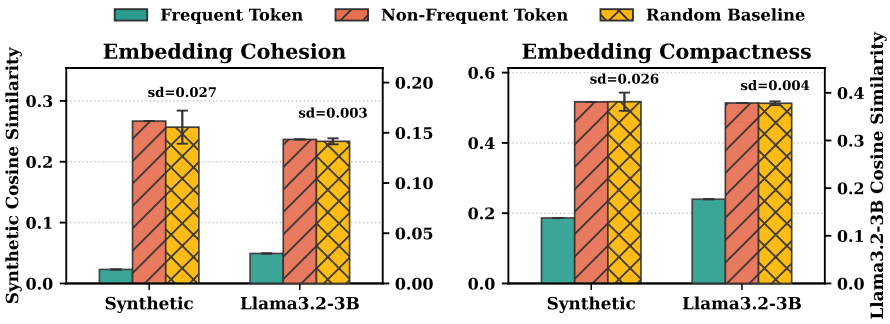



Figure 48: Frequent/non-frequent cohesion and compactness in the synthetic model and Llama3.2-3B.

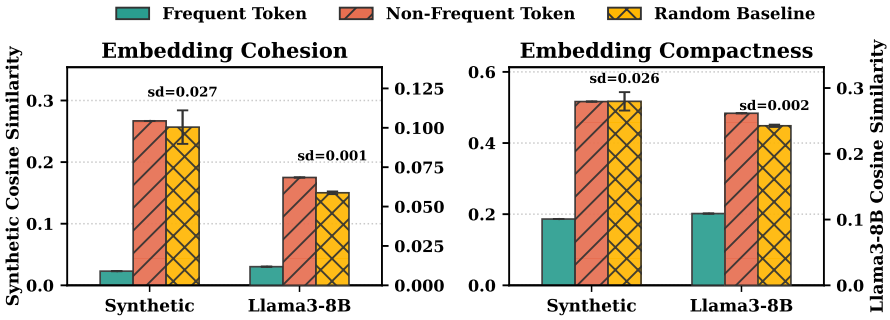



Figure 49: Frequent/non-frequent cohesion and compactness in the synthetic model and Llama3-8B.

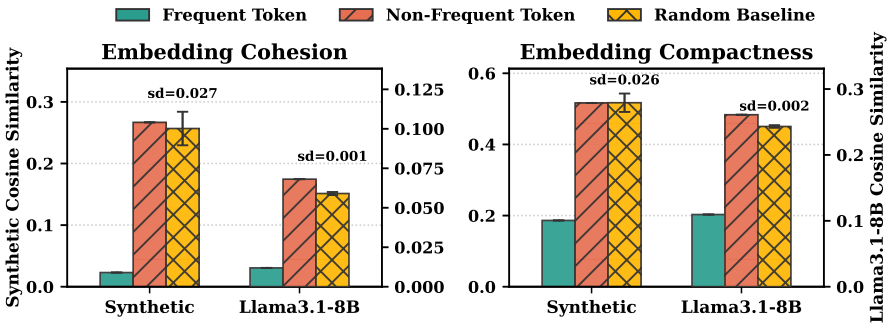



Figure 50: Frequent/non-frequent cohesion and compactness in the synthetic model and Llama3.1-8B.

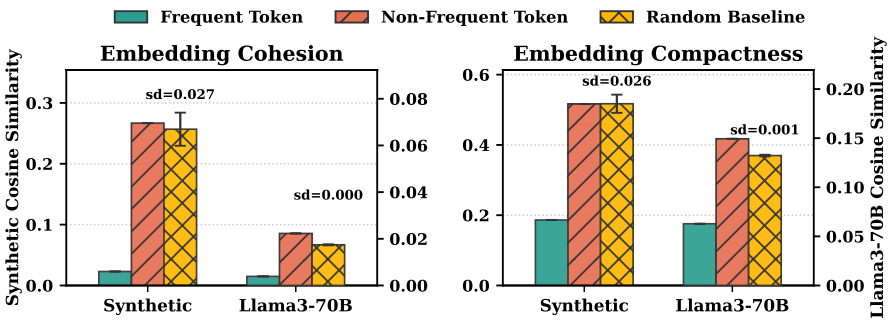



Figure 51: Frequent/non-frequent cohesion and compactness in the synthetic model and Llama3-70B.

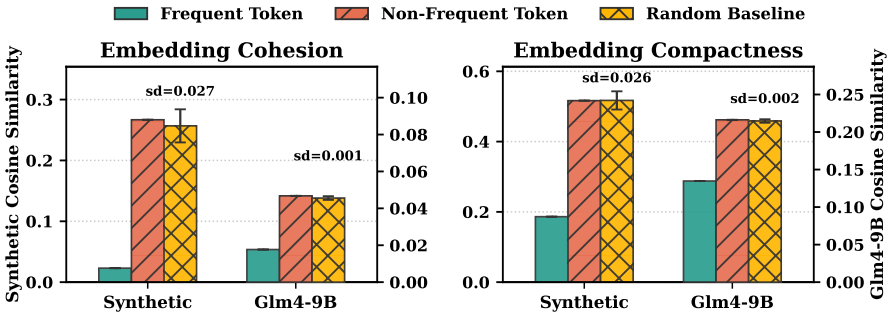



Figure 52: Frequent/non-frequent cohesion and compactness in the synthetic model and GLM4-9B.

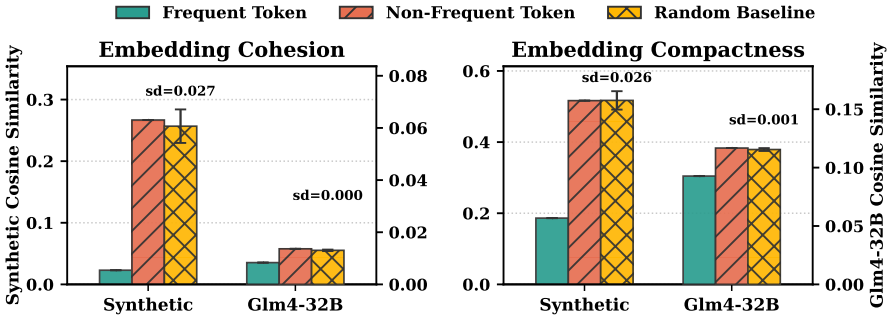



Figure 53: Frequent/non-frequent cohesion and compactness in the synthetic model and GLM4-32B.

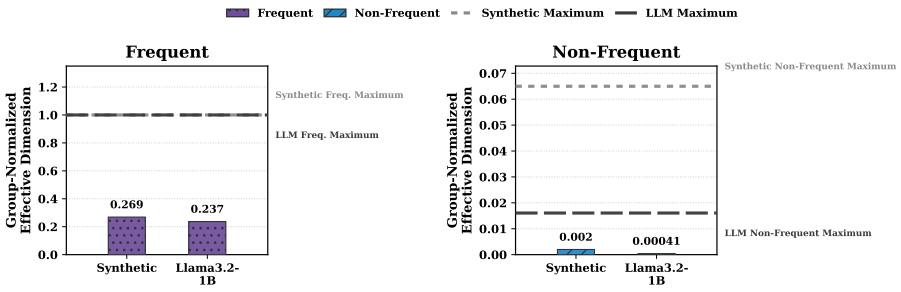



Figure 54: Frequent/non-frequent group-normalized effective dimension in the synthetic model and Llama3.2-1B.

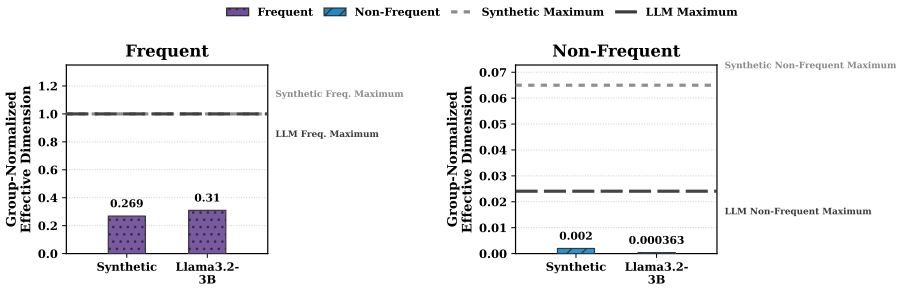



Figure 55: Frequent/non-frequent group-normalized effective dimension in the synthetic model and Llama3.2-3B.

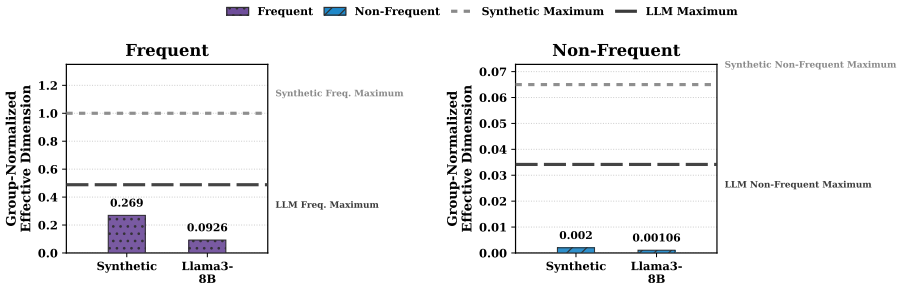



Figure 56: Frequent/non-frequent group-normalized effective dimension in the synthetic model and Llama3-8B.

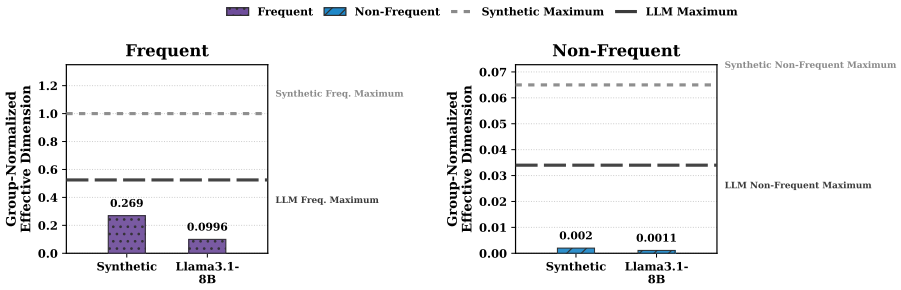



Figure 57: Frequent/non-frequent group-normalized effective dimension in the synthetic model and Llama3.1-8B.

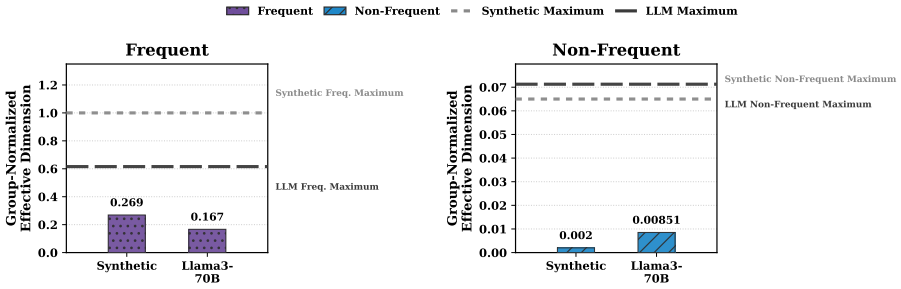



Figure 58: Frequent/non-frequent group-normalized effective dimension in the synthetic model and Llama3-70B.

