IN-CONTEXT LEARNING OPERATES AS CONCEPT SUBSPACE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Regression and Bayesian accounts of in-context learning (ICL) explain how demonstrations can induce predictors, while mechanistic analyses often identify compact activation directions that steer prompted behavior. However, it remains unclear whether structured demonstrations induce low-dimensional concept inference. We study this question through a concept-subspace view of ICL, in which tasks vary only along intrinsic concept coordinates, although inputs are observed in a high-dimensional ambient space. For ridge and least-squares ICL proxies, prediction decomposes exactly into concept-coordinate regression and off-subspace leakage. We show that an idealized single-layer linear attention head optimal for this task family instead stores the concept subspace in its weights and reads the prompt only through it, so patching its written subspace carries the head's entire pooled clean–corrupted effect. Testing this prediction in Llama-3-8B on CounterFact-derived multi-relation prompts, we find that a 73-dimensional subspace of the 4096-dimensional residual stream restores 78.8% of the clean–corrupted accuracy gap, whereas patching the complementary subspace restores 0%. Concept swaps redirect predictions toward injected relations, while random and cross-task matched-rank controls are largely ineffective. Additional experiments on Qwen2.5-7B, a controlled cross-lingual rule task, context-defined nonce mappings, and two-step compositional rules show the same qualitative pattern. These results support concept subspaces as compact, task-aligned mediators of recoverable ICL behavior in structured task families, without implying full-circuit recovery. Our code is available at the following anonymous repository `https://anonymous.4open.science/r/ICL-Concept-Subspace`.

1 INTRODUCTION

In-context learning (ICL) enables a pretrained language model to infer a new task from a few demonstrations without updating its weights (Brown et al., 2020; Wu & Yang, 2024; Wei et al., 2023). Demonstrations do more than provide supervised examples: they can specify label spaces, input distributions, output formats, and semantic priors (Min et al., 2022; Chan et al., 2022). Theoretical work has shown that, in controlled settings, ICL can implement least-squares or ridge regression, Bayesian inference, gradient descent, and algorithm selection (Xie et al., 2022; Bai et al., 2023). These accounts explain how a predictor can be fit from a prompt, but they often treat the inferred task as an unconstrained vector in a high-dimensional ambient space.

This issue is most salient when prompts share latent structure. Many demonstrations may instantiate a shared relation, rule, transformation, or semantic attribute (Pan et al., 2023). In such cases, treating the task as an unconstrained ambient parameter may obscure the relevant geometry. This raises a fundamental structural question: *When tasks share a low-dimensional conceptual structure, does ICL operate by explicitly inferring that low-dimensional concept?* This question is related to recent evidence that transformer representations can contain disentangled latent concepts or language-agnostic concept directions (Hong et al., 2026; Dumas et al., 2025). If ICL exploits shared task geometry, estimation should scale with the concept dimension, while orthogonal perturbations act mainly as nuisances. Existing regression theories, activation-space analyses, and causal-intervention tools each support parts of this view (Huben et al., 2024), but do not identify a common latent object induced by structured ICL prompts.

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.

To make this question precise, we introduce *concept-subspace ICL*. In our model, task parameters satisfy $w = U\beta$, where $U \in \mathbb{R}^{d\times r}$ has orthonormal columns and $r \ll d$. Inputs live in the ambient space, but the task depends on them only through concept coordinates $z = U^\top x$. For ridge and least-squares ICL proxies, ambient prediction decomposes into regression over the r-dimensional concept coordinates plus an off-subspace leakage term. Under block-diagonal or near-block-diagonal covariance assumptions, the leading estimation term depends on r, while residual error is controlled by cross-subspace coupling. The same formulation also admits a Bayesian interpretation as posterior inference over β and identifies the concept subspace up to rotation (Stuehmer et al., 2020).

In this paper, we show that a cost in d is intrinsic to not knowing the subspace: for isotropic inputs, every subspace-agnostic predictor pays for d on average. An idealized single-layer linear attention head that is optimal for the task family instead stores the subspace in its weights, so it reads the prompt only through this subspace and its risk scales with r. Patching its r-dimensional written subspace carries the head's entire pooled clean–corrupted effect, and patching the orthogonal complement carries none. For deep transformers this mediation is a hypothesis, which we test by patching the query-token residual state with complementary and matched-rank controls. On CounterFact-derived multi-relation ICL tasks (Meng et al., 2022) with Llama-3-8B (Llama Team, 2024), a 73-dimensional subspace of the 4096-dimensional residual stream recovers 78.8% of the clean–corrupted accuracy gap, while patching the complementary subspace restores 0%. Concept-swap interventions redirect predictions toward the injected relation, whereas random and cross-task matched-rank controls are largely ineffective. Additional experiments show consistent patterns on Qwen2.5-7B (Qwen Team, 2024), cross-lingual tasks (Lample et al., 2018), context-defined nonce mappings, and two-step compositional rules. These results do not imply that the full ICL circuit is low-dimensional. Rather, they indicate that, for the structured task families, a compact concept subspace carries a substantial portion of the recoverable causal signal.

Our contributions can be summarized as follows. First, we formulate ICL with shared task variation as estimation in an intrinsic concept subspace rather than unconstrained ambient regression. Second, we analyze ridge and least-squares proxies and show that, under explicit covariance and regularization assumptions, the concept-estimation error depends on the concept dimension r, while leakage is controlled by cross-subspace coupling. Third, we show that subspace-agnostic predictors pay for d, and derive exact results for subspace patching, subspace estimators and the decodability of the unread complement in the single-layer head that is optimal for the task family. Fourth, we turn the theory into a subspace-patching hypothesis and test it with concept-only patching, complementary-space controls, matched-rank controls, and concept swaps on controlled multi-relation tasks, context-defined nonce mappings and two-step compositional rules, and examine transfer to unseen relations, rank-threshold sensitivity and nonlinear probes.

2 Related Work

In-context learning. Empirical studies of ICL show that demonstrations provide more than a small supervised training set. A complementary theoretical line asks which algorithm a transformer executes in-context. Linear regression studies show that transformers can implement or approximate least-squares-style predictors in controlled settings Garg et al. (2022); Akyürek et al. (2023). Bayesian and algorithm-selection views explain ICL as posterior inference or as selecting among latent algorithms from the prompt Xie et al. (2022); Bai et al. (2023). Gradient-descent and meta-optimizer interpretations analyze how transformer computations can emulate iterative optimization procedures von Oswald et al. (2023); Dai et al. (2023); Ahn et al. (2023). Other analyses study how pretraining task diversity, meta-learning objectives, and learnability conditions shape the emergence of ICL Raventós et al. (2023); Kirsch et al. (2022); Wies et al. (2023). Mechanistic work further identifies induction-head circuits and task-vector phenomena as compact carriers of prompt-induced behavior Olsson et al. (2022); Yang et al. (2025). Closest to our view, Hu et al. (2026) find that in-context addition is transported by low-dimensional subspaces of a few attention heads, in line with our findings. These works support a task-inference view of ICL. Our focus is different: we ask when the inferred task should be treated as a low-dimensional concept coordinate rather than as an arbitrary ambient parameter.

Concept learning. Concept-based methods seek representations in which semantically meaningful variables can be inspected, constrained, or intervened on Koh et al. (2020); Hu et al. (2025b;a).

Post-hoc concept directions and whitening methods identify or align internal features with human-interpretable concepts Kim et al. (2018); Chen et al. (2020). Ante-hoc architectures make concept variables explicit through bottlenecks or concept embeddings Koh et al. (2020); Espinosa Zarlenga et al. (2022). However, unsupervised latent factors are generally unidentifiable without structural assumptions, weak supervision, or an appropriate invariance class Locatello et al. (2019); Khemakhem et al. (2020). Recent analyses also caution that concept bottlenecks may leak non-concept information rather than enforcing a clean semantic factorization Almudévar et al. (2026). In contrast, our concept variable is not a named human label and is not imposed during model training. It is a latent task coordinate inferred online from demonstrations.

Activation patching. Activation patching and interchange interventions test whether an internal variable causally mediates behavior by replacing activations between counterfactual runs. This logic underlies causal mediation and causal abstraction analyses Vig et al. (2020); Geiger et al. (2021). Circuit-level work studies how transformer components implement specific behaviors and how such circuits can be discovered automatically Wang et al. (2023); Conmy et al. (2023). Patching conclusions depend on corruption schemes, metrics, intervention granularity, and controls; in particular, subspace patching can steer behavior without faithfully identifying the subspace used by the original computation Zhang & Nanda (2024); Makelov et al. (2024). Our intervention target is not an entire residual stream or an arbitrary direction found by search, but the low-dimensional activation subspace predicted by the concept-subspace theory and tested with matched controls.

Overall, prior work largely treats ICL, concept-level representation, and causal intervention as separate phenomena. We connect them through a theory-first latent concept subspace that links statistical task structure, activation geometry, and causal behavior.

3 Concept-Subspace In-Context Learning

We develop the theory in four steps. First, we formalize concept-subspace ICL (Section 3.1) and analyze a ridge proxy that estimates the task without knowing the concept subspace (Section 3.2). We then characterize when the subspace is identifiable (Section 3.3). Finally, we show that a single-layer attention head trained on the task family reads the prompt only through this subspace, which yields the intervention predictions (Section 3.4).

3.1 Problem Formalization

We study a linear regression in-context learning (ICL) model with a shared low-dimensional task structure, following the stylized linear regression settings widely used to analyze ICL in transformers and linear attention models Guo et al. (2024); He et al. (2025). Let $d \geq 1$ and let $\Lambda \in \mathbb{R}^{d\times d}$ be positive definite. For a task parameter $w \in \mathbb{R}^d$, demonstrations are sampled as

$$x_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Lambda), \qquad y_i = \langle w, x_i \rangle, \qquad i = 1, \ldots, M. \tag{1}$$

A query $x_q \sim \mathcal{N}(0, \Lambda)$ is sampled independently, and the goal is to predict $y_q = \langle w, x_q \rangle$.

Assumption 1 (Concept-subspace task family)**.** Let $U \in \mathbb{R}^{d\times r}$ have orthonormal columns, where $1 \leq r \leq d$. The task family varies only in the concept subspace $\mathrm{span}(U)$: $w = U\beta$, with $\beta \sim \mathcal{N}(0, I_r)$. Equivalently, $y_i = \langle \beta, U^\top x_i \rangle$. We call $z := U^\top x \in \mathbb{R}^r$ the concept coordinates of x.

Let $U_\perp \in \mathbb{R}^{d\times(d-r)}$ be any orthonormal complement of U, and write the covariance in the basis $[U\ U_\perp]$:

$$\Lambda = \begin{bmatrix} \Lambda_{11} & \Lambda_{12} \\ \Lambda_{21} & \Lambda_{22} \end{bmatrix}, \qquad \Lambda_{11} = U^\top \Lambda U, \quad \Lambda_{22} = U_\perp^\top \Lambda U_\perp, \quad \Lambda_{12} = U^\top \Lambda U_\perp. \tag{2}$$

We use two covariance regimes: **(BD)** exact block diagonality, $\Lambda_{12} = 0$; and **(NBD)** near block diagonality, $\|\Lambda_{12}\|_{\mathrm{op}} \leq \rho$. Under (BD), concept and nuisance coordinates are independent because the joint distribution is Gaussian.