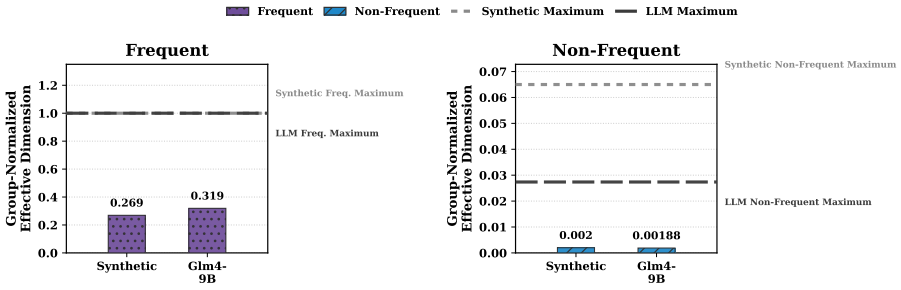



Figure 59: Frequent/non-frequent group-normalized effective dimension in the synthetic model and GLM4-9B.

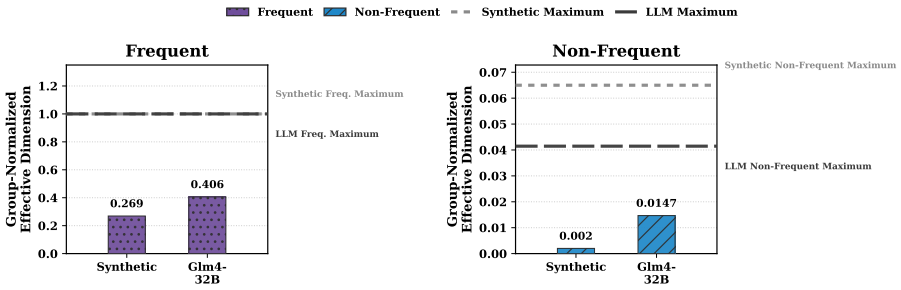



Figure 60: Frequent/non-frequent group-normalized effective dimension in the synthetic model and GLM4-32B.

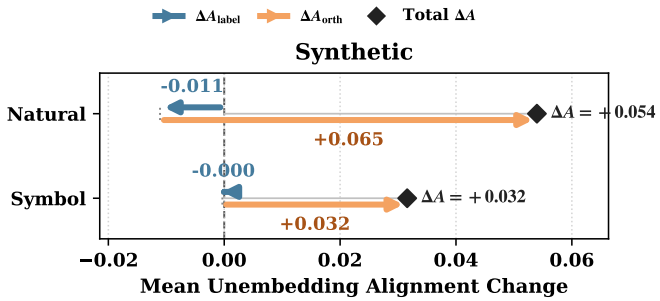



Figure 61: Exact mean-alignment decomposition under natural and symbol labels in the synthetic model.

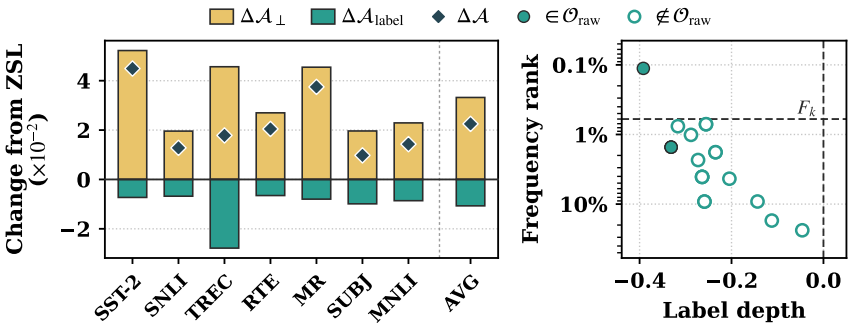



Figure 62: Per-task decomposition of the ICL alignment shift and label status, as in Figure 5, for Llama3.2-1B.

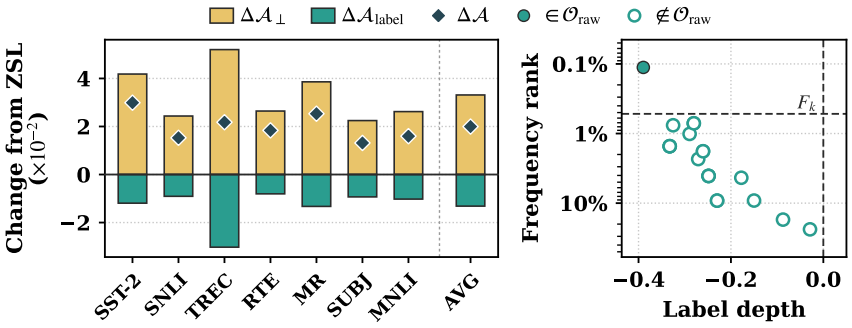



Figure 63: Per-task decomposition of the ICL alignment shift and label status, as in Figure 5, for Llama3.2-3B.

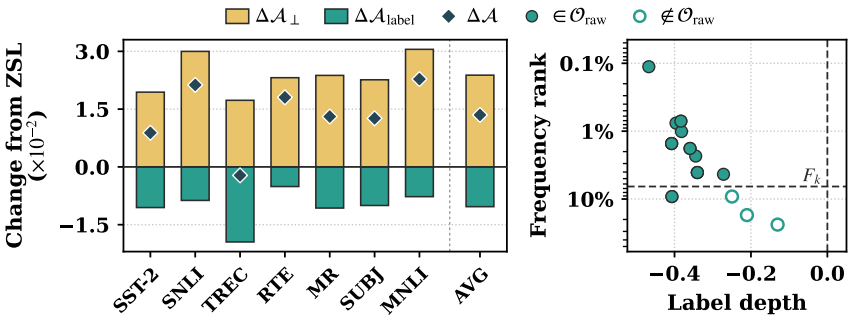



Figure 64: Per-task decomposition of the ICL alignment shift and label status, as in Figure 5, for Llama3-8B.

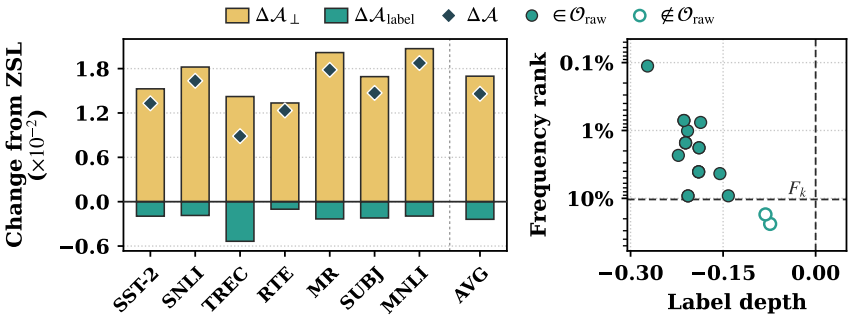



Figure 65: Per-task decomposition of the ICL alignment shift and label status, as in Figure 5, for Llama3-70B.

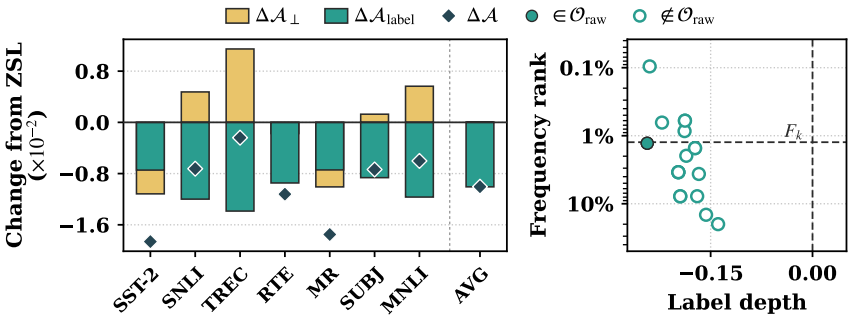



Figure 66: Per-task decomposition of the ICL alignment shift and label status, as in Figure 5, for GLM4-9B.

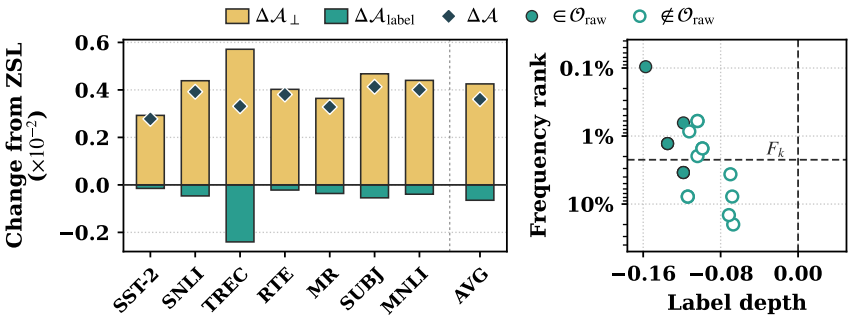



Figure 67: Per-task decomposition of the ICL alignment shift and label status, as in Figure 5, for GLM4-32B.

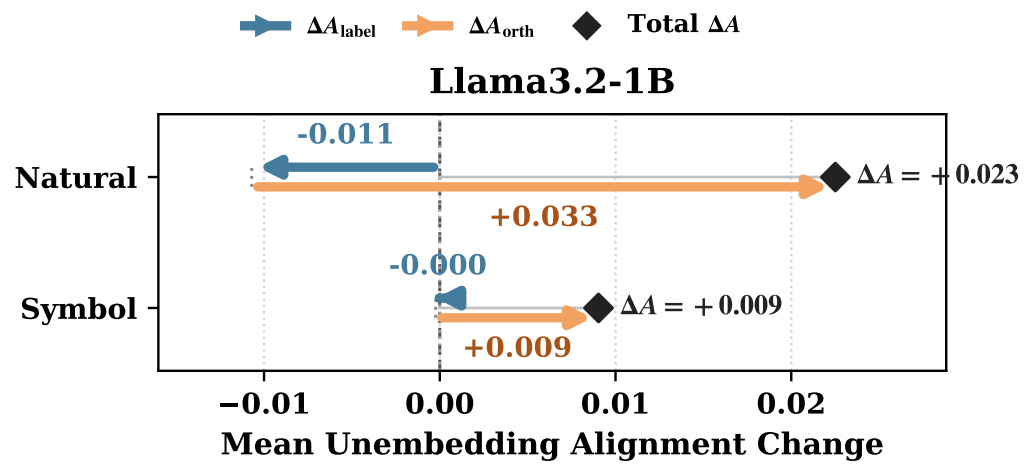



Figure 68: Exact mean-alignment decomposition under natural and symbol labels for Llama3.2-1B.