3.2 Ridge Proxy: Concept Estimation and Off-Subspace Leakage

For a demonstration set $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^M$, we focus on the ridge regression proxy for linear ICL:

$$\hat{w}_\mathcal{D} = (S + \lambda I_d)^{-1} \left(\frac{1}{M} \sum_{i=1}^M x_i y_i \right), \qquad f_\mathcal{D}(x) = \langle \hat{w}_\mathcal{D}, x \rangle, \qquad S = \frac{1}{M} \sum_{i=1}^M x_i x_i^\top. \tag{3}$$

Ridge and least-squares estimators are standard statistical proxies for test-time regression Hoerl & Kennard (1970); in the ICL literature, least-squares, ridge, gradient-descent, and preconditioned-gradient-descent predictors arise as explicit or implicit algorithms implemented by trained transformers. All ambient-space statements below assume $\lambda > 0$, so that $S + \lambda I_d$ is invertible. The unregularized least-squares case can be recovered in concept space under the usual full-rank condition on $Z^\top Z$, or in ambient space by replacing inverses with Moore–Penrose pseudoinverses under additional rank assumptions.

Lemma 1 (Exact subspace decomposition)**.** *Assume $w = U\beta$ and $\lambda > 0$. Let $Z \in \mathbb{R}^{M\times r}$ have rows $z_i^\top = (U^\top x_i)^\top$, let $y = (y_1, \dots, y_M)^\top$, and write $x_q = Uz + U_\perp t$, where $z = U^\top x_q$ and $t = U_\perp^\top x_q$. Define the empirical blocks $S_{11} = U^\top SU$, $S_{12} = U^\top SU_\perp$, $S_{21} = S_{12}^\top$, $S_{22} = U_\perp^\top SU_\perp$, and set $A_{11} = S_{11} + \lambda I_r$, $A_{12} = S_{12}$, $A_{22} = S_{22} + \lambda I_{d-r}$, $H = A_{22} - S_{21}A_{11}^{-1}A_{12}$. Then $H \succ 0$, and*

$$f_{\mathcal{D}}(x_q) = z^\top \hat{\beta}_{\mathcal{D}} + \Delta_{\mathcal{D}}(x_q), \qquad \hat{\beta}_{\mathcal{D}} = A_{11}^{-1}S_{11}\beta = (Z^\top Z + M\lambda I_r)^{-1}Z^\top y, \tag{4}$$

with

$$\Delta_{\mathcal{D}}(x_q) = \big(t^\top - z^\top A_{11}^{-1}A_{12}\big)\gamma_{\mathcal{D}}, \qquad \gamma_{\mathcal{D}} = H^{-1}S_{21}(\beta - \hat{\beta}_{\mathcal{D}}). \tag{5}$$

Thus the off-subspace contribution is controlled by the bottom coordinate $\gamma_{\mathcal{D}} = U_\perp^\top \hat{w}_{\mathcal{D}}$, which is induced through empirical cross-subspace coupling. In particular, under (BD), the population cross-covariance is zero, while the finite-sample term S_{21} remains.

For compactness, write $L_\delta := r + \log(a_0/\delta)$, $R_\perp(\delta) := \sqrt{\mathrm{tr}(\Lambda_{22})} + \sqrt{\|\Lambda_{22}\|_{\mathrm{op}}\log(a_0/\delta)}$, and $\eta_M(\delta) := R_\perp(\delta)\sqrt{L_\delta/M}$, where $a_0 > 0$ is a universal numerical constant.

Theorem 1 (Concept-coordinate estimation and off-subspace leakage)**.** *Under the data model in Equation (1) and the concept-subspace model in Assumption 1, take $\lambda = \lambda_0/M$ for a fixed $\lambda_0 > 0$. Under (BD), if $M \geq K_0 L_\delta$, for a constant K_0 depending only on Λ_{11} and λ_0, then with probability at least $1 - \delta$ over β, the demonstrations, and the query, the following hold. Let $z = U^\top x_q$, $t = U_\perp^\top x_q$, and let $\Delta_{\mathcal{D}}(x_q)$ be the off-subspace term in Lemma 1. Then*

$$\|\hat{\beta}_{\mathcal{D}} - \beta\| \leq K_\beta \frac{\sqrt{L_\delta}}{M}, \qquad |\Delta_{\mathcal{D}}(x_q)| \leq K_\Delta \left(\|t\| + \frac{R_\perp(\delta)}{\sqrt{M}}\|z\|\right)\eta_M(\delta). \tag{6}$$

Consequently, the prediction error splits into concept estimation and off-subspace leakage:

$$|f_{\mathcal{D}}(x_q) - \langle w, x_q\rangle| \leq \underbrace{K_\beta\|z\|\frac{\sqrt{L_\delta}}{M}}_{\text{concept estimation error}} + \underbrace{K_\Delta\left(\|t\| + \frac{R_\perp(\delta)}{\sqrt{M}}\|z\|\right)\eta_M(\delta)}_{\text{off-space leakage}}. \tag{7}$$

Moreover, for every $v \in \mathrm{span}(U)^\perp$,

$$|f_{\mathcal{D}}(x_q + v) - f_{\mathcal{D}}(x_q)| \leq K_\Delta \eta_M(\delta)\|v\|. \tag{8}$$

Under (NBD), the same bounds hold after replacing $\eta_M(\delta)$ by $\eta_M(\delta) + K_\rho\rho\sqrt{L_\delta}$ and $R_\perp(\delta)/\sqrt{M}$ by $R_\perp(\delta)/\sqrt{M} + K'_\rho\rho$, where K_ρ, K'_ρ depend only on Λ_{11} and λ_0.

Remark 1 (Interpretation of the dimension dependence). The leading concept-estimation term in Equation (6) depends on r, not on the ambient dimension d. The ambient predictor remains sensitive to nuisance directions through the cross-subspace coefficient $\gamma_{\mathcal{D}}$. Ambient ridge regression cannot avoid this sensitivity unless the nuisance covariance scale is controlled, e.g., through whitening or an explicit bound on $R_\perp(\delta)$. The theorem is therefore a decomposition result: concept inference is r-dimensional, while residual nuisance sensitivity is governed by cross-subspace coupling. Two further consequences are given in Appendix A.1: the concept-space estimator is a conjugate Bayesian posterior mean over β and is exactly invariant to nuisance perturbations. Appendix A.2 covers the extension to noisy labels.

3.3 Identifiability of the Concept Subspace

We next formalize when the concept subspace is statistically recoverable. As for latent factors in general, this requires additional structure or auxiliary information (Locatello et al., 2019; Khemakhem et al., 2020). Consider T tasks with coefficients $\beta_\tau \in \mathbb{R}^r$, observations $x_{\tau i} \sim \mathcal{N}(0, \Lambda)$, and labels $y_{\tau i} = \langle U\beta_\tau, x_{\tau i}\rangle + \varepsilon_{\tau i}$, where $\mathbb{E}[\varepsilon_{\tau i} \mid x_{\tau i}, \tau] = 0$. The concept subspace is *identifiable up to rotation* if every orthonormal $\tilde{U} \in \mathbb{R}^{d\times r}$ that induces the same task-conditioned linear predictors satisfies $\mathrm{span}(\tilde{U}) = \mathrm{span}(U)$. In that case $\tilde{U} = UO$ and $\tilde{\beta}_\tau = O^\top\beta_\tau$ for some orthogonal O.

Proposition 1 (Task-conditioned identifiability). *Assume $\Lambda \succ 0$ is known and $\mathbb{E}[x\varepsilon \mid \tau] = 0$. Let $B = [\beta_1, \ldots, \beta_T] \in \mathbb{R}^{r\times T}$ have rank r. For each task define $m_\tau := \mathbb{E}[xy \mid \tau] \in \mathbb{R}^d$ and $G_{xy} := [m_1, \ldots, m_T]$. Then*

$$m_\tau = \Lambda U \beta_\tau, \qquad G_{xy} = \Lambda U B, \qquad \operatorname{rank}(G_{xy}) = r, \tag{9}$$

and therefore

$$\operatorname{span}(\Lambda^{-1} G_{xy}) = \operatorname{span}(U). \tag{10}$$

If $\beta_\tau \overset{i.i.d.}{\sim} \mathcal{N}(0, I_r)$ and $T \geq r$, then $\operatorname{rank}(B) = r$ almost surely. Hence the concept subspace is identifiable up to rotation from task-conditioned first moments. In contrast, the unconditional first moment satisfies $\mathbb{E}[xy] = \Lambda U \mathbb{E}[\beta] = 0$ under the zero-mean task prior and cannot identify U by itself.

This justifies estimating a concept subspace $\widehat{U}$ from task-conditioned cross-moments: the statistically meaningful target is the subspace, not a particular basis for its coordinates. Section 3.4 shows which subspace such estimators recover inside an attention head.

3.4 From the Ridge Proxy to a Single-Layer Attention Head

The ridge proxy estimates the task without knowing $\operatorname{span}(U)$. Its concept-estimation bound does not depend on d, but its leakage bound grows with $\operatorname{tr}(\Lambda_{22})$ through $R_\perp(\delta)$, and hence with $d - r$ when $\Lambda = I_d$ (Theorem 1). This is an upper bound for one estimator, so it leaves open whether a cost in d is intrinsic to not knowing the subspace. Proposition 2 answers this for $\Lambda = I_d$, a uniformly random U and $M \leq d$. We call a predictor subspace-agnostic if it uses only the prompt and does not depend on U. Averaged over U, the risk of every subspace-agnostic predictor is bounded below by a quantity that grows with d, and the $\lambda \to 0$ limit of the ridge proxy attains this bound. In this regime, the cost comes from the missing subspace knowledge rather than from the choice of estimator.

Proposition 2 (Subspace knowledge is necessary). *Let $\Lambda = I_d$ and $M \leq d$, and let U be uniformly distributed on the Stiefel manifold, independent of β and of the data. Every predictor $\hat{g}(x_{1:M}, y_{1:M}, x_q)$ satisfies $\mathbb{E}(\hat{g} - \langle w, x_q\rangle)^2 \geq r(1 - M/d)$, with equality for the $\lambda \to 0$ limit of the ridge proxy.*

A transformer trained on the task family can instead store $\operatorname{span}(U)$ in its weights. We characterize a single-layer head that is optimal for the family; for this head, the representation behind task, function and in-context vectors (Hendel et al., 2023; Todd et al., 2024; Liu et al., 2024; Merullo et al., 2024; Dong et al., 2026; Bu et al., 2025) is derived rather than assumed. As in Section 4, we intervene on the query-token residual state, clean and corrupted prompts share the query, and interventions are orthogonal projections. Tokens are $e_i = (x_i, y_i)$, and we write $\hat{m}_{\mathcal{D}} := \frac{1}{M}\sum_i x_i y_i$ and $\overline{y^2} := \frac{1}{M}\sum_i y_i^2$.

Assumption 2 (Single-layer head). A linear attention head without softmax assigns demonstration i the score $\langle \kappa, e_i\rangle$ with $\kappa = (\kappa_x, \kappa_y) \in \mathbb{R}^d \times \mathbb{R}$, independent of the query, and writes

$$h(\mathcal{P}) = h_0(x_q) + W_{OV}\,\frac{1}{M}\sum_{i=1}^{M} \langle \kappa, e_i\rangle\, e_i \;\in \mathbb{R}^D. \tag{11}$$

The query embedding h_0 is continuous with $\mathbb{E}\|h_0(x_q)\|^4 < \infty$ and takes values in a subspace $\mathcal{H}_0$, with $\operatorname{range}(W_{OV}) \subseteq \mathcal{H}_0^\perp$ and $\dim \mathcal{H}_0^\perp \geq d$.