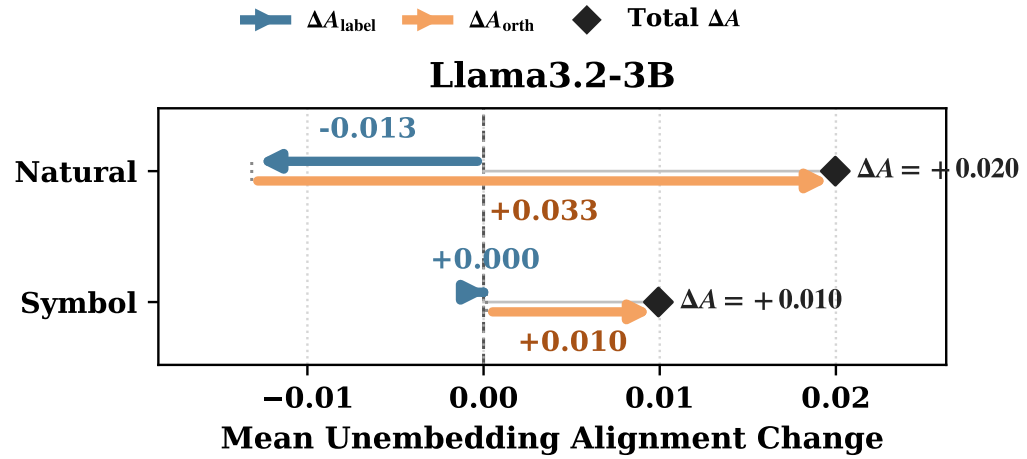



Figure 69: Exact mean-alignment decomposition under natural and symbol labels for Llama3.2-3B.

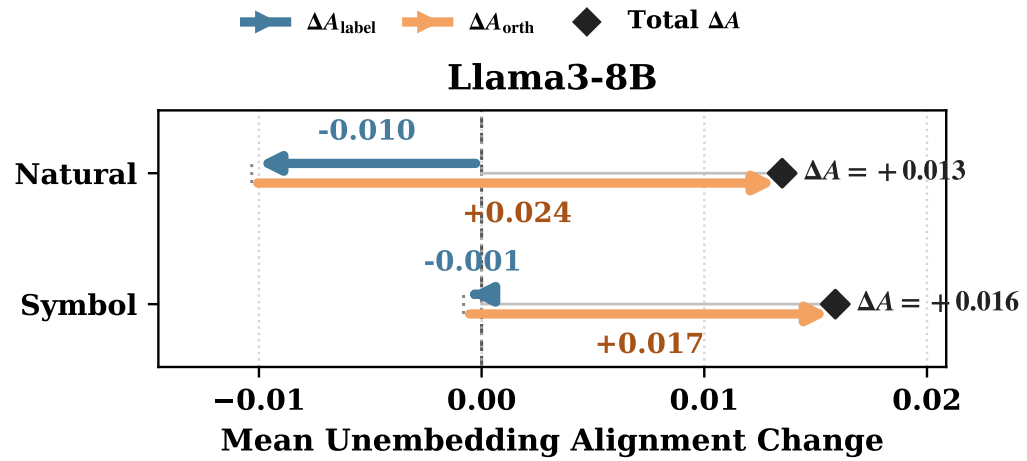



Figure 70: Exact mean-alignment decomposition under natural and symbol labels for Llama3-8B.

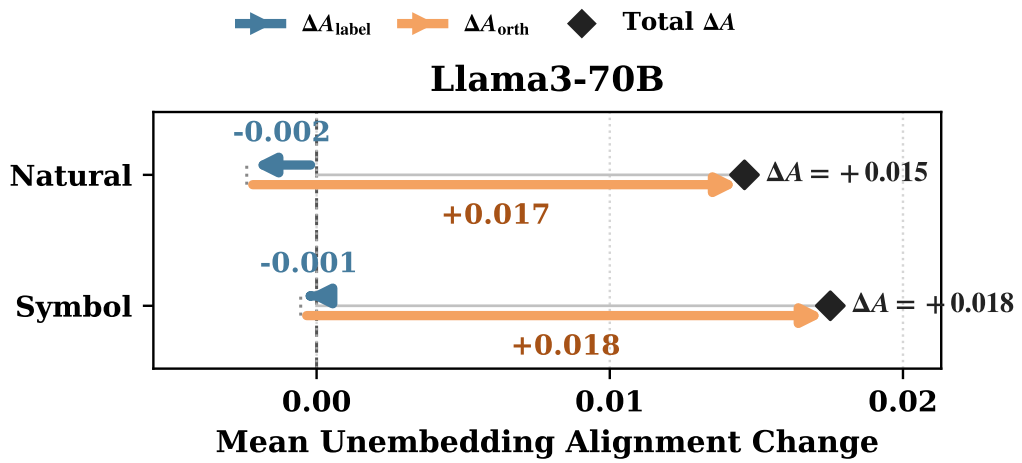



Figure 71: Exact mean-alignment decomposition under natural and symbol labels for Llama3-70B.

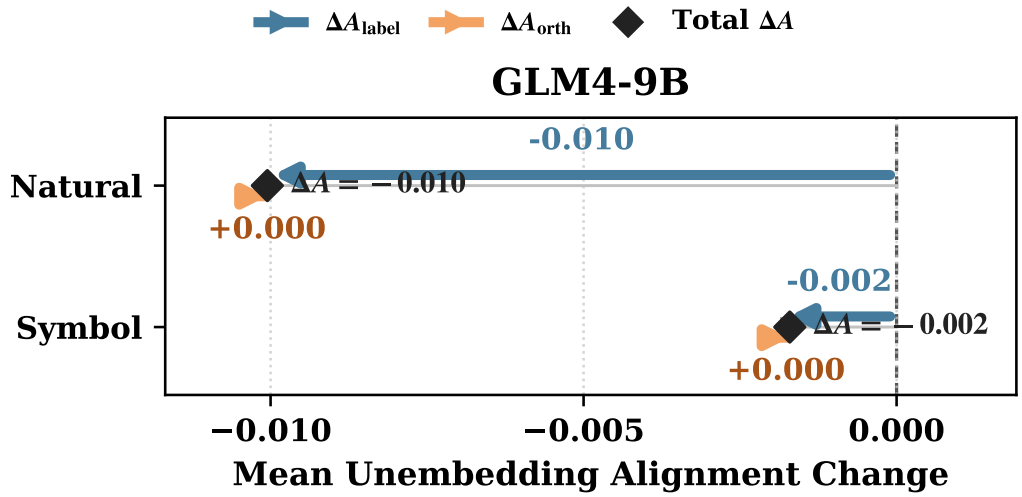



Figure 72: Exact mean-alignment decomposition under natural and symbol labels for GLM4-9B.

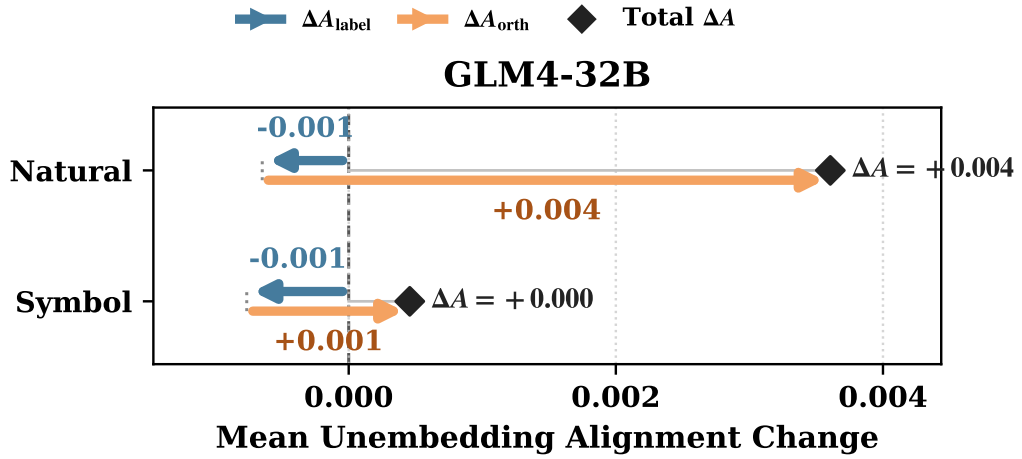



Figure 73: Exact mean-alignment decomposition under natural and symbol labels for GLM4-32B.

Table 15: Demonstration labels in the outlier set $\mathcal{O}_{\mathrm{raw}}$ and in the size-matched frequent set $F_k$, written as $\mathcal{O}_{\mathrm{raw}}$ / $F_k$ counts for every model, task, and label type. Column headers give the number of labels per task; the last column gives the share of the vocabulary in $\mathcal{O}_{\mathrm{raw}}$, which equals that in $F_k$.

| Model | Labels | SST-2 (2) | SNLI (3) | TREC (6) | RTE (2) | MR (2) | SUBJ (2) | MNLI (3) | Total (20) | $\lvert\mathcal{O}_{\mathrm{raw}}\rvert/V$ (%) |
|---|---|---|---|---|---|---|---|---|---|---|
| Llama3.1-8B | Natural | 2/2 | 3/2 | 5/4 | 2/2 | 2/2 | 1/1 | 3/2 | 18/15 | 6.1 |
| | Symbol | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | |
| Llama3.2-1B | Natural | 0/0 | 1/0 | 1/1 | 1/0 | 0/0 | 0/0 | 1/0 | 4/1 | 0.6 |
| | Symbol | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | |
| Llama3.2-3B | Natural | 0/0 | 0/0 | 1/1 | 0/0 | 0/0 | 0/0 | 0/0 | 1/1 | 0.5 |
| | Symbol | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | |
| Llama3-8B | Natural | 2/2 | 3/2 | 4/4 | 2/2 | 2/2 | 1/1 | 3/2 | 17/15 | 6.5 |
| | Symbol | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | |
| Llama3-70B | Natural | 2/2 | 3/3 | 5/5 | 2/2 | 2/2 | 1/1 | 3/3 | 18/18 | 10.4 |
| | Symbol | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | |
| GLM4-9B | Natural | 0/1 | 1/0 | 0/3 | 1/0 | 0/1 | 0/0 | 1/0 | 3/5 | 1.2 |
| | Symbol | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | |
| GLM4-32B | Natural | 0/2 | 2/1 | 2/4 | 2/1 | 0/2 | 0/0 | 2/1 | 8/11 | 2.2 |
| | Symbol | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | 0/0 | |

Table 16: Label depth $\cos(u_\ell - \bar{u}_{\bar{L}}, m_{\mathrm{unit}})$ for every model, task, and label type. Task columns average the depth over the task's labels; the last column gives its range over all 20 labels.

| Model | Labels | SST-2 | SNLI | TREC | RTE | MR | SUBJ | MNLI | Range over labels |
|---|---|---|---|---|---|---|---|---|---|
| Llama3.1-8B | Natural | $-0.37$ | $-0.38$ | $-0.33$ | $-0.37$ | $-0.37$ | $-0.23$ | $-0.38$ | $-0.47$ to $-0.13$ |
| | Symbol | $-0.06$ | $-0.01$ | $-0.01$ | $-0.06$ | $-0.06$ | $-0.06$ | $-0.01$ | $-0.13$ to $0.11$ |
| Llama3.2-1B | Natural | $-0.25$ | $-0.28$ | $-0.24$ | $-0.30$ | $-0.25$ | $-0.16$ | $-0.28$ | $-0.39$ to $-0.05$ |
| | Symbol | $-0.02$ | $0.00$ | $-0.01$ | $-0.02$ | $-0.02$ | $-0.02$ | $0.00$ | $-0.08$ to $0.04$ |
| Llama3.2-3B | Natural | $-0.27$ | $-0.27$ | $-0.24$ | $-0.29$ | $-0.27$ | $-0.13$ | $-0.27$ | $-0.39$ to $-0.03$ |
| | Symbol | $-0.02$ | $0.01$ | $-0.01$ | $-0.02$ | $-0.02$ | $-0.02$ | $0.01$ | $-0.08$ to $0.05$ |
| Llama3-8B | Natural | $-0.37$ | $-0.38$ | $-0.33$ | $-0.37$ | $-0.37$ | $-0.24$ | $-0.38$ | $-0.47$ to $-0.13$ |
| | Symbol | $-0.05$ | $0.01$ | $-0.01$ | $-0.05$ | $-0.05$ | $-0.05$ | $0.01$ | $-0.15$ to $0.12$ |
| Llama3-70B | Natural | $-0.20$ | $-0.20$ | $-0.18$ | $-0.20$ | $-0.20$ | $-0.12$ | $-0.20$ | $-0.27$ to $-0.07$ |
| | Symbol | $-0.04$ | $-0.01$ | $-0.02$ | $-0.04$ | $-0.04$ | $-0.04$ | $-0.01$ | $-0.08$ to $0.05$ |
| GLM4-9B | Natural | $-0.18$ | $-0.21$ | $-0.19$ | $-0.22$ | $-0.18$ | $-0.16$ | $-0.21$ | $-0.24$ to $-0.14$ |
| | Symbol | $-0.04$ | $-0.02$ | $-0.02$ | $-0.04$ | $-0.04$ | $-0.04$ | $-0.02$ | $-0.06$ to $0.07$ |
| GLM4-32B | Natural | $-0.10$ | $-0.12$ | $-0.10$ | $-0.13$ | $-0.10$ | $-0.07$ | $-0.12$ | $-0.16$ to $-0.07$ |
| | Symbol | $-0.04$ | $-0.03$ | $-0.04$ | $-0.04$ | $-0.04$ | $-0.04$ | $-0.03$ | $-0.07$ to $-0.01$ |

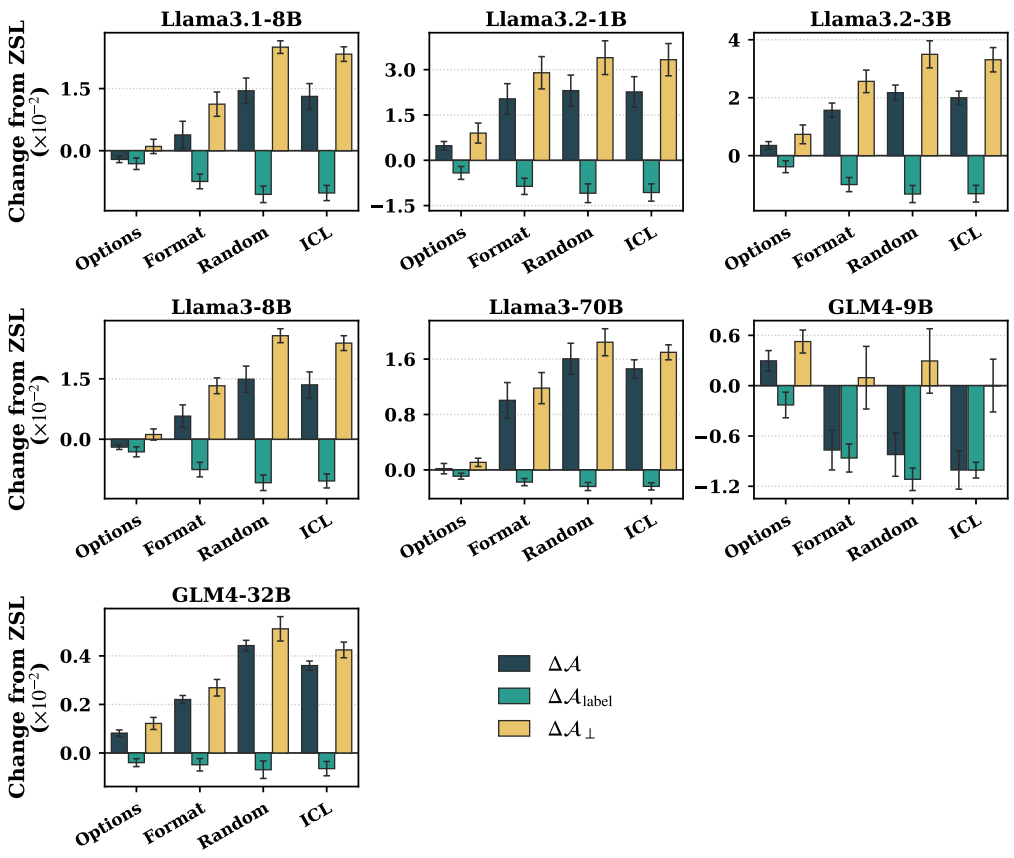



Figure 74: Change of mean alignment from ZSL and its label and orthogonal parts under the three prompt controls (Options: label options only; Format: label-format records without inputs; Random: demonstrations with random labels) and 10-shot ICL, averaged over tasks with ±1 SEM.

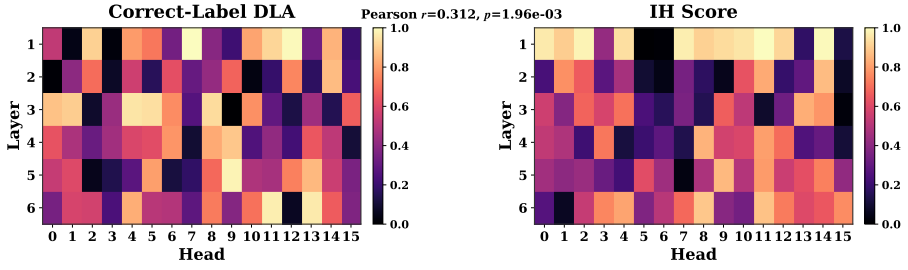



Figure 75: Correct-label DLA and induction-head scores for all heads in the synthetic model. Scores are displayed by layer and head after separate percentile-rank normalization; the annotation reports their Pearson correlation before normalization.

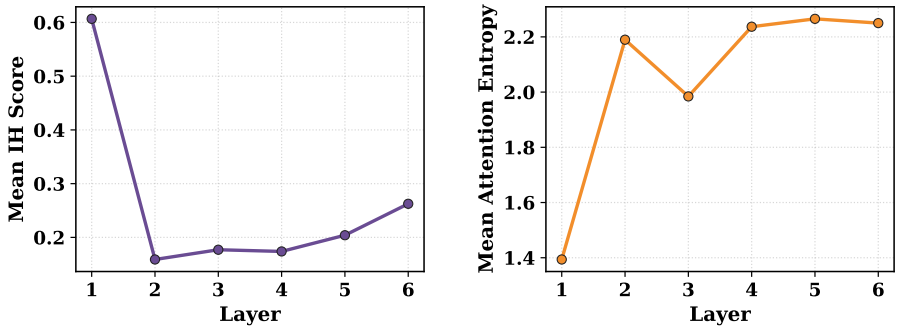



Figure 76: Layerwise attention signatures in the synthetic model. Left: IH score averaged across the 16 heads in each layer. Right: mean Shannon entropy of the final-position attention distribution, averaged across prompts and heads.

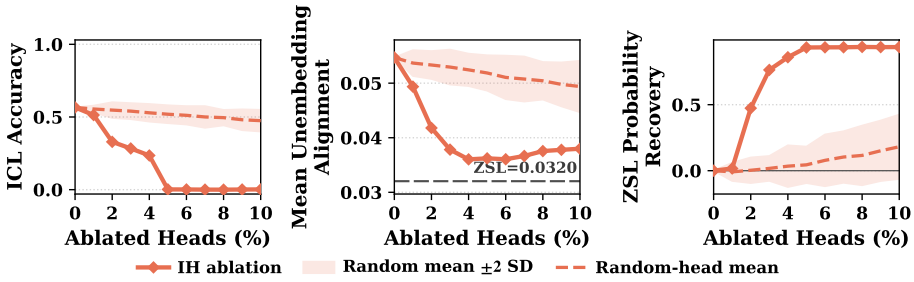



Figure 77: Targeted-IH and random-head ablation dose sweep for Llama3.2-1B, aggregated over seven datasets. Random curves and bands use ten equal-size random-head sets.

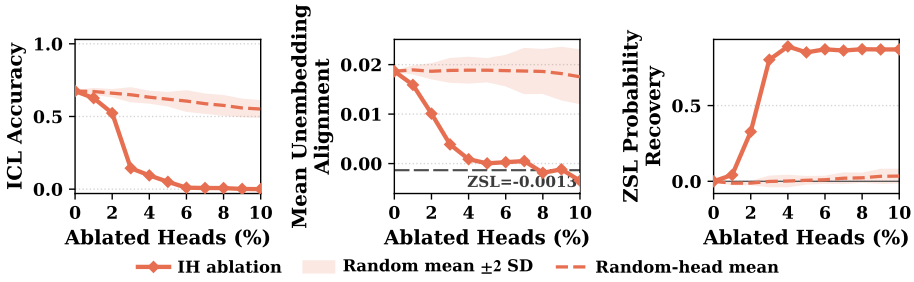



Figure 78: Targeted-IH and random-head ablation dose sweep for Llama3.2-3B, aggregated over seven datasets. Random curves and bands use ten equal-size random-head sets.

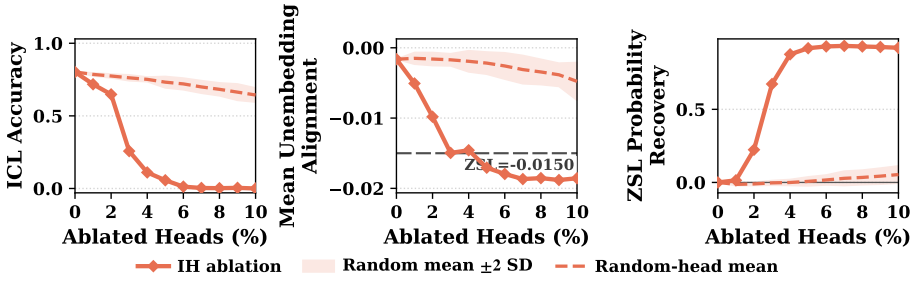



Figure 79: Targeted-IH and random-head ablation dose sweep for Llama3-8B, aggregated over seven datasets. Random curves and bands use ten equal-size random-head sets.

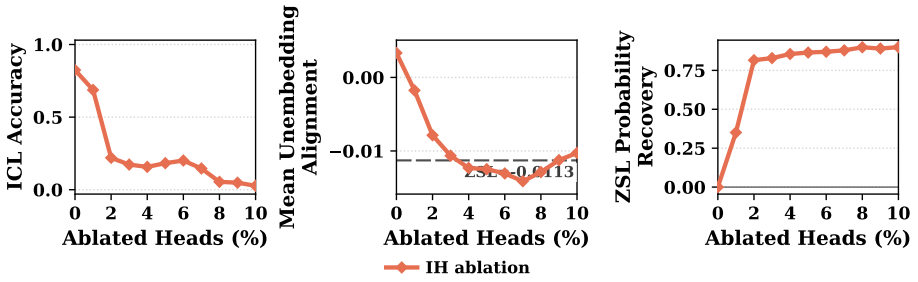



Figure 80: Targeted-IH ablation dose sweep for Llama3-70B, aggregated over seven datasets.