Assumption 3 (Optimal read-out). The prediction is $f(\mathcal{P}) = x_q^\top \Omega\, h(\mathcal{P})$, and $\theta = (\kappa, W_{OV}, \Omega)$ minimizes $\mathcal{L}(\theta) := \mathbb{E}\big(f(\mathcal{P}) - \langle w, x_q\rangle\big)^2$. The *minimum-norm read-out* has the smallest $\|\Omega\|_F$ among minimizers sharing (κ, W_{OV}).

Query-independent scores make the head output a task state by construction, which excludes standard linear self-attention. Assumption 3 concerns population optima, not training dynamics.

Patching recovery. The corrupted prompt $\tilde{\mathcal{P}}$ keeps the query and permutes the demonstration labels by a random permutation π that is independent of the data and differs from the identity with positive probability. Patching a subspace $V \subseteq \mathbb{R}^D$ with orthogonal projector P_V gives the prediction

$f_V := x_q^\top \Omega\big(h(\tilde{\mathcal{P}}) + P_V(h(\mathcal{P}) - h(\tilde{\mathcal{P}}))\big)$. With $f := f(\mathcal{P})$ and $\tilde{f} := f(\tilde{\mathcal{P}})$, the recovery and the difference energy of V are

$$\mathrm{Rec}(V) := \frac{\mathbb{E}[(f_V - \tilde{f})(f - \tilde{f})]}{\mathbb{E}[(f - \tilde{f})^2]}, \qquad E(V) := \frac{\mathbb{E}\|P_V(h(\mathcal{P}) - h(\tilde{\mathcal{P}}))\|^2}{\mathbb{E}\|h(\mathcal{P}) - h(\tilde{\mathcal{P}})\|^2}. \tag{12}$$

Rec is the pooled analogue of the recovery rate in Section 4, and E is its natural null.

Theorem 2 (Concept-subspace mediation in a single-layer head)**.** *Assume Equation (1), Assumptions 1–3, $r \geq 2$ and $\Lambda \succ 0$, and let $\Gamma_c := \big((1 + \frac{1}{M})\Lambda_{11} + \frac{\mathrm{tr}\,\Lambda_{11}}{M} I_r\big)^{-1}$.*

- *(i)* Structure. *Every minimizer attends in proportion to the labels, $\kappa = (0, \kappa_y)$ with $\kappa_y \neq 0$, and writes $h(\mathcal{P}) = h_0(x_q) + W\hat{m}_{\mathcal{D}} + \overline{y^2}\, v_y$ with $\Omega W = U\Gamma_c U^\top$ and $\Omega v_y = 0$. Its prediction is $f(\mathcal{P}) = (U^\top x_q)^\top \hat{\beta}(\mathcal{P})$ with $\hat{\beta}(\mathcal{P}) := \Gamma_c S_{11}\beta$, and its risk is $\mathrm{tr}\,\Lambda_{11} - \mathrm{tr}(\Gamma_c \Lambda_{11}^2)$, which equals $\frac{r(r+1)}{M+r+1}$ for $\Lambda = I_d$.*
- *(ii)* Mediation. *Let $\mathcal{W} := W\,\mathrm{span}(\Lambda U)$; then $\dim \mathcal{W} = r$. For every minimizer and every subspace V: $\mathrm{Rec}(V) + \mathrm{Rec}(V^\perp) = 1$; $\mathrm{Rec}(V) = 1$ if $V \supseteq \mathcal{W}$ and $\mathrm{Rec}(V) = 0$ if $V \perp \mathcal{W}$; and $\mathbb{E}\,\mathrm{Rec}(V) = k/D$ for a uniformly random k-dimensional V independent of the data. For the minimum-norm read-out, whose row space $\mathcal{R}$ has dimension r, transplanting the $\mathcal{R}$-component of $h(\mathcal{P}_2)$ into $h(\mathcal{P}_1)$ yields exactly $x_{q,1}^\top U\Gamma_c U^\top \hat{m}_{\mathcal{D}}(\mathcal{P}_2)$, even when the two queries differ; $\mathcal{R} = \mathcal{W}$ under (BD) if $W^\top W \propto I_d$ and $v_y \perp \mathrm{range}(W)$.*
- *(iii)* Estimators. *If a target $Y \in \mathbb{R}^p$ has $\mathbb{E}[Y \mid \beta, x_{1:M}, x_q] = \Theta(x_q)\beta$ with $\mathrm{rank}\,\mathbb{E}\Theta(x_q) = r$, then $\mathbb{E}[hY^\top] = W\Lambda U\,\mathbb{E}[\Theta(x_q)]^\top$ has column space $\mathcal{W}$. So do the task centroids $\mathbb{E}[h(\mathcal{P}) - h(\tilde{\mathcal{P}}) \mid \beta] = c_\pi W\Lambda U\beta$, $c_\pi > 0$, when the tasks' β span $\mathbb{R}^r$, whereas randomly re-paired targets give $\mathbb{E}[hY^\top] = 0$. At the population level, the rank selected at any cumulative-energy threshold is at most r.*
- *(iv)* Complement. *Let $h_\perp := P_{\mathcal{W}^\perp} h$ and $q_\beta := \beta^\top \Lambda_{11}\beta$. The conditional law of $h_\perp$ given β depends on β only through q_β, and $\mathbb{E}[h_\perp \mid \beta] = \mathbb{E}\, h_0(x_q) + q_\beta P_{\mathcal{W}^\perp} v_y$. Task centroids of $h_\perp$ are therefore collinear, and no decoder can separate tasks with equal q_β from $h_\perp$.*
- *(v)* Use versus decodability. *Let $\Lambda = I_d$, $W^\top W = sI_d$, $v_y \perp \mathrm{range}(W)$, π uniform, and minimum-norm read-outs. A minimizer for the unstructured prior $w \sim \mathcal{N}(0, I_d)$ with the same W writes the same state, so the estimators in (iii) return the same $\mathcal{W}$. Yet it has $\mathrm{Rec}(V) = E(V)$ for every V and risk $\frac{r(d+1)}{M+d+1}$ on the concept family, whereas the structured head has $\mathrm{Rec}(\mathcal{W}) = 1$ although $E(\mathcal{W}) < 1$ whenever $r < d$.*

Interpretation. The optimal head writes the cross-moment $\hat{m}_{\mathcal{D}}$, a query-independent task state, at the query position, and the read-out combines it with the query. The subspace is stored in the weights ($\Omega W = U\Gamma_c U^\top$), so the in-context computation only estimates β: its risk scales with r (i), whereas an unstructured head (v) and any subspace-agnostic predictor (Proposition 2) pay for d. Theorem 2 separates where the task is written from how it is read. Estimators recover the written subspace $\mathcal{W}$ from states alone (iii), and patching $\mathcal{W}$ carries the entire pooled clean–corrupted effect for every optimal read-out, without (BD) (ii). Because Rec is a slope, per-example agreement is a stronger property; under (NBD) it degrades only at order ρ^2. The complement can remain decodable, but only through the scalar q_β (iv). An unstructured head writes the same $\mathcal{W}$, so the evidence for structured reading is the gap $\mathrm{Rec}(V) - E(V)$, not the dimension of V (v). The map $U\Gamma_c U^\top$ coincides with known one-layer optima (Zhang et al., 2024; Ahn et al., 2023; Mahankali et al., 2024; Li et al., 2024) under a rank-r task prior (cf. Kwon et al., 2026; Takanami et al., 2026); items (ii)–(v) and Proposition 2 are new. Proofs are in Appendices A.11 and A.12.

Testable predictions. For deep transformers, Theorem 2(i) is a hypothesis, which we probe by patching at the query token. For a subspace $\widehat{U}$ estimated as in (iii), the theorem predicts (a) *selectivity*: $\mathrm{Rec}(\widehat{U})$ close to 1 and well above $E(\widehat{U})$ (v); (b) *query independence*: the $\widehat{U}$-component transfers across queries and redirects predictions toward the donor task (ii); (c) a *scalar complement*: task centroids in $\widehat{U}^\perp$ are nearly collinear (iv); and (d) a *task-level rank*: after centering across tasks, the estimated rank is at most the number of tasks minus one. Additivity and the k/D law hold for any linear read-out, so we use them only as linearity checks, measured on logit differences.

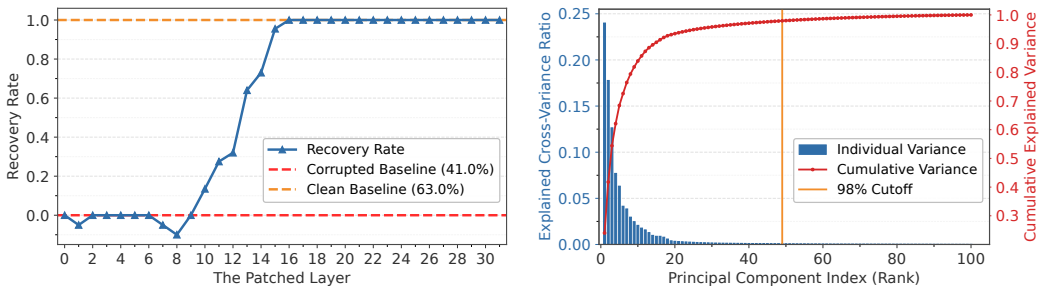

Figure 1: Layerwise localization and low-rank extraction of task-aligned residual directions (Llama-3-8B and CounterFact). *Left*: clean-to-corrupted patching localizes layers where task information is causally accessible. *Right*: the 98% explained-cross-variance threshold selects a low-dimensional concept subspace at layer 30.

4 Mechanistic Validation for Concept-Subspace ICL

In this section, we test whether the theory-motivated concept subspace appears as a causally effective activation subspace in large language models. The goal is not full circuit recovery. Instead, we ask whether a compact, task-aligned residual-stream subspace carries a substantial fraction of the recoverable ICL signal under controlled interventions.

4.1 Experimental Protocol and Metrics

Tasks. We evaluate structured multi-relation tasks from CounterFact Meng et al. (2022). Retaining 34 mapping relations (e.g., country $\rightarrow$ capital), we evaluate 90 held-out queries per relation using 12-shot prompts. For each query, we contrast two conditions: a *clean context* governed by a consistent concept and a *corrupted context* that destroys concept consistency (e.g., via label shuffling) while preserving prompt format, length, and token statistics. This controlled corruption isolates concept mediation from superficial pattern matching. Moreover, Appendix C.2 evaluates implicit cross-lingual translation tasks, confirming that this mediation extends from factual relations to abstract rules. The same intervention pattern holds on context-defined nonce mappings and two-step compositional rules (Appendices C.3 and C.4), and a subspace estimated from 24 relations recovers 48.8% of the gap on 10 unseen relations, against 56.1% for one estimated from all 34 (Appendix C.5).

Evaluation Settings. We compare five primary conditions: (i) the corrupted baseline, (ii) the clean few-shot upper bound, (iii) full clean-to-corrupted patching, (iv) concept-subspace-only patching, and (v) complementary-subspace patching. We also evaluate two same-rank controls: a random orthogonal subspace and a cross-task subspace extracted from unrelated relations.

Evaluation Metrics. We report top-1 accuracy for standard evaluation. For patching experiments, we additionally report the *recovery rate*, i.e., the fraction of the clean–corrupted accuracy gap restored by the intervention: $\text{Recovery Rate}(\%) = 100 \cdot \frac{\text{Acc}_{\text{patch}} - \text{Acc}_{\text{corr}}}{\text{Acc}_{\text{clean}} - \text{Acc}_{\text{corr}}}$. For concept swapping and same-rank controls, we report *override success*, i.e., the fraction of examples whose final prediction follows the injected target relation. More implementation details are provided in Appendix B.1.