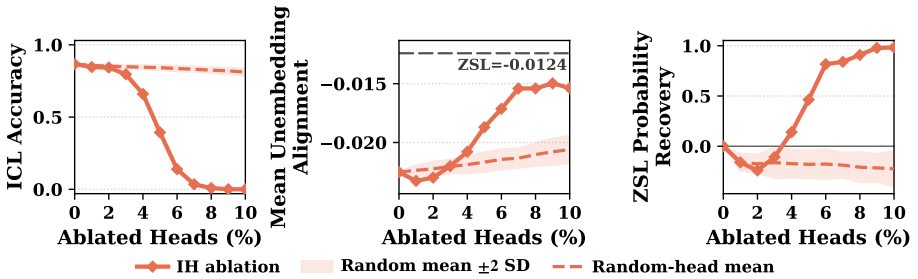



Figure 81: Targeted-IH and random-head ablation dose sweep for GLM4-9B, aggregated over seven datasets. Random curves and bands use ten equal-size random-head sets.

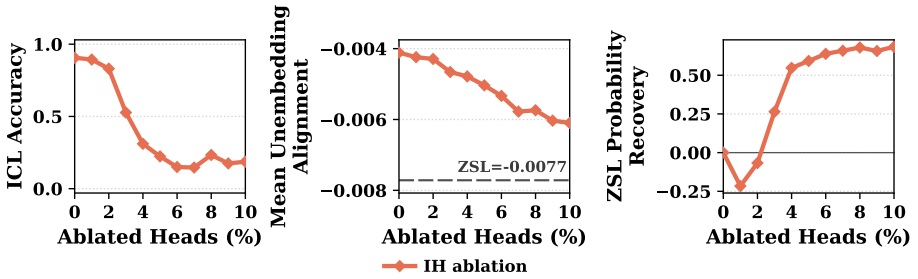



Figure 82: Targeted-IH ablation dose sweep for GLM4-32B, aggregated over seven datasets.

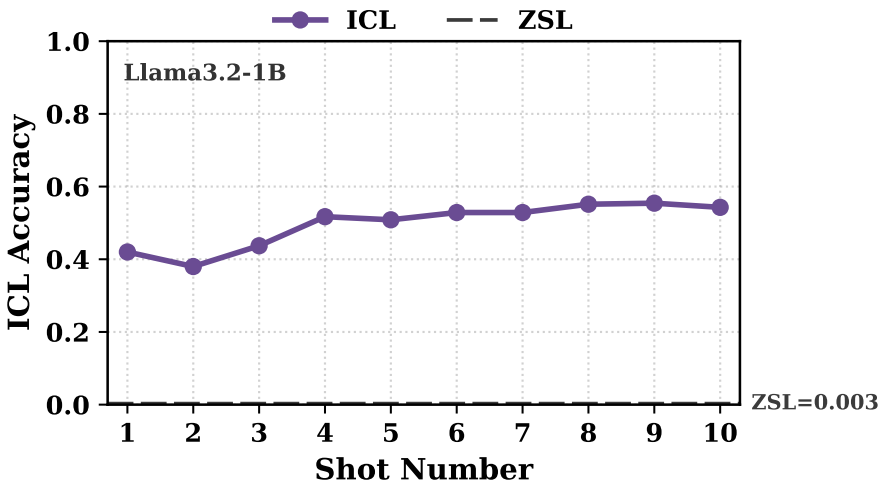



Figure 83: One-to-ten-shot ICL accuracy and the ZSL reference for Llama3.2-1B.

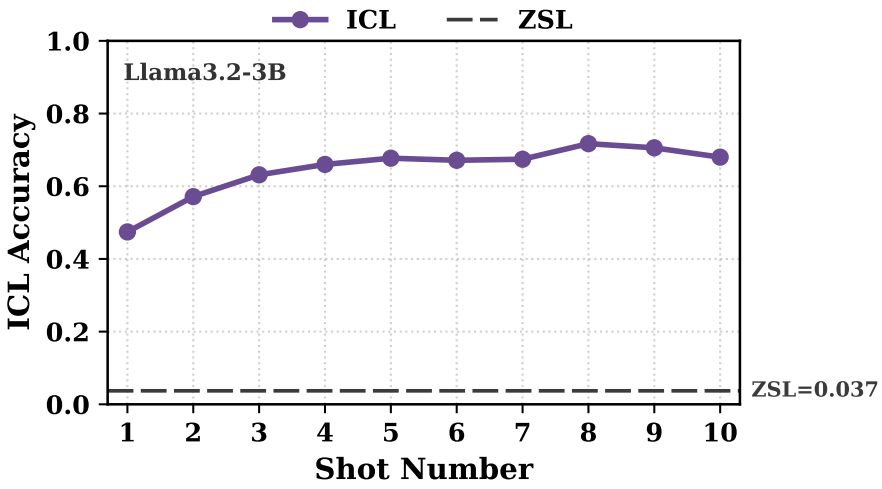



Figure 84: One-to-ten-shot ICL accuracy and the ZSL reference for Llama3.2-3B.

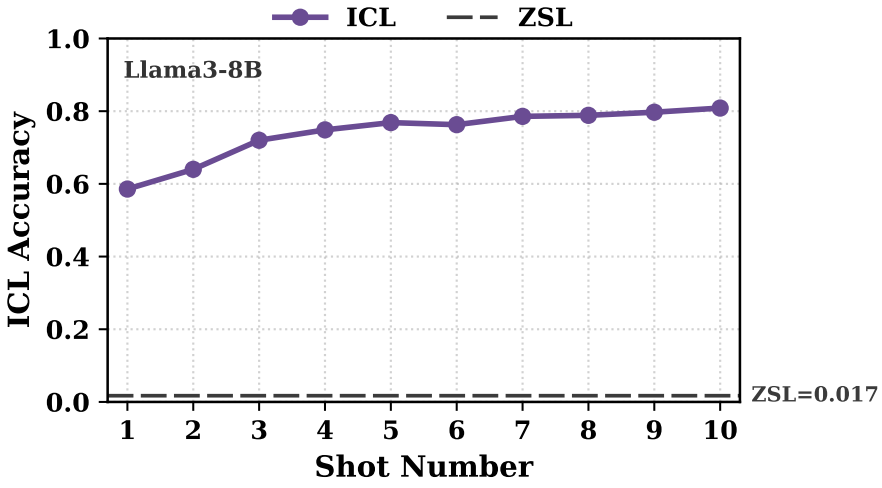



Figure 85: One-to-ten-shot ICL accuracy and the ZSL reference for Llama3-8B.

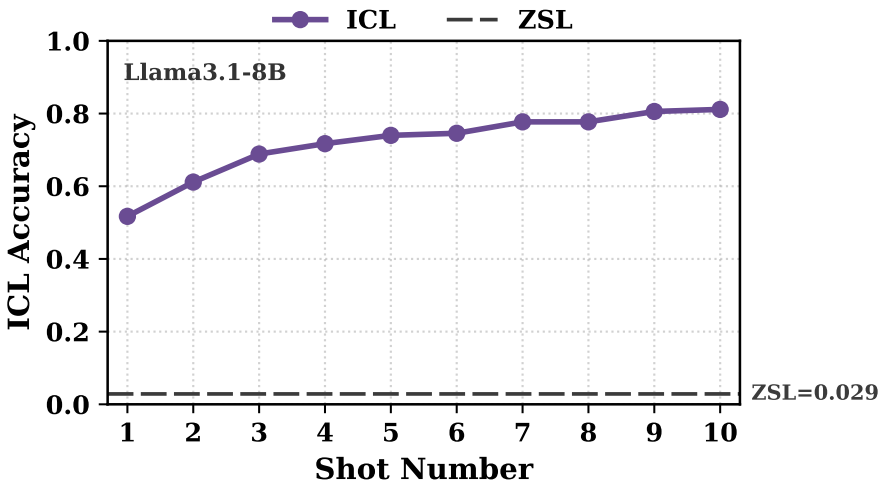



Figure 86: One-to-ten-shot ICL accuracy and the ZSL reference for Llama3.1-8B.

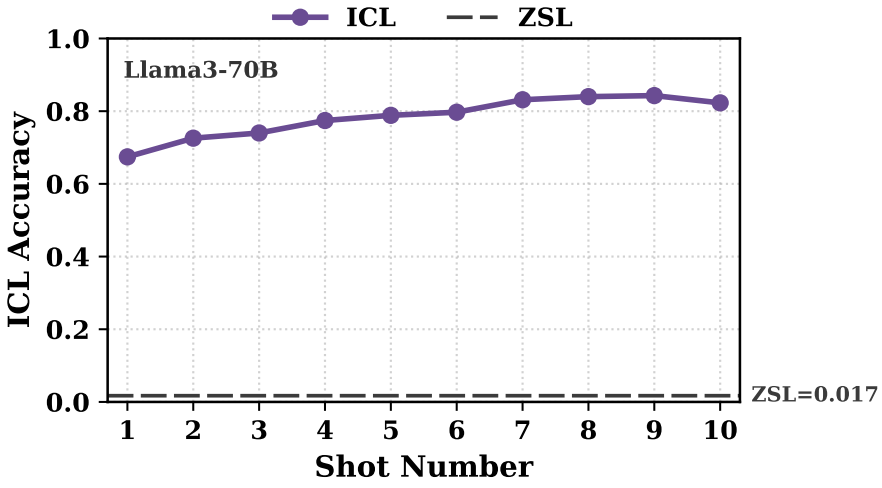



Figure 87: One-to-ten-shot ICL accuracy and the ZSL reference for Llama3-70B.

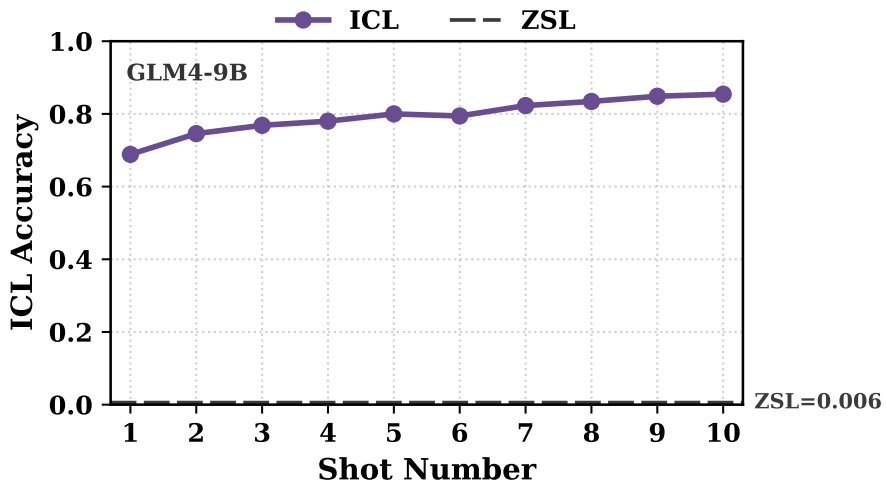



Figure 88: One-to-ten-shot ICL accuracy and the ZSL reference for GLM4-9B.

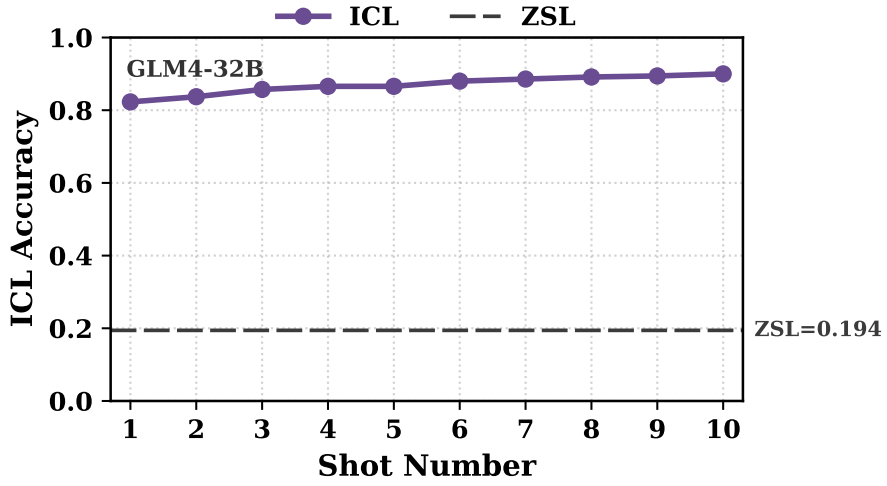



Figure 89: One-to-ten-shot ICL accuracy and the ZSL reference for GLM4-32B.

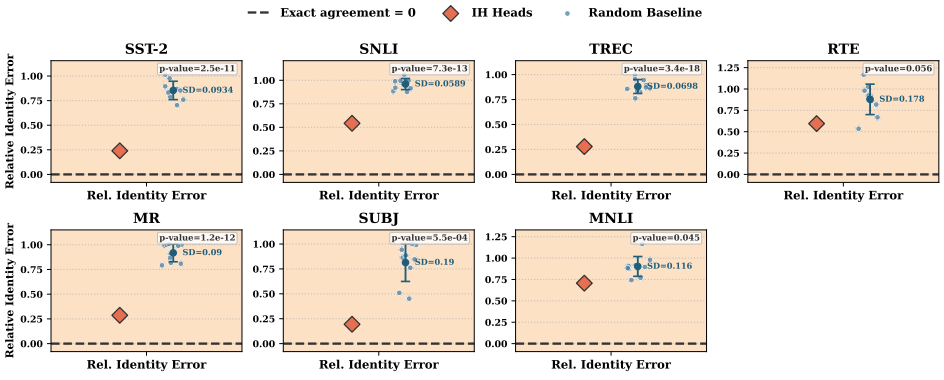



Figure 90: Relative identity error between selected-IH attribution and the ICL–ZSL correct-label log-probability difference for Llama3.1-8B.

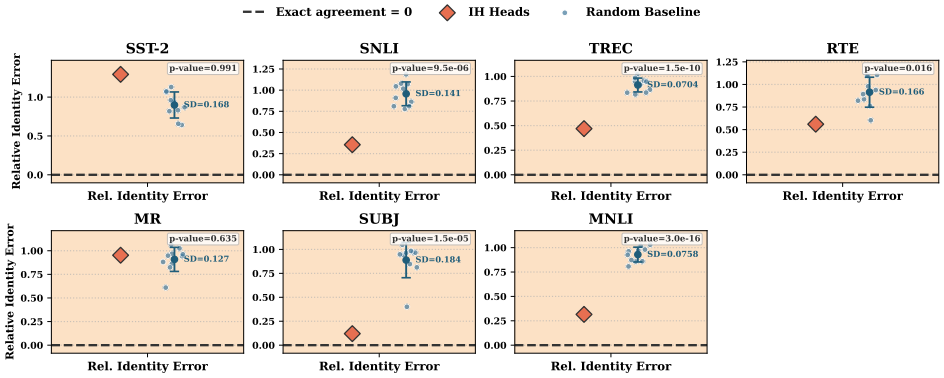



Figure 91: Relative identity error between selected-IH attribution and the ICL–ZSL correct-label log-probability difference for Llama3.2-1B.

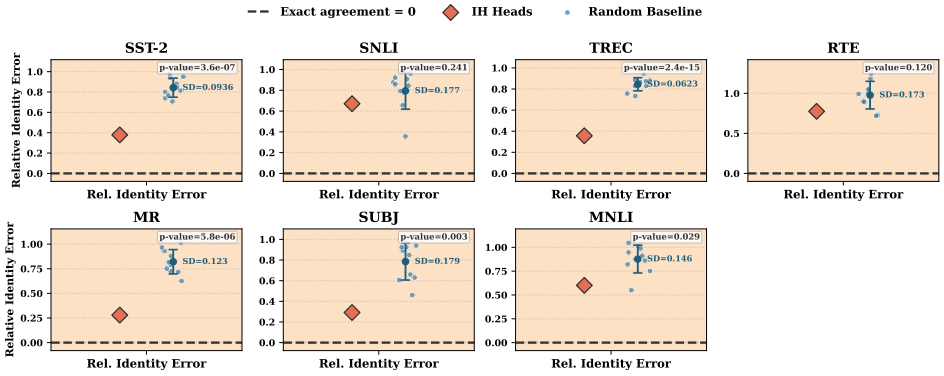



Figure 92: Relative identity error between selected-IH attribution and the ICL–ZSL correct-label log-probability difference for Llama3.2-3B.

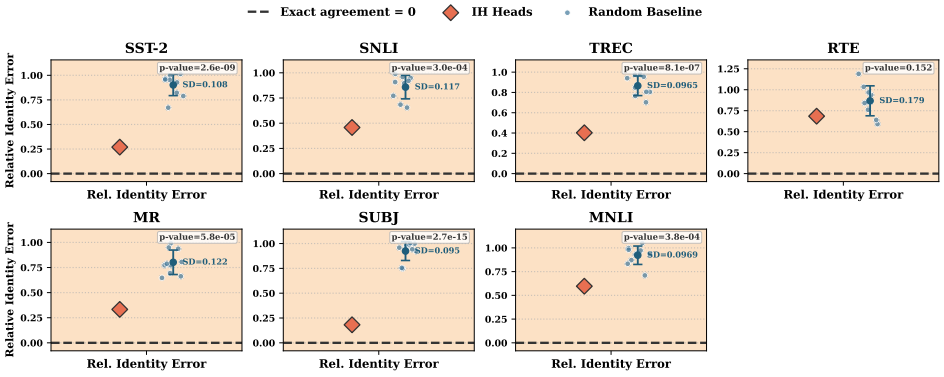



Figure 93: Relative identity error between selected-IH attribution and the ICL–ZSL correct-label log-probability difference for Llama3-8B.

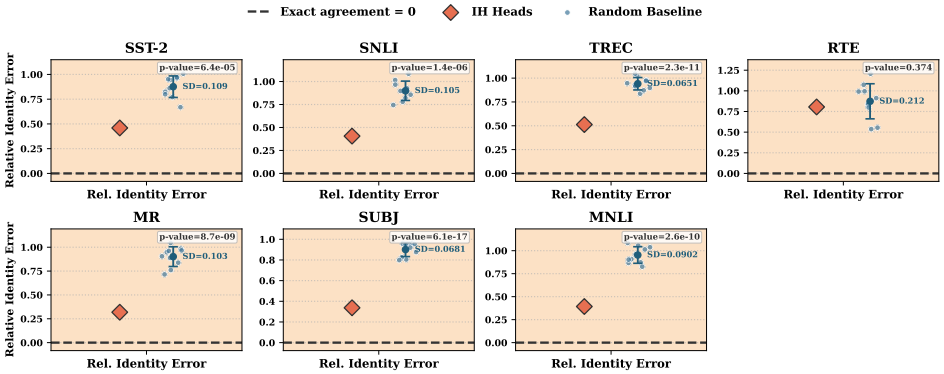



Figure 94: Relative identity error between selected-IH attribution and the ICL–ZSL correct-label log-probability difference for Llama3-70B.

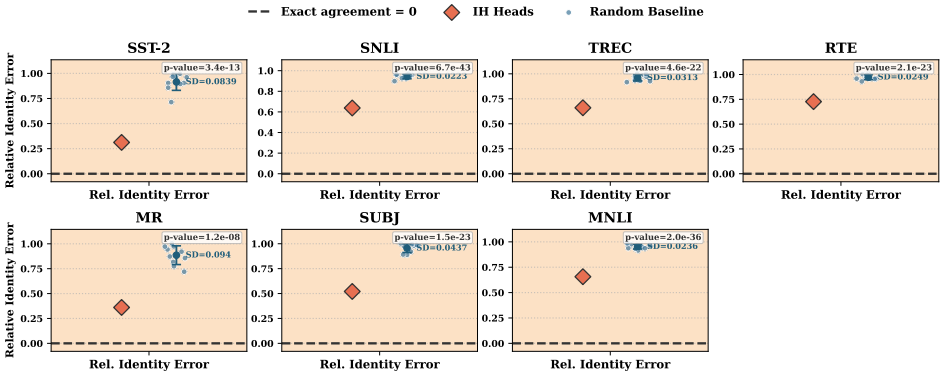



Figure 95: Relative identity error between selected-IH attribution and the ICL–ZSL correct-label log-probability difference for GLM4-9B.

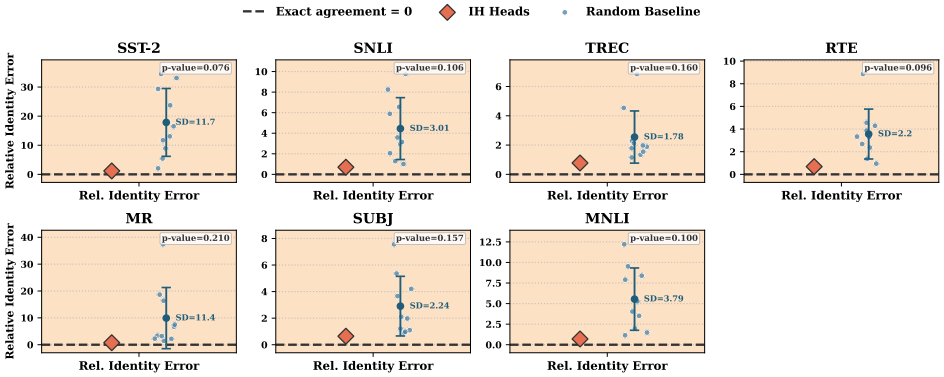



Figure 96: Relative identity error between selected-IH attribution and the ICL–ZSL correct-label log-probability difference for GLM4-32B.