4.2 Localizing and Estimating the Concept Subspace

Localizing the concept subspace. We first localize where task-relevant information becomes available for in-context prediction. For each layer ℓ, we run the Llama-3-8B model on clean and corrupted prompts, patch the clean query-token residual state h_ℓ^{clean} into the corrupted run, and measure the resulting recovery. Figure 1 (left) reveals a strict late-layer emergence of recovery: patching effects remain near zero through layers 0–14, become visible around layers 15–20, and rise steeply after layer 21. This suggests that task-level information is not uniformly distributed, but consolidates only after deep contextual integration. Consequently, we select layer 30 ($\ell^\star = 30$) for all subsequent concept-subspace estimation and causal patching experiments. Further layerwise analyses confirm the late-layer emergence of this causal concept bottleneck (see Appendix B.3). This design choice is methodologically important: a linear subspace estimator is only meaningful once the task variable has become explicitly represented in the residual stream. We further

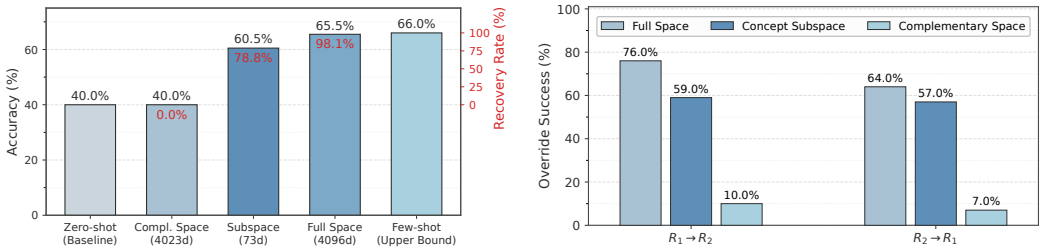

Figure 2: Concept-subspace patching and swapping. *Left:* the learned subspace recovers most of the clean-corrupted gap. *Right:* subspace swaps redirect predictions toward the injected relation.

discuss the results on Qwen2.5-7B Qwen Team (2024) in Appendix C.1 and on the cross-lingual task Lample et al. (2018) in Appendix C.2.

Estimating the concept subspace. Guided by earlier localization, we estimate the task-aligned subspace at layer $\ell^\star = 30$ using clean 12-shot prompts across 34 structured CounterFact relations (e.g., country $\to$ capital, entity $\to$ language), with 90 held-out queries per relation. To isolate relation-level directions from entity-specific variations, we take the query-token residual activations $\mathbf{H} \in \mathbb{R}^{n \times D}$ and the aligned target-supervision matrix Y to compute their cross-covariance $C_{\mathbf{H}Y} = \frac{1}{n}\mathbf{H}^\top Y$. Performing SVD ($C_{\mathbf{H}Y} = Q\Sigma Q'^\top$), we define the subspace by selecting the minimal rank $\hat{r}$ that captures $\geq 98\%$ of the cumulative cross-variance: $\frac{\sum_{i=1}^{\hat{r}} s_i^2}{\sum_i s_i^2} \geq 0.98$. We then define the empirical concept subspace as $\widehat{U}_{\ell^\star} = Q_{[:,1:\hat{r}]}$. Equivalently, if one first fits a linear probe from $\mathbf{H}$ to Y, $\widehat{U}_{\ell^\star}$ is given by the leading left-singular directions of the probe weight matrix. This construction is important because it suppresses entity-specific redundancy while preserving only those activation directions that are consistently associated with relation-level output structure across prompts. Figure 1 (right) reveals a sharply concentrated spectrum at layer 30. A 98% explained-cross-variance threshold selects only a minimal fraction of directions in the representative run, and the cutoff remains in a narrow range across runs (see Appendix B.2 for evidence of the estimated rank's stability across varying estimation sizes and few-shot counts). This low rank relative to the 4096-dimensional stream suggests that ICL compresses demonstrations into a compact task representation rather than a diffuse code. While this spectral signature is purely correlational, it enables our causal test: if the concept-subspace hypothesis holds, intervening solely on $\widehat{U}_{\ell^\star}$ should recover most of the lost ICL behavior, whereas intervening on its orthogonal complement should prove ineffective.

4.3 Concept-Subspace Patching Recovers ICL Behavior

Let $P_{\widehat{U}_{\ell^\star}}$ denote the orthogonal projector onto $\widehat{U}_{\ell^\star}$. Given clean and corrupted runs for the same query, we intervene on the corrupted residual stream by patching only the learned concept component $\tilde{h}_{\ell^\star} = h^{\text{corr}}_{\ell^\star} + P_{\widehat{U}_{\ell^\star}}\left(h^{\text{clean}}_{\ell^\star} - h^{\text{corr}}_{\ell^\star}\right)$. We compare this intervention with full-state patching, complementary-space patching using $I - P_{\widehat{U}_{\ell^\star}}$, and the unpatched zero-shot baseline. Figure 2 (left) shows a strong asymmetry between the learned subspace and its orthogonal complement. The zero-shot baseline obtains 40.0% accuracy, while the clean few-shot upper bound reaches 66.0%. Full activation patching raises accuracy to 65.5%, recovering 98.1% of the clean–corrupted gap. Patching only the 73-dimensional concept subspace still reaches 60.5%, corresponding to 78.8% recovery, despite modifying only $73/4096 \approx 1.8\%$ of the dimensions. In contrast, patching the 4023-dimensional complement leaves accuracy unchanged at 40.0%, yielding 0.0% recovery. Thus, the intervention effect is not proportional to the number of patched dimensions. Most recoverable task-level signal is concentrated in a small task-aligned subspace, while the high-dimensional complement is insufficient to restore behavior. Directional noise perturbation experiments confirm that this low-dimensional concept subspace acts as the central causal bottleneck governing the model's logical inference (see Figure 4). Additionally, the concept subspace demonstrates higher stability against few-shot label noise than the full space (Appendix B.4). The result supports a more precise claim: the learned subspace captures the dominant recoverable causal component of the ICL computation in these structured relation tasks (see Appendix B.3 for scaling analyses confirming this pattern across 1 to 12 demonstrations). Nonlinear probes decode part of the relation from the complement, yet the complement carries little patchable signal, consistent with Theorem 2(iv) (Appendix C.7).

4.4 Concept Swaps and Matched-Rank Specificity Controls

Concept swaps. To stringently test causal mediation via behavioral redirection, we extract clean layer-30 residual states for entities participating in two distinct relations (R_1, R_2) with corresponding concept parameters (β_1, β_2). During a source-relation forward pass, we inject target-relation information using three parallel interventions: (i) *full-space replacement* (swapping the entire residual state); (ii) *concept-subspace replacement* (swapping solely the $\widehat{U}_{\ell^\star}$ projection); and (iii) *complementary-space replacement* (swapping only the orthogonal complement while retaining the source projection). We evaluate *override success*—the fraction of predictions successfully flipped to the target relation. Systematic logical shifts induced specifically by subspace injection would firmly establish this low-dimensional representation as a causal mediator of in-context learning. Figure 2 (right) shows that directional control is concentrated in the learned subspace. Full-space replacement achieves 76.0% for $R_1 \to R_2$ and 64.0% for $R_2 \to R_1$, close to the native few-shot accuracies of the target relations (75.0% for R_2, 64.0% for R_1). Replacing only the concept subspace still yields substantial override, whereas replacing the complementary space is largely ineffective. Therefore, manipulating just the minimal concept subspace is sufficient to invert the model's logical outputs, capturing the vast majority of the adversarial override achieved by full-space replacement. The inertness of the complementary space supports the learned subspace as the dominant causal bottleneck governing in-context task execution.

Matched-rank specificity controls. We ask whether the overwrite effect is specific to the learned concept subspace or can be reproduced by an arbitrary low-rank intervention. Using the same source–target relation setup as in the concept-swap experiment, we patch the target hidden state into the source inference trace after projecting the update onto one of three matched-dimensional subspaces at the target layer: the learned concept subspace $\widehat{U}_{\ell^\star}$, a random subspace U_{rand}, and a cross-task subspace U_{cross} estimated from an unrelated task. We report the adversarial overwrite success rate. Table 1 shows that only the learned subspace ($\widehat{U}_{\ell^\star}$) achieves substantial override (58.8%), while matched-rank controls (U_{rand}, U_{cross}) remain near the 2.5% baseline ($\leq 3.3\%$). Since dimensionality is controlled, this gap suggests that causal efficacy stems specifically from task-aligned directions rather than arbitrary low-rank edits. Alongside the inert complementary space, this establishes the learned concept subspace as a highly specific causal mediator of in-context behavior. Appendix B.6 further shows that debiased concept-coordinate estimates contract as the number of demonstrations grows.

Table 1: Matched-rank controls for override success.

Intervention	Success (%)
None	2.5
$\widehat{U}_{\ell^\star}$	58.8
U_{rand}	2.3
U_{cross}	3.3

5 Conclusion

We studied ICL through the lens of concept subspace learning. Ridge and least-squares ICL proxies decompose into concept-coordinate regression and off-subspace leakage: the intrinsic task estimate is governed by the concept dimension under explicit covariance assumptions. A single-layer linear attention head optimal for the task family instead stores the subspace in its weights, so patching its written subspace carries its entire pooled clean–corrupted effect. Our experiments support this mediation hypothesis for deep transformers: on CounterFact-derived relation tasks, a low-dimensional residual-stream subspace recovers most of the clean–corrupted ICL gap, while complementary and matched-rank controls have little effect, and concept swaps redirect predictions toward an injected relation. The same pattern appears on another model, a cross-lingual rule, context-defined nonce mappings, and two-step compositional rules. These findings suggest that concept subspaces provide a useful bridge between statistical accounts of ICL and causal activation-space analyses.

Limitations and future work. Our theory covers an idealized single-layer linear head at its population optimum, and extending it to multi-layer softmax transformers and to training dynamics remains open. As in other patching studies, our interventions show partial sufficiency but cannot certify that the subspace is the model's own variable (Makelov et al., 2024). Measuring the difference energy and cross-query transfer would test the selectivity and query-independence predictions of Section 3.4. The selected rank does not estimate r and depends on a threshold that held-out recovery could replace. Our evidence covers structured task families, 7–8B models and one or two intervention layers, and does not imply that full ICL circuits are low-dimensional. Open-ended tasks, larger models and more intervention layers would test how far concept-subspace mediation extends.

AI USE STATEMENT

In this work, we did not use generative AI tools for any of the tasks that require disclosure under the ICLR 2027 AI Policy for Authors, including formulating hypotheses, developing theoretical models, formulating or proving mathematical claims, designing or implementing methods and experiments, generating or processing data, interpreting results, and translation. We did, however, use generative AI tools for searching for and identifying relevant literature, and for checking grammar and improving the readability of author-written text. We have reviewed all AI-assisted work: retrieved references were verified against their original sources, and language edits were checked to ensure that the technical content remained unchanged. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work does not involve human subjects, and all experiments are conducted on publicly available datasets used in accordance with their licenses. We are not aware of any immediate ethical concerns or potential for harmful applications arising from this work.

REPRODUCIBILITY STATEMENT

To facilitate reproducibility, we release the source code for all experiments in an anonymous repository at `https://anonymous.4open.science/r/ICL-Concept-Subspace`. The repository includes documentation and code for reproducing the reported results. Experimental settings and hyperparameters are described in Section 4 and Appendix B.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems 36*, pp. 45614–45650, 2023.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The 11th International Conference on Learning Representations*, 2023.

Antonio Almudévar, José Miguel Hernández-Lobato, and Alfonso Ortega. There was never a bottleneck in concept bottleneck models. In *The 14th International Conference on Learning Representations*, 2026.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems 36*, pp. 57125–57211, 2023.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

Tom B. Brown, Benjamin Mann, Nick Ryder, and et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33*, pp. 1877–1901, 2020.

Dake Bu, Wei Huang, Andi Han, Atsushi Nitanda, Qingfu Zhang, Hau-San Wong, and Taiji Suzuki. Provable in-context vector arithmetic via retrieving task concepts. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 5669–5724, 2025.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems 35*, pp. 18878–18891, 2022.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems 36*, pp. 16318–16352, 2023.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models secretly perform gradient descent as meta-optimizers. In *Findings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 4005–4019, 2023.

Yuxin Dong, Jiachen Jiang, Zhihui Zhu, and Xia Ning. Understanding task vectors in in-context learning: Emergence, functionality, and limitations. In *The 14th International Conference on Learning Representations*, 2026.

Clément Dumas, Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Separating tongue from thought: Activation patching reveals language-agnostic concept representations in transformers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 31822–31841, 2025.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off. In *Advances in Neural Information Processing Systems 35*, pp. 21400–21413, 2022.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems 35*, pp. 30583–30598, 2022.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. Causal abstractions of neural networks. In *Advances in Neural Information Processing Systems 34*, pp. 9574–9586, 2021.

Tianyu Guo, Wei Hu, Song Mei, Huan Wang, Caiming Xiong, Silvio Savarese, and Yu Bai. How do transformers learn in-context beyond simple functions? a case study on learning with representations. In *The 12th International Conference on Learning Representations*, 2024.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2009.

Jianliang He, Xintian Pan, Siyu Chen, and Zhuoran Yang. In-context linear regression demystified: Training dynamics and mechanistic interpretability of multi-head softmax attention. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 22686–22742, 2025.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

Guan Zhe Hong, Bhavya Vasudeva, Vatsal Sharan, Cyrus Rashtchian, Prabhakar Raghavan, and Rina Panigrahy. Latent concept disentanglement in transformer-based language models. In *The 14th International Conference on Learning Representations*, 2026.

Lijie Hu, Tianhao Huang, Huanyi Xie, Xilin Gong, Chenyang Ren, Zhengyu Hu, Lu Yu, Ping Ma, and Di Wang. Semi-supervised concept bottleneck models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2110–2119, 2025a.

Lijie Hu, Chenyang Ren, Zhengyu Hu, Hongbin Lin, Cheng-Long Wang, Zhen Tan, Weimin Lyu, Jingfeng Zhang, Hui Xiong, and Di Wang. Editable concept bottleneck models. In *Proceedings of the 42nd International Conference on Machine Learning*, pp. 24678–24726, 2025b.

Xinyan Hu, Kayo Yin, Michael I. Jordan, Jacob Steinhardt, and Lijie Chen. Understanding in-context learning of addition via activation subspaces. In *Proceedings of the Third Conference on Language Modeling*, 2026.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *The 12th International Conference on Learning Representations*, 2024.

Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

Ilyes Khemakhem, Diederik P. Kingma, Ricardo Pio Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 2207–2217, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2668–2677, 2018.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, 2022.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 5338–5348, 2020.

Soo Min Kwon, Alec S. Xu, Can Yaras, Laura Balzano, and Qing Qu. Out-of-distribution generalization of in-context learning: A low-dimensional subspace perspective. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *The 6th International Conference on Learning Representations*, 2018.

Yingcong Li, Ankit Singh Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. In *Advances in Neural Information Processing Systems 37*, pp. 138324–138364, 2024.

Sheng Liu, Haotian Ye, Lei Xing, and James Y. Zou. In-context vectors: Making in context learning more effective and controllable through latent space steering. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 32287–32307, 2024.

Llama Team. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 4114–4124, 2019.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The 12th International Conference on Learning Representations*, 2024.

Aleksandar Makelov, Georg Lange, Atticus Geiger, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *The 12th International Conference on Learning Representations*, 2024.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems 35*, pp. 17359–17372, 2022.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. Language models implement simple Word2Vec-style vector arithmetic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5030–5047, 2024.

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, 2022.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.

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

Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning "learns" in-context: Disentangling task recognition and task learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8298–8319, 2023.

Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems 36*, pp. 14228–14246, 2023.

Jan Stuehmer, Richard Turner, and Sebastian Nowozin. Independent subspace analysis for unsupervised learning of disentangled representations. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 1200–1210, 2020.

Kaito Takanami, Takashi Takahashi, and Yoshiyuki Kabashima. Learning linear regression with low-rank tasks in-context. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *The 12th International Conference on Learning Representations*, 2024.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems 33*, pp. 12388–12401, 2020.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 35151–35174, 2023.

Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. In *The 11th International Conference on Learning Representations*, 2023.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning. In *Advances in Neural Information Processing Systems 36*, pp. 36637–36651, 2023.

Yongliang Wu and Xu Yang. A glance at in-context learning. *Frontiers of Computer Science*, 18(5): 185347, 2024.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit Bayesian inference. In *The 10th International Conference on Learning Representations*, 2022.

Liu Yang, Ziqian Lin, Kangwook Lee, Dimitris Papailiopoulos, and Robert D. Nowak. Task vectors in in-context learning: Emergence, formation, and benefits. In *Proceedings of the Second Conference on Language Modeling*, 2025.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *The 12th International Conference on Learning Representations*, 2024.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

A Additional Theory and Proofs

A.1 Bayesian Inference and Nuisance Stability

Theorem 1 shows that the intrinsic estimation problem is r-dimensional and that deviations from exact nuisance invariance are controlled by cross-subspace coupling. We record two consequences and the extension to noisy labels.

(i) Bayesian concept inference.

Corollary 1 (Posterior interpretation)**.** *Assume the noisy linear-Gaussian concept model $y = Z\beta+\varepsilon$, where $\beta \sim \mathcal{N}(0, I_r)$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_M)$. Then*

$$\beta \mid \mathcal{D} \sim \mathcal{N}(\mu_{\mathcal{D}}, \Sigma_{\mathcal{D}}), \qquad \Sigma_{\mathcal{D}} = \left(I_r + \frac{1}{\sigma^2} Z^\top Z\right)^{-1}, \qquad \mu_{\mathcal{D}} = \Sigma_{\mathcal{D}} \frac{1}{\sigma^2} Z^\top y. \tag{13}$$

The Bayes predictor is $\mathbb{E}[y_q \mid x_q, \mathcal{D}] = \mu_{\mathcal{D}}^\top U^\top x_q$, and the posterior mean coincides with the concept-space ridge estimator in Lemma 1 when $\lambda = \sigma^2/M$. This is the standard conjugate Bayesian linear regression posterior (Bishop, 2006; Murphy, 2012), specialized to the concept coordinate space.

This posterior lives in the concept coordinate space rather than in the ambient parameter space.

(ii) Stability under nuisance perturbations.

Corollary 2 (Nuisance stability)**.** *Under the assumptions of Theorem 1 and under (BD), with probability at least $1-\delta$, for every $v \in \mathrm{span}(U)^\perp$,*

$$|f_{\mathcal{D}}(x_q + v) - f_{\mathcal{D}}(x_q)| \le K_{\mathrm{stab}} R_\perp(\delta) \|v\| \sqrt{\frac{r + \log(a_0/\delta)}{M}}. \tag{14}$$

For the Bayes concept predictor $x \mapsto \mu_{\mathcal{D}}^\top U^\top x$, the same perturbation has exactly zero effect: $\mu_{\mathcal{D}}^\top U^\top (x_q + v) = \mu_{\mathcal{D}}^\top U^\top x_q$.

A.2 Noisy labels in concept space

For concept-space ridge or Bayesian inference, homoscedastic label noise adds the usual $\sigma\sqrt{r/M}$ estimation term, as in standard finite-sample analyses of linear regression Hastie et al. (2009); Wainwright (2019). The ambient ridge predictor satisfies the same exact block decomposition as Lemma 1, but with an additional off-subspace noise term; any ambient-space noisy-label stability bound must therefore control this term explicitly. We extend Equation (1) to homoscedastic noise:

$$y_i = \langle U\beta, x_i \rangle + \varepsilon_i, \qquad \varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2), \qquad \beta \sim \mathcal{N}(0, I_r). \tag{15}$$

Let $Z \in \mathbb{R}^{M \times r}$ have rows $z_i^\top = (U^\top x_i)^\top$, let $y = (y_1, \ldots, y_M)^\top$, and define the concept-space ridge estimator $\hat{\beta}_{\mathcal{D}} = (Z^\top Z + M\lambda I_r)^{-1} Z^\top y$.

Corollary 3 (Noisy concept inference)**.** *Assume Equation (15), $\Lambda_{11} \succ 0$, and $\lambda = \lambda_0/M$ for fixed $\lambda_0 > 0$. Let $L_\delta = r + \log(a_0/\delta)$. If $M \ge K_0 L_\delta$, then with probability at least $1-\delta$,*

$$\|\hat{\beta}_{\mathcal{D}} - \beta\| \le K_1 \left(\frac{\sqrt{L_\delta}}{M} + \sigma \sqrt{\frac{L_\delta}{M}} \right). \tag{16}$$

Consequently, for the concept predictor $\tilde{f}_{\mathcal{D}}(x_q) := (U^\top x_q)^\top \hat{\beta}_{\mathcal{D}}$,

$$|\tilde{f}_{\mathcal{D}}(x_q) - \langle U\beta, x_q \rangle| \le \|U^\top x_q\| K_1 \left(\frac{\sqrt{L_\delta}}{M} + \sigma \sqrt{\frac{L_\delta}{M}} \right). \tag{17}$$

Moreover, $\tilde{f}_{\mathcal{D}}(x_q + v) = \tilde{f}_{\mathcal{D}}(x_q)$ for every $v \in \mathrm{span}(U)^\perp$.

Remark 2 (Ambient ridge with noisy labels). For the ambient ridge predictor $\hat{w}_{\mathcal{D}} = (S + \lambda I_d)^{-1} M^{-1} \sum_i x_i y_i$, the same block decomposition as Lemma 1 holds, but the bottom coordinate obtains an additional off-subspace noise term. This term can depend on the nuisance covariance scale and regularization. Therefore, the clean $\sigma\sqrt{r/M}$ rate in Corollary 3 should be interpreted as a concept-space or Bayesian inference statement unless additional assumptions control off-subspace noise amplification.

A.3 AUXILIARY CONCENTRATION FACTS

We use two standard Gaussian concentration facts for non-asymptotic covariance and Gaussian tail bounds. First, if $z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \Lambda_{11})$ and $S_{11} = M^{-1}Z^\top Z$, then for $M \geq K(r + \log(a_0/\delta))$,

$$S_{11} \succeq \frac{1}{2}\lambda_{\min}(\Lambda_{11})I_r \tag{18}$$

with probability at least $1 - \delta$, where K depends only on Λ_{11}. Second, if $g \sim \mathcal{N}(0, \Sigma)$, then with probability at least $1 - \delta$,

$$\|g\| \leq \sqrt{\mathrm{tr}(\Sigma)} + \sqrt{2\|\Sigma\|_{\mathrm{op}} \log(1/\delta)}. \tag{19}$$

Throughout the appendix, constants may change from line to line. The universal constant a_0 inside $L_\delta = r + \log(a_0/\delta)$ is chosen large enough to absorb union bounds.