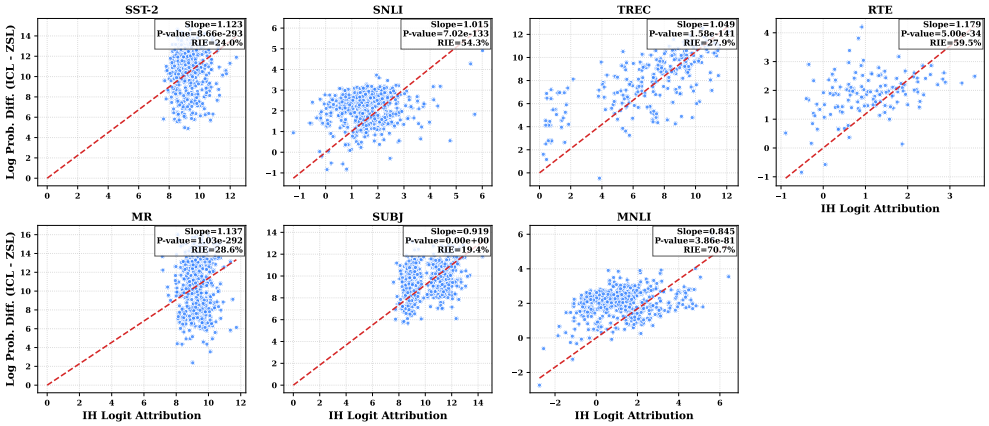



Figure 97: Prompt-level IH attribution versus probability change across all seven datasets for Llama3.1-8B.

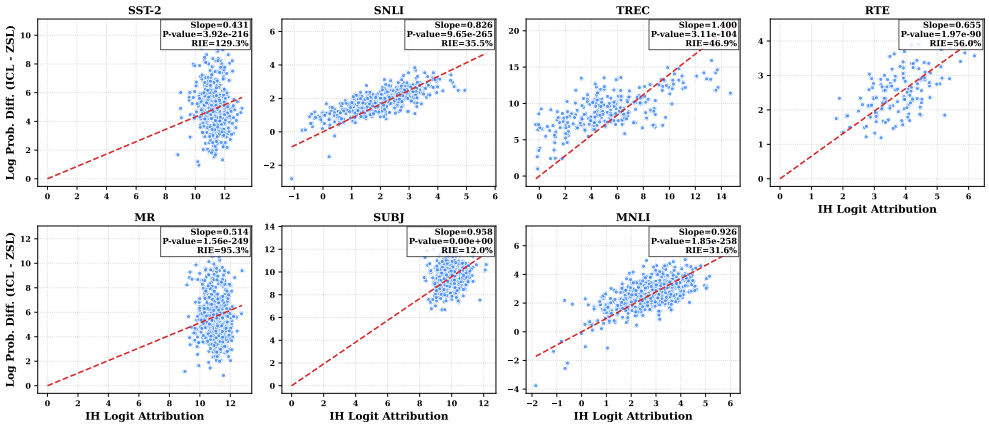



Figure 98: Prompt-level IH attribution versus probability change across datasets for Llama3.2-1B.

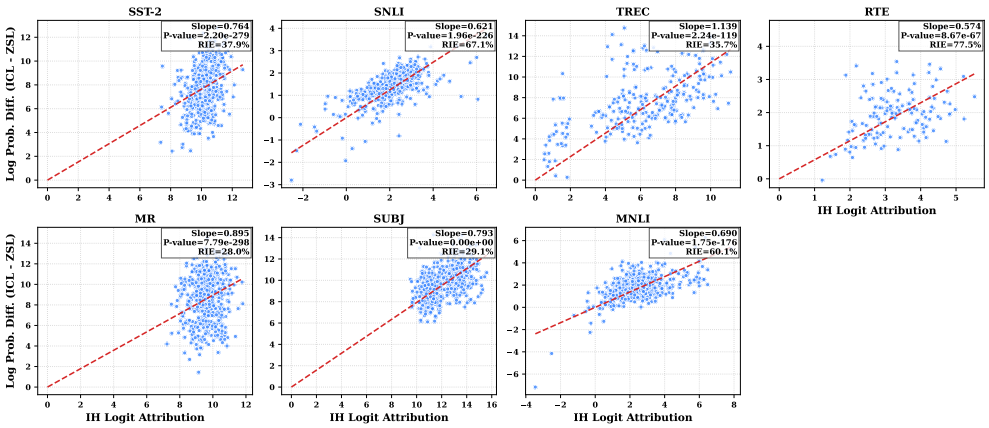



Figure 99: Prompt-level IH attribution versus probability change across datasets for Llama3.2-3B.

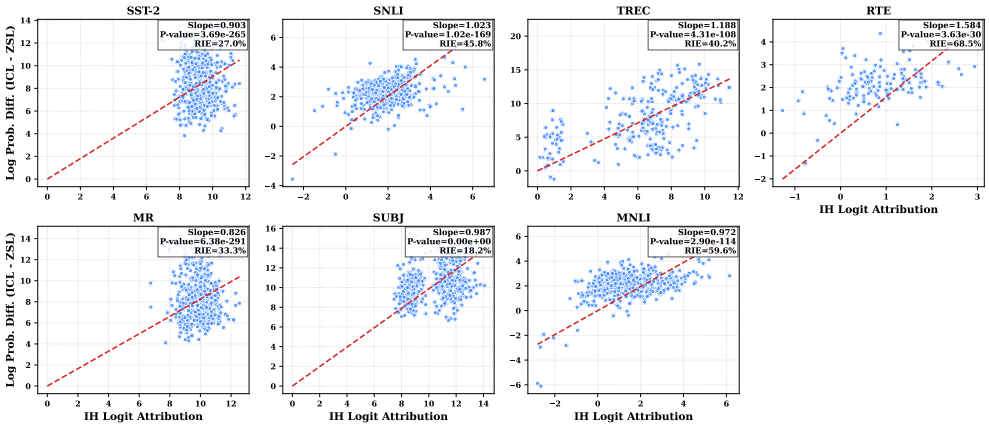



Figure 100: Prompt-level IH attribution versus probability change across datasets for Llama3-8B.

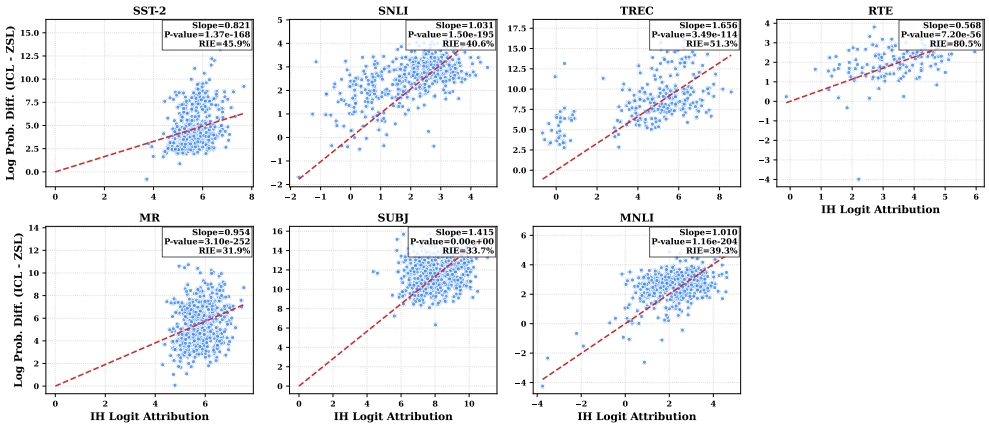



Figure 101: Prompt-level IH attribution versus probability change across datasets for Llama3-70B.

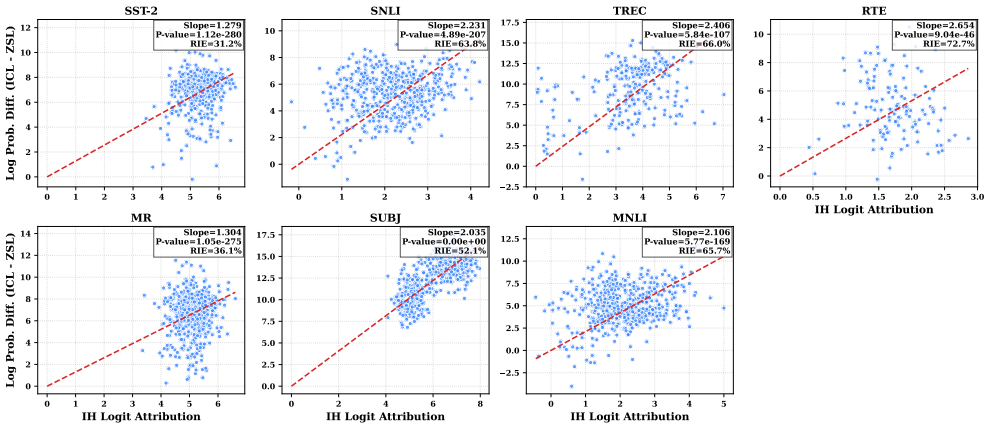



Figure 102: Prompt-level IH attribution versus probability change across datasets for GLM4-9B.

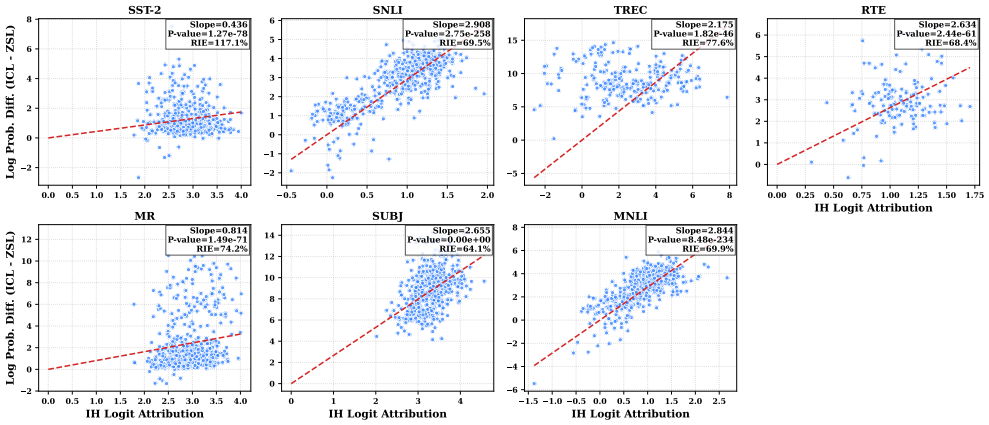



Figure 103: Prompt-level IH attribution versus probability change across datasets for GLM4-32B.

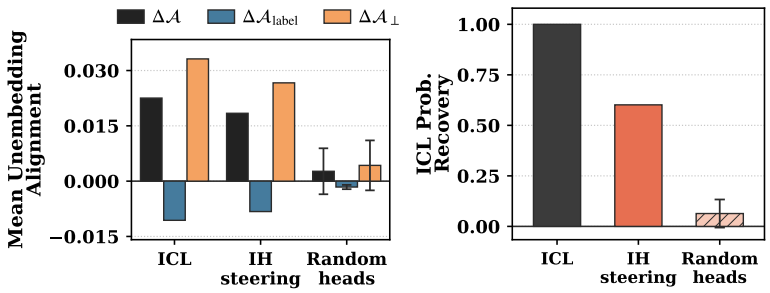



Figure 104: IH steering of held-out ZSL prompts in Llama3.2-1B, averaged over datasets. Left: mean-alignment shift from ZSL and its label and orthogonal parts. Right: ICL probability recovery. Random heads: mean ±2 SD over three (left) and ten (right) size-matched head sets.