A.4 PROOF OF LEMMA 1

Proof. Let $Q = [U\ U_\perp]$. In the Q-basis, write $Q^\top \hat{w}_\mathcal{D} = (\alpha^\top, \gamma^\top)^\top$ and

$$Q^\top S Q = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}. \tag{20}$$

Since $w = U\beta$, the ridge normal equations are

$$\begin{bmatrix} S_{11} + \lambda I_r & S_{12} \\ S_{21} & S_{22} + \lambda I_{d-r} \end{bmatrix} \begin{bmatrix} \alpha \\ \gamma \end{bmatrix} = \begin{bmatrix} S_{11}\beta \\ S_{21}\beta \end{bmatrix}. \tag{21}$$

Set $A_{11} = S_{11} + \lambda I_r$, $A_{12} = S_{12}$, $A_{22} = S_{22} + \lambda I_{d-r}$, and $H = A_{22} - S_{21}A_{11}^{-1}A_{12}$. Since $S + \lambda I_d \succ 0$, its Schur complement satisfies $H \succ 0$. The first block equation gives $\alpha = A_{11}^{-1}S_{11}\beta - A_{11}^{-1}A_{12}\gamma = \hat{\beta}_\mathcal{D} - A_{11}^{-1}A_{12}\gamma$, where $\hat{\beta}_\mathcal{D} = A_{11}^{-1}S_{11}\beta$. Substituting this into the second block equation yields $H\gamma = S_{21}(\beta - \hat{\beta}_\mathcal{D})$, hence $\gamma = \gamma_\mathcal{D} = H^{-1}S_{21}(\beta - \hat{\beta}_\mathcal{D})$. Finally, for $x_q = Uz + U_\perp t$,

$$f_\mathcal{D}(x_q) = z^\top \alpha + t^\top \gamma = z^\top \hat{\beta}_\mathcal{D} + \big(t^\top - z^\top A_{11}^{-1}A_{12}\big)\gamma_\mathcal{D}. \tag{22}$$

The identity $\hat{\beta}_\mathcal{D} = (Z^\top Z + M\lambda I_r)^{-1}Z^\top y$ follows from $S_{11} = M^{-1}Z^\top Z$ and $y = Z\beta$. □

A.5 PROOF OF THEOREM 1

Proof. Let $\ell_{11} := \lambda_{\min}(\Lambda_{11})$. On the event in Equation (18), $S_{11} \succeq \ell_{11}I_r/2$. Also, with probability at least $1 - \delta/a_0$, $\|\beta\| \leq K\sqrt{L_\delta}$. Since $\hat{\beta}_\mathcal{D} = (S_{11} + \lambda I_r)^{-1}S_{11}\beta$, we have

$$\hat{\beta}_\mathcal{D} - \beta = -\lambda(S_{11} + \lambda I_r)^{-1}\beta. \tag{23}$$

With $\lambda = \lambda_0/M$, this gives

$$\|\hat{\beta}_\mathcal{D} - \beta\| \leq \frac{\lambda_0/M}{\ell_{11}/2 + \lambda_0/M}\|\beta\| \leq K_\beta \frac{\sqrt{L_\delta}}{M}, \tag{24}$$

which proves the concept-estimation bound.

We next control the off-subspace leakage. By Lemma 1, $\Delta_\mathcal{D}(x_q) = (t^\top - z^\top A_{11}^{-1}A_{12})\gamma_\mathcal{D}$, where $\gamma_\mathcal{D} = H^{-1}S_{21}(\beta - \hat{\beta}_\mathcal{D})$. Since $S + \lambda I_d \succeq \lambda I_d$, the Schur complement satisfies $H \succeq \lambda I_{d-r}$, hence $\|H^{-1}\|_{\mathrm{op}} \leq 1/\lambda$. Combining this with $\beta - \hat{\beta}_\mathcal{D} = \lambda A_{11}^{-1}\beta$ gives

$$\|\gamma_\mathcal{D}\| \leq \|S_{21}A_{11}^{-1}\beta\|. \tag{25}$$

Under (BD), conditional on Z and β, the nuisance rows $t_i = U_\perp^\top x_i$ are independent of z_i and distributed as $\mathcal{N}(0, \Lambda_{22})$. Therefore, for any fixed $a \in \mathbb{R}^r$,

$$S_{21}a = \frac{1}{M}\sum_{i=1}^{M} t_i z_i^\top a \;\Big|\; Z, a \sim \mathcal{N}\left(0, \frac{a^\top S_{11}a}{M}\Lambda_{22}\right). \tag{26}$$

Taking $a = A_{11}^{-1}\beta$ and using $S_{11} \succeq \ell_{11} I_r/2$, we obtain $a^\top S_{11} a \le K\|\beta\|^2$. Applying Equation (19) conditionally and then the bound on $\|\beta\|$ gives

$$\|\gamma_{\mathcal{D}}\| \le K R_\perp(\delta)\sqrt{\frac{L_\delta}{M}} = K\eta_M(\delta). \tag{27}$$

Similarly, conditional on Z and the query concept coordinate z, applying the same argument with $a = A_{11}^{-1} z$ gives

$$\|S_{21}A_{11}^{-1}z\| \le K R_\perp(\delta)\frac{\|z\|}{\sqrt{M}}. \tag{28}$$

Consequently,

$$\begin{aligned} |\Delta_{\mathcal{D}}(x_q)| &\le \big(\|t\| + \|S_{21}A_{11}^{-1}z\|\big)\|\gamma_{\mathcal{D}}\| \\ &\le K_\Delta\left(\|t\| + \frac{R_\perp(\delta)}{\sqrt{M}}\|z\|\right)\eta_M(\delta). \end{aligned} \tag{29}$$

The prediction decomposition follows from

$$f_{\mathcal{D}}(x_q) - \langle U\beta, x_q\rangle = z^\top(\hat{\beta}_{\mathcal{D}} - \beta) + \Delta_{\mathcal{D}}(x_q), \tag{30}$$

and the final prediction bound follows by substituting the two estimates above.

It remains to prove invariance. If $v \in \mathrm{span}(U)^\perp$, then $U^\top(x_q + v) = U^\top x_q$ and $U_\perp^\top(x_q + v) = t + U_\perp^\top v$. The concept term and the $z^\top A_{11}^{-1}A_{12}\gamma_{\mathcal{D}}$ correction cancel, so

$$f_{\mathcal{D}}(x_q + v) - f_{\mathcal{D}}(x_q) = \langle \gamma_{\mathcal{D}}, U_\perp^\top v\rangle. \tag{31}$$

Using Equation (27) gives $|f_{\mathcal{D}}(x_q + v) - f_{\mathcal{D}}(x_q)| \le K_\Delta \eta_M(\delta)\|v\|$.

For (NBD), use the Gaussian conditioning formula

$$t_i = \Lambda_{21}\Lambda_{11}^{-1}z_i + \xi_i, \qquad \xi_i \sim \mathcal{N}(0, \Lambda_{22} - \Lambda_{21}\Lambda_{11}^{-1}\Lambda_{12}), \tag{32}$$

where ξ_i is independent of z_i. The residual part is bounded as in the BD case, since its covariance is dominated by Λ_{22}. The conditional mean contributes

$$\|\Lambda_{21}\Lambda_{11}^{-1}S_{11}A_{11}^{-1}\beta\| \le K\rho\|\beta\| \le K_\rho\rho\sqrt{L_\delta} \tag{33}$$

to $\|\gamma_{\mathcal{D}}\|$, and similarly contributes at most $K'_\rho\rho\|z\|$ to $\|S_{21}A_{11}^{-1}z\|$. This yields the stated replacements in the NBD bound. □

A.6 Proof of Corollary 1

Proof. The model $y = Z\beta + \varepsilon$, $\beta \sim \mathcal{N}(0, I_r)$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_M)$ is conjugate Gaussian linear regression Bishop (2006); Murphy (2012). Completing the square in the posterior density gives

$$\Sigma_{\mathcal{D}} = \left(I_r + \frac{1}{\sigma^2}Z^\top Z\right)^{-1}, \qquad \mu_{\mathcal{D}} = \Sigma_{\mathcal{D}}\frac{1}{\sigma^2}Z^\top y. \tag{34}$$

The posterior predictive mean at query x_q is $\mu_{\mathcal{D}}^\top U^\top x_q$. Since

$$\mu_{\mathcal{D}} = (Z^\top Z + \sigma^2 I_r)^{-1}Z^\top y, \tag{35}$$

it coincides with the concept-space ridge estimator $(Z^\top Z + M\lambda I_r)^{-1}Z^\top y$ when $M\lambda = \sigma^2$, i.e., $\lambda = \sigma^2/M$. □

A.7 Proof of Corollary 2

Proof. The ambient ridge statement is exactly the invariance bound in Theorem 1, after renaming constants:

$$|f_{\mathcal{D}}(x_q + v) - f_{\mathcal{D}}(x_q)| \le K_{\mathrm{stab}} R_\perp(\delta)\|v\|\sqrt{\frac{r + \log(a_0/\delta)}{M}}. \tag{36}$$

For the Bayes concept predictor, if $v \in \mathrm{span}(U)^\perp$, then $U^\top(x_q + v) = U^\top x_q$. Hence $\mu_{\mathcal{D}}^\top U^\top(x_q + v) = \mu_{\mathcal{D}}^\top U^\top x_q$, so the Bayes concept predictor is exactly invariant to nuisance perturbations. □

A.8 Proof of Corollary 3

Proof. Let $S_{11} = M^{-1}Z^\top Z$. On the event $S_{11} \succeq \lambda_{\min}(\Lambda_{11})I_r/2$, which holds with probability at least $1 - \delta/a_0$ when $M \geq K_0 L_\delta$, and on the event $\|\beta\| \leq K\sqrt{L_\delta}$, write $y = Z\beta + \varepsilon$. Then

$$\hat{\beta}_{\mathcal{D}} - \beta = -\lambda(S_{11} + \lambda I_r)^{-1}\beta + (S_{11} + \lambda I_r)^{-1}\frac{1}{M}Z^\top\varepsilon. \tag{37}$$

For the bias term, $\lambda = \lambda_0/M$ gives

$$\left\|\lambda(S_{11} + \lambda I_r)^{-1}\beta\right\| \leq K\frac{\sqrt{L_\delta}}{M}. \tag{38}$$

For the variance term, condition on Z. Since $M^{-1}Z^\top\varepsilon \sim \mathcal{N}(0, \sigma^2 S_{11}/M)$, we may write

$$(S_{11} + \lambda I_r)^{-1}\frac{1}{M}Z^\top\varepsilon = (S_{11} + \lambda I_r)^{-1}S_{11}^{1/2}g, \qquad g \sim \mathcal{N}\left(0, \frac{\sigma^2}{M}I_r\right). \tag{39}$$

On $S_{11} \succeq \lambda_{\min}(\Lambda_{11})I_r/2$, the operator norm $\|(S_{11} + \lambda I_r)^{-1}S_{11}^{1/2}\|_{\mathrm{op}}$ is bounded by a constant depending only on Λ_{11}. A χ^2 tail for $\|g\|$ therefore gives

$$\left\|(S_{11} + \lambda I_r)^{-1}\frac{1}{M}Z^\top\varepsilon\right\| \leq K\sigma\sqrt{\frac{L_\delta}{M}} \tag{40}$$

with probability at least $1 - \delta/a_0$. Combining the bias and variance bounds proves Equation (16). The prediction bound follows from

$$|\tilde{f}_{\mathcal{D}}(x_q) - \langle U\beta, x_q\rangle| = |z^\top(\hat{\beta}_{\mathcal{D}} - \beta)| \leq \|z\|\|\hat{\beta}_{\mathcal{D}} - \beta\|. \tag{41}$$

Finally, if $v \in \mathrm{span}(U)^\perp$, then $U^\top(x_q + v) = U^\top x_q$, hence $\tilde{f}_{\mathcal{D}}(x_q + v) = \tilde{f}_{\mathcal{D}}(x_q)$. □

A.9 Noisy ambient decomposition

For completeness, we record the exact ambient decomposition with noisy labels. Let $Z_\perp \in \mathbb{R}^{M\times(d-r)}$ have rows $t_i^\top = (U_\perp^\top x_i)^\top$, and define $e_z = M^{-1}Z^\top\varepsilon$ and $e_t = M^{-1}Z_\perp^\top\varepsilon$. In the basis $[U\ U_\perp]$, the normal equations are

$$\begin{bmatrix} A_{11} & A_{12} \\ S_{21} & A_{22} \end{bmatrix}\begin{bmatrix}\alpha \\ \gamma\end{bmatrix} = \begin{bmatrix} S_{11}\beta + e_z \\ S_{21}\beta + e_t\end{bmatrix}. \tag{42}$$

The concept-space noisy ridge estimator is $\hat{\beta}_{\mathcal{D}} = A_{11}^{-1}(S_{11}\beta + e_z)$. Solving the block equations gives

$$\alpha = \hat{\beta}_{\mathcal{D}} - A_{11}^{-1}A_{12}\gamma_{\mathcal{D}}, \qquad \gamma_{\mathcal{D}} = H^{-1}\big(S_{21}(\beta - \hat{\beta}_{\mathcal{D}}) + e_t\big). \tag{43}$$

Thus

$$f_{\mathcal{D}}(x_q) = z^\top\hat{\beta}_{\mathcal{D}} + \big(t^\top - z^\top A_{11}^{-1}A_{12}\big)\gamma_{\mathcal{D}}. \tag{44}$$

Compared with the noise-free case, the additional term $H^{-1}e_t$ is the off-subspace noise amplification. It is not governed solely by r without further assumptions on the nuisance design, covariance scale, or regularization.

A.10 Proof of Proposition 1

Proof. For each task τ, the label model is $y = \langle U\beta_\tau, x\rangle + \varepsilon$. Since $\mathbb{E}[x\varepsilon \mid \tau] = 0$,

$$m_\tau = \mathbb{E}[xy \mid \tau] = \mathbb{E}[xx^\top]U\beta_\tau + \mathbb{E}[x\varepsilon \mid \tau] = \Lambda U\beta_\tau. \tag{45}$$

Stacking the task-conditioned moments gives $G_{xy} = [m_1, \ldots, m_T] = \Lambda UB$. Because $\Lambda \succ 0$, U has full column rank, and B has rank r, we have $\mathrm{rank}(G_{xy}) = r$. Moreover,

$$\mathrm{span}(\Lambda^{-1}G_{xy}) = \mathrm{span}(UB) = \mathrm{span}(U). \tag{46}$$

If $\beta_\tau \overset{i.i.d.}{\sim} \mathcal{N}(0, I_r)$ and $T \geq r$, then B has rank r almost surely. The remaining ambiguity is only the choice of orthonormal basis within the recovered subspace: any two orthonormal bases of the same r-dimensional subspace differ by an orthogonal matrix O, giving $\tilde{U} = UO$ and $\tilde{\beta}_\tau = O^\top\beta_\tau$. Finally, under the zero-mean task prior, the unconditional first moment is

$$\mathbb{E}[xy] = \mathbb{E}_\beta[\Lambda U\beta] = \Lambda U\mathbb{E}[\beta] = 0, \tag{47}$$

so task conditioning is necessary for this first-moment identification argument. □

A.11 Proof of Proposition 2

Let $X := [x_1, \dots, x_M]^\top \in \mathbb{R}^{M\times d}$, so that $y = Xw$ for $y = (y_1, \dots, y_M)^\top$, let $\mathcal{X} := \operatorname{span}\{x_1, \dots, x_M\}$, and write $y_q := \langle w, x_q\rangle$.

Proof. Step 1: the design and the task prior. First, $\dim \mathcal{X} = M$ almost surely. For each $k \le M$, the subspace $\operatorname{span}\{x_1, \dots, x_{k-1}\}$ has dimension at most $k-1 < d$ and is independent of $x_k \sim \mathcal{N}(0, I_d)$, so it contains x_k with probability zero. Consequently, $XX^\top \succ 0$ and $\operatorname{tr} P_{\mathcal{X}^\perp} = d - M$ almost surely. Second,

$$\mathbb{E}[ww^\top] = \frac{r}{d} I_d. \tag{48}$$

Since β is independent of U and $\mathbb{E}[\beta\beta^\top] = I_r$, we have $\mathbb{E}[ww^\top] = \mathbb{E}[UU^\top]$. For every orthogonal O, the invariance of the law of U gives $O\,\mathbb{E}[UU^\top]O^\top = \mathbb{E}[UU^\top]$. As in the proof of Corollary 4(b) (Appendix A.12), a matrix that commutes with every orthogonal matrix is a multiple cI_d of the identity, and taking traces gives $c = \mathbb{E}\operatorname{tr}(U^\top U)/d = r/d$. In particular, $\mathbb{E}\|w\|^2 = r$.

Step 2: a reflection that fixes the prompt. Let $Q_{\mathcal{X}} := P_{\mathcal{X}} - P_{\mathcal{X}^\perp}$ be the reflection across $\mathcal{X}$. It is orthogonal and a function of X alone. Consider the map

$$(U, \beta, X, x_q) \;\mapsto\; (Q_{\mathcal{X}}U, \beta, X, x_q). \tag{49}$$

It preserves the joint law. Indeed, U is independent of (β, X, x_q), and for each fixed value of (β, X, x_q), left multiplication by the fixed orthogonal matrix $Q_{\mathcal{X}}$ preserves the law of U; Fubini's theorem gives the claim. The map sends $w = U\beta$ to $Q_{\mathcal{X}}w$. Since $XP_{\mathcal{X}} = X$ and $XP_{\mathcal{X}^\perp} = 0$, we have $XQ_{\mathcal{X}}w = Xw$, so the labels $y_{1:M}$ are unchanged. Hence the map leaves the prompt, the query and every predictor $\hat{g}(x_{1:M}, y_{1:M}, x_q)$ unchanged. It also leaves $P_{\mathcal{X}}w$ unchanged and sends $P_{\mathcal{X}^\perp}w$ to $-P_{\mathcal{X}^\perp}w$. It plays the role of the sign flip $\beta \mapsto -\beta$ in the proof of Lemma 4 (Appendix A.12), but acts only on the component of w that the prompt does not reveal.

Step 3: risk decomposition. Let

$$g_\star := \langle P_{\mathcal{X}}w, x_q\rangle \quad \text{and} \quad g_\perp := \langle P_{\mathcal{X}^\perp}w, x_q\rangle, \tag{50}$$

so that $y_q = g_\star + g_\perp$. Since $P_{\mathcal{X}}w = X^+Xw = X^+y$, the quantity $g_\star = \langle X^+y, x_q\rangle$ is itself a predictor, that is, a function of $(x_{1:M}, y_{1:M}, x_q)$. Let $\hat{g}$ be a measurable function of $(x_{1:M}, y_{1:M}, x_q)$ with $\mathbb{E}(\hat{g} - y_q)^2 < \infty$; otherwise there is nothing to prove. Both $|g_\star|$ and $|g_\perp|$ are at most $\|w\|\|x_q\|$, which is square integrable because w and x_q are independent. Hence $\hat{g} - g_\star = (\hat{g} - y_q) + g_\perp$ is square integrable, and

$$\mathbb{E}(\hat{g} - y_q)^2 = \mathbb{E}(\hat{g} - g_\star)^2 - 2\,\mathbb{E}\big[(\hat{g} - g_\star)g_\perp\big] + \mathbb{E}g_\perp^2. \tag{51}$$

The map in Equation (49) preserves the joint law, leaves $\hat{g} - g_\star$ unchanged and changes the sign of $g_\perp$. Therefore $\mathbb{E}[(\hat{g} - g_\star)g_\perp] = -\mathbb{E}[(\hat{g} - g_\star)g_\perp]$, the cross term in Equation (51) vanishes, and

$$\mathbb{E}(\hat{g} - y_q)^2 = \mathbb{E}(\hat{g} - g_\star)^2 + \mathbb{E}g_\perp^2 \;\ge\; \mathbb{E}g_\perp^2, \tag{52}$$

with equality if and only if $\hat{g} = g_\star$ almost surely. The same argument applies if $\hat{g}$ also uses internal randomness that is independent of (U, β, X, x_q), since this randomness can be appended to (β, X, x_q) in Equation (49).

Step 4: the value of the bound. Since x_q is independent of (U, β, X) and $\mathbb{E}[x_q x_q^\top] = I_d$, we have $\mathbb{E}g_\perp^2 = \mathbb{E}\|P_{\mathcal{X}^\perp}w\|^2$. Since w is independent of X, Step 1 gives

$$\begin{aligned}\mathbb{E}g_\perp^2 &= \mathbb{E}\operatorname{tr}\left(P_{\mathcal{X}^\perp}\,\mathbb{E}[ww^\top]\right) = \frac{r}{d}\,\mathbb{E}\operatorname{tr} P_{\mathcal{X}^\perp} \\ &= \frac{r}{d}(d - M) = r\Big(1 - \frac{M}{d}\Big).\end{aligned} \tag{53}$$

Together with Equation (52), this proves the lower bound.

Step 5: the ridge limit attains the bound. With $S = \frac{1}{M}X^\top X$ and $\frac{1}{M}\sum_i x_i y_i = \frac{1}{M}X^\top y$, the ridge proxy in Equation (3) is

$$\hat{w}_{\mathcal{D}} = (X^\top X + M\lambda I_d)^{-1}X^\top y. \tag{54}$$

The identity $(X^\top X + M\lambda I_d)X^\top = X^\top(XX^\top + M\lambda I_M)$ gives

$$\hat{w}_\mathcal{D} = X^\top(XX^\top + M\lambda I_M)^{-1}y. \tag{55}$$

Since $XX^\top \succ 0$ almost surely by Step 1,

$$\hat{w}_\mathcal{D} \to X^\top(XX^\top)^{-1}y = X^+y = P_\mathcal{X}w \quad \text{as } \lambda \to 0. \tag{56}$$

Hence $f_\mathcal{D}(x_q) \to g_\star$ almost surely, and by Equation (52) the limit predictor $g_\star$ attains the bound. The limit can also be taken in the risk. Since $y = Xw$, we have

$$\hat{w}_\mathcal{D} = P_\lambda w \quad \text{with} \quad P_\lambda := X^\top(XX^\top + M\lambda I_M)^{-1}X. \tag{57}$$

This matrix vanishes on $\mathcal{X}^\perp$, and its eigenvalues on $\mathcal{X}$ are $\mu/(\mu + M\lambda)$, where μ ranges over the eigenvalues of $XX^\top$. The computation of Step 4 in Equation (53) then gives

$$\begin{aligned}\mathbb{E}\big(f_\mathcal{D}(x_q) - y_q\big)^2 &= \frac{r}{d}\,\mathbb{E}\,\mathrm{tr}(I_d - P_\lambda)^2 \\ &= r\Big(1 - \frac{M}{d}\Big) + \frac{r}{d}\,\mathbb{E}\sum_\mu \Big(\frac{M\lambda}{\mu + M\lambda}\Big)^2.\end{aligned} \tag{58}$$

Each summand lies in $(0, 1)$ and decreases to 0 as $\lambda \to 0$. The last term in Equation (58) is therefore positive for every $\lambda > 0$ and tends to 0 as $\lambda \to 0$ by dominated convergence. □

Remarks. The zero predictor has risk $\mathbb{E}y_q^2 = \mathbb{E}\|w\|^2 = r$, so without knowledge of the subspace, $M \le d$ demonstrations remove at most a fraction M/d of this risk. Because the bound holds on average over U, every predictor $\hat{g}$ has conditional risk at least $r(1 - M/d)$ given U on a set of subspaces of positive probability, and in particular for the worst-case subspace. The unstructured head of Theorem 2(v) does not depend on U, and its risk $\frac{r(d+1)}{M+d+1}$ exceeds the bound by $\frac{rM(M+1)}{d(M+d+1)}$. The head of Theorem 2(i) stores $\mathrm{span}(U)$ in its weights, so it is not subspace-agnostic. Its risk $\frac{r(r+1)}{M+r+1}$ does not depend on d and is smaller than the bound exactly when $d > M + r + 1$.