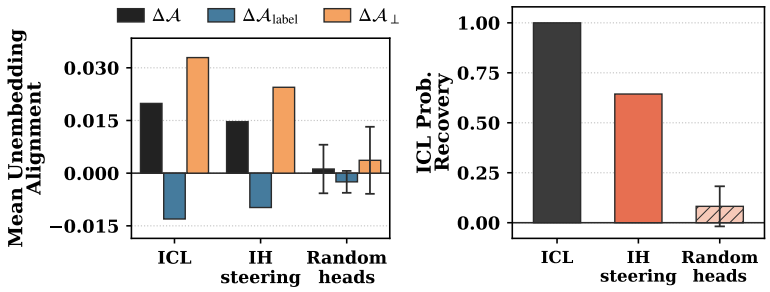



Figure 105: IH steering of held-out ZSL prompts in Llama3.2-3B, averaged over datasets. Left: mean-alignment shift from ZSL and its label and orthogonal parts. Right: ICL probability recovery. Random heads: mean ±2 SD over three (left) and ten (right) size-matched head sets.

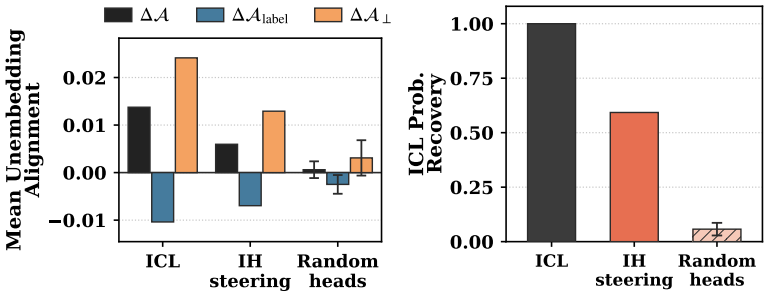



Figure 106: IH steering of held-out ZSL prompts in Llama3-8B, averaged over datasets. Left: mean-alignment shift from ZSL and its label and orthogonal parts. Right: ICL probability recovery. Random heads: mean ±2 SD over three (left) and ten (right) size-matched head sets.

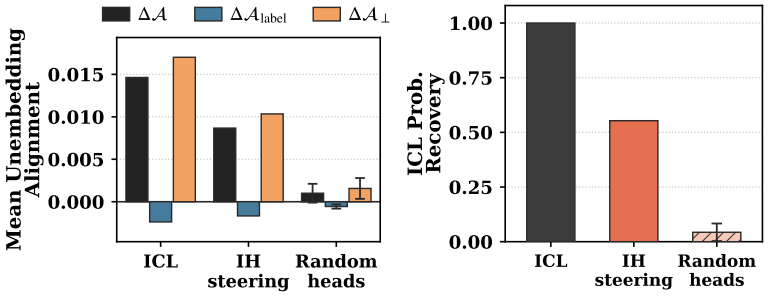



Figure 107: IH steering of held-out ZSL prompts in Llama3-70B, averaged over datasets. Left: mean-alignment shift from ZSL and its label and orthogonal parts. Right: ICL probability recovery. Random heads: mean ±2 SD over three (left) and ten (right) size-matched head sets.

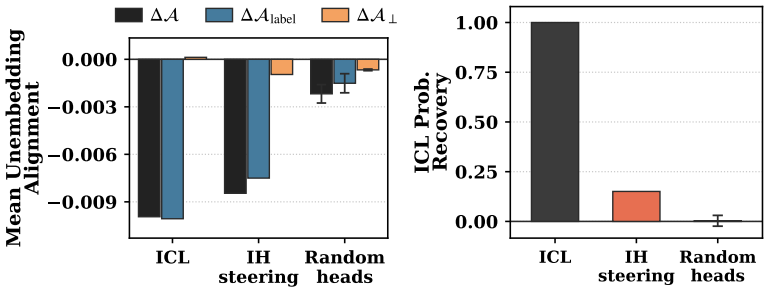



Figure 108: IH steering of held-out ZSL prompts in GLM4-9B, averaged over datasets. Left: mean-alignment shift from ZSL and its label and orthogonal parts. Right: ICL probability recovery. Random heads: mean ±2 SD over three (left) and ten (right) size-matched head sets.

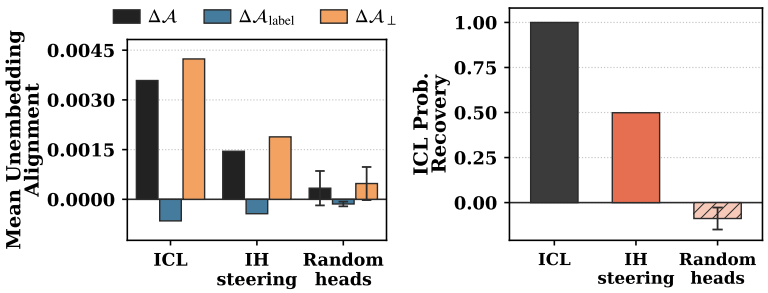



Figure 109: IH steering of held-out ZSL prompts in GLM4-32B, averaged over datasets. Left: mean-alignment shift from ZSL and its label and orthogonal parts. Right: ICL probability recovery. Random heads: mean ±2 SD over three (left) and ten (right) size-matched head sets.

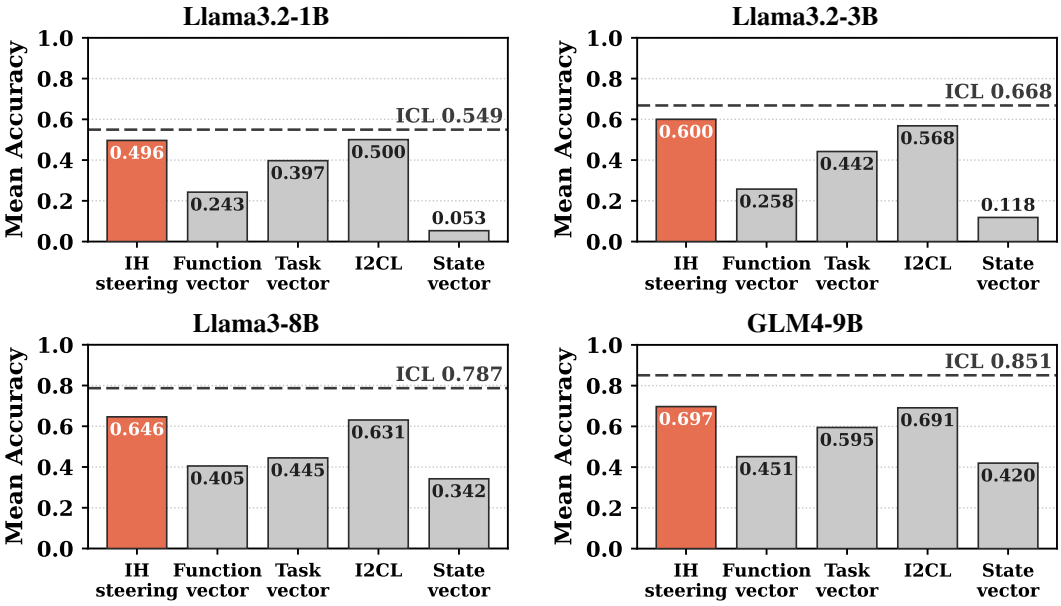



Figure 110: Accuracy of steering methods on four other LLMs, averaged over seven tasks; the dashed line marks 10-shot ICL.

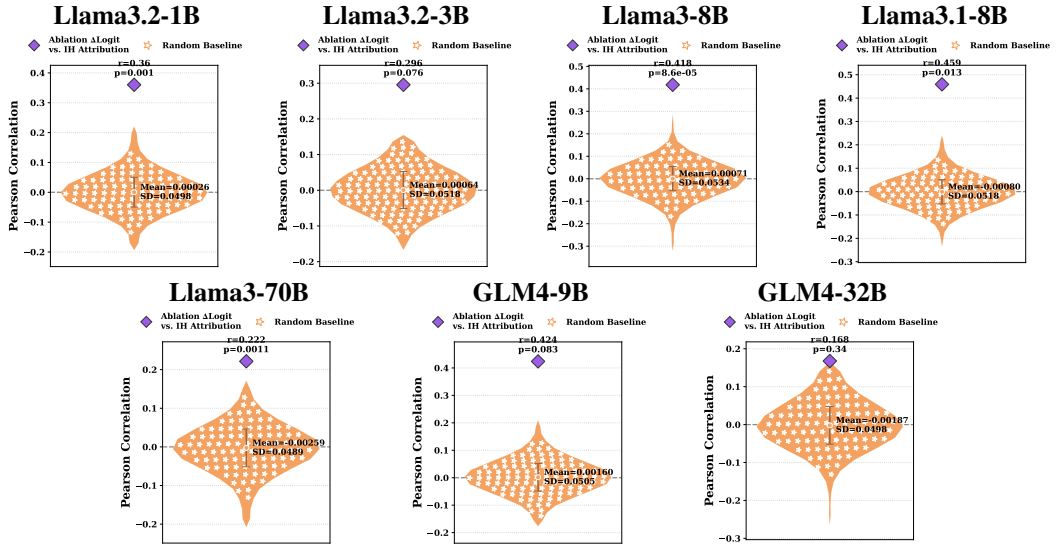



Figure 111: Cross-task correlation between induction-head attribution and ablation effects across all seven LLMs. Diamonds show the selected-head correlation; the orange distributions show matched random-head baselines.

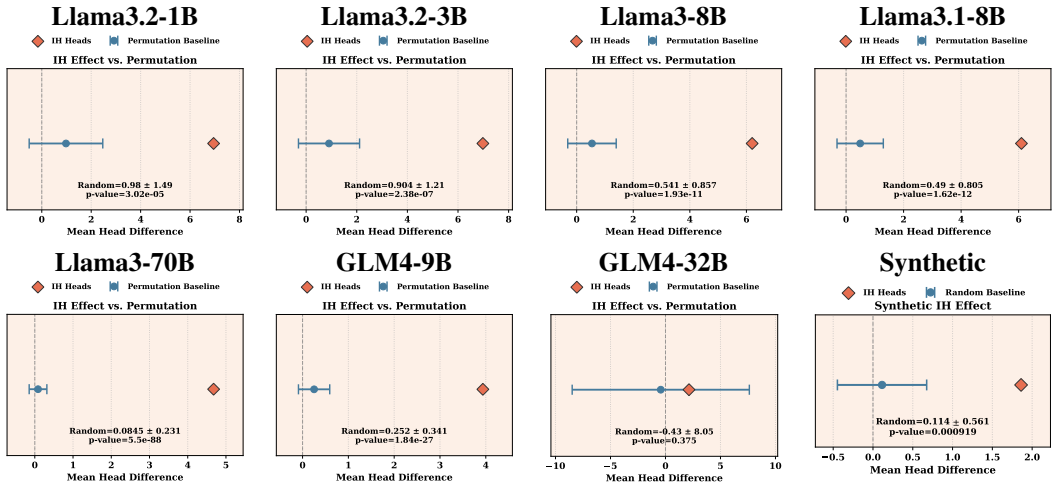



Figure 112: Significance of selected induction heads against random-head baselines across all seven LLMs and the synthetic model. Each panel compares the selected-head ICL–ZSL attribution with its permutation reference.