A.12 Proof of Theorem 2

Throughout this subsection, we assume Equation (1), Assumption 1 and Assumption 2. We write $W_{\mathrm{OV}} = [W_x, u_y]$ with $W_x \in \mathbb{R}^{D\times d}$ and $u_y \in \mathbb{R}^D$, $P_U := UU^\top$ and $\hat{m} := \hat{m}_\mathcal{D}$, and $y_q := \langle w, x_q\rangle$ for the query target. Since $y_i = \langle U\beta, x_i\rangle$,

$$\hat{m} = SU\beta, \qquad U^\top\hat{m} = S_{11}\beta, \qquad U_\perp^\top\hat{m} = S_{21}\beta. \tag{59}$$

All moments used below are finite. The random quantities are polynomials in Gaussian vectors, and $\mathbb{E}\|h_0(x_q)\|^4 < \infty$.

Lemma 2 (Moments of the head state)**.** *The cross-moment satisfies $\mathbb{E}[\hat{m}\beta^\top] = \Lambda U$, and*

$$\Phi := \mathbb{E}[\hat{m}\hat{m}^\top] = \Big(1 + \frac{1}{M}\Big)\Lambda P_U\Lambda + \frac{\mathrm{tr}\,\Lambda_{11}}{M}\Lambda \succ 0. \tag{60}$$

Moreover, $\Phi = \Lambda J$ with

$$J := \Big(1 + \frac{1}{M}\Big)P_U\Lambda + \frac{\mathrm{tr}\,\Lambda_{11}}{M}I_d \tag{61}$$

invertible, and $J^{-1}U = U\Gamma_c$.

Proof. Since β is independent of S, $\mathbb{E}[S] = \Lambda$ and $\mathbb{E}[\beta\beta^\top] = I_r$, we have $\mathbb{E}[\hat{m}\beta^\top] = \mathbb{E}[S]U = \Lambda U$ and $\Phi = \mathbb{E}[SP_US]$. For a symmetric $C \in \mathbb{R}^{d\times d}$ and $x \sim \mathcal{N}(0, \Lambda)$, Isserlis' theorem (Isserlis, 1918) gives

$$\mathbb{E}[xx^\top Cxx^\top] = 2\Lambda C\Lambda + \mathrm{tr}(C\Lambda)\Lambda. \tag{62}$$

We expand

$$SP_US = \frac{1}{M^2}\sum_{i,j} x_ix_i^\top P_Ux_jx_j^\top \tag{63}$$

into $M(M-1)$ terms with $i \neq j$ and M terms with $i = j$. This yields

$$\mathbb{E}[SP_US] = \frac{M-1}{M}\Lambda P_U\Lambda + \frac{1}{M}\big(2\Lambda P_U\Lambda + \mathrm{tr}(P_U\Lambda)\Lambda\big), \tag{64}$$

which is Equation (60) because $\mathrm{tr}(P_U\Lambda) = \mathrm{tr}\,\Lambda_{11}$. Since $\Lambda P_U\Lambda \succeq 0$ and $\Lambda \succ 0$, we have $\Phi \succ 0$. The factorization $\Phi = \Lambda J$ is immediate, and $J = \Lambda^{-1}\Phi$ is invertible. Finally, $P_U\Lambda U = U\Lambda_{11}$ gives $JU = U\Gamma_{\mathrm{c}}^{-1}$, that is, $J^{-1}U = U\Gamma_{\mathrm{c}}$. □

Lemma 3 (Label permutations)**.** *Let π be a random permutation of $\{1,\dots,M\}$, independent of $(x_{1:M}, \beta)$. Let $N_\pi := \{i : \pi(i) \neq i\}$, let $n_\pi := |N_\pi|$, and let b_π be the number of two-cycles of π. Let $\tilde{m} := \frac{1}{M}\sum_i x_i y_{\pi(i)}$ and $\Delta\hat{m} := \hat{m} - \tilde{m}$. Then*

$$\mathbb{E}[\Delta\hat{m}\,\Delta\hat{m}^\top \mid \pi] = \frac{n_\pi^2 + n_\pi + 2b_\pi}{M^2}\,\Lambda P_U\Lambda + \frac{2n_\pi\,\mathrm{tr}\,\Lambda_{11}}{M^2}\,\Lambda. \tag{65}$$

Proof. Condition on π and on $w = U\beta$, and let $a_i := (x_i - x_{\pi(i)})^\top w$, so that

$$\Delta\hat{m} = \frac{1}{M}\sum_{i\in N_\pi} a_i x_i. \tag{66}$$

Diagonal terms. For $i \in N_\pi$, the vectors x_i and $x_{\pi(i)}$ are independent, so the cross term in a_i^2 has zero mean. Isserlis' theorem in Equation (62) then gives

$$\begin{aligned}\mathbb{E}[a_i^2 x_i x_i^\top] &= \mathbb{E}[(x_i^\top w)^2 x_i x_i^\top] + \mathbb{E}[(x_{\pi(i)}^\top w)^2]\,\Lambda \\ &= 2\Lambda w w^\top\Lambda + 2(w^\top\Lambda w)\Lambda.\end{aligned} \tag{67}$$

Off-diagonal terms. For distinct $i, j \in N_\pi$, we expand $a_i a_j$ into four products. The term $\mathbb{E}[(x_i^\top w)(x_j^\top w)x_i x_j^\top]$ equals $\Lambda w w^\top\Lambda$. The term $\mathbb{E}[(x_{\pi(i)}^\top w)(x_{\pi(j)}^\top w)x_i x_j^\top]$ equals $\Lambda w w^\top\Lambda$ if $\pi(i) = j$ and $\pi(j) = i$, and zero otherwise. The two remaining terms vanish, because each contains a zero-mean vector that appears only once.

Summation. Summing over $i, j \in N_\pi$ gives

$$\big(2n_\pi + n_\pi(n_\pi - 1) + 2b_\pi\big)\Lambda w w^\top\Lambda + 2n_\pi(w^\top\Lambda w)\Lambda.$$

We divide by M^2 and use $\mathbb{E}[ww^\top] = P_U$ and $\mathbb{E}[w^\top\Lambda w] = \mathrm{tr}\,\Lambda_{11}$, which proves Equation (65). □

Lemma 4 (Reduction to a bilinear read-out)**.** *For every θ allowed by Assumption 2, $f(\mathcal{P}) = x_q^\top\Psi(\theta)\hat{m} + g_\theta$, where*

$$\begin{aligned}\Psi(\theta) &:= \Omega(\kappa_y W_x + u_y\kappa_x^\top), \\ g_\theta &:= x_q^\top\Omega h_0(x_q) + x_q^\top\Omega W_x S\kappa_x + \kappa_y\overline{y^2}\,x_q^\top\Omega u_y.\end{aligned} \tag{68}$$

Moreover,

$$\mathcal{L}(\theta) = \mathcal{L}_{\mathrm{odd}}(\Psi(\theta)) + \mathbb{E}\,g_\theta^2, \qquad \mathcal{L}_{\mathrm{odd}}(\Psi) := \mathbb{E}\big(x_q^\top\Psi\hat{m} - y_q\big)^2. \tag{69}$$

The function $\mathcal{L}_{\mathrm{odd}}$ is strictly convex on $\mathbb{R}^{d\times d}$ with unique minimizer $\Psi_\star := U\Gamma_{\mathrm{c}}U^\top$. Consequently, the minimum of $\mathcal{L}$ equals $\mathcal{L}_{\mathrm{odd}}(\Psi_\star)$ and is attained. Moreover, θ minimizes $\mathcal{L}$ if and only if $\Psi(\theta) = \Psi_\star$ and $g_\theta = 0$ almost surely.

Proof. The blocks of $\frac{1}{M}\sum_i e_i e_i^\top$ are S, $\hat{m}$ and $\overline{y^2}$. Hence

$$\frac{1}{M}\sum_i \langle\kappa, e_i\rangle e_i = (S\kappa_x + \kappa_y\hat{m},\ \hat{m}^\top\kappa_x + \kappa_y\overline{y^2}), \tag{70}$$

and

$$h(\mathcal{P}) = h_0(x_q) + W_x S\kappa_x + (\kappa_y W_x + u_y\kappa_x^\top)\hat{m} + \kappa_y\overline{y^2}\,u_y. \tag{71}$$

Applying $x_q^\top\Omega$ gives Equation (68).

Risk decomposition. The map $\beta \mapsto -\beta$ preserves the joint law of $(x_{1:M}, x_q, \beta)$. It changes the signs of $\hat{m}$ and y_q and leaves S, $\overline{y^2}$ and x_q unchanged. Hence $x_q^\top \Psi(\theta)\hat{m} - y_q$ is odd and g_θ is even under this map. Their product therefore has zero mean, which gives Equation (69).

The odd part. Since x_q is independent of the prompt, Lemma 2 gives

$$\mathcal{L}_{\mathrm{odd}}(\Psi) = \mathrm{tr}(\Lambda\Psi\Phi\Psi^\top) - 2\,\mathrm{tr}(\Lambda\Psi\Lambda P_U) + \mathrm{tr}(\Lambda P_U). \tag{72}$$

Under column-major vectorization, its Hessian is $2(\Phi \otimes \Lambda)$. This is positive definite because $\Phi \succ 0$ and $\Lambda \succ 0$, so $\mathcal{L}_{\mathrm{odd}}$ is strictly convex. The first-order condition $\Lambda\Psi\Phi = \Lambda P_U \Lambda$ gives $\Psi = P_U \Lambda \Phi^{-1}$. By Lemma 2,

$$\Psi^\top = \Phi^{-1}\Lambda P_U = J^{-1}P_U = U\Gamma_{\mathrm{c}}U^\top, \tag{73}$$

which is symmetric; hence $\Psi_\star = U\Gamma_{\mathrm{c}}U^\top$. This is the rank-deficient case of the one-layer linear-attention optimum (Zhang et al., 2024; Li et al., 2024).

Attainment. Take $\kappa = (0, 1)$ and $u_y = 0$. Take W_x injective with range in $\mathcal{H}_0^\perp$, which is possible because $\dim \mathcal{H}_0^\perp \geq d$. Take $\Omega := \Psi_\star W_x^+$. Then $\Psi(\theta) = \Psi_\star W_x^+ W_x = \Psi_\star$. Moreover, Ω vanishes on $\mathrm{range}(W_x)^\perp \supseteq \mathcal{H}_0$, so $g_\theta = x_q^\top \Omega h_0(x_q) = 0$.

Conclusion. For every θ,

$$\mathcal{L}(\theta) \geq \mathcal{L}_{\mathrm{odd}}(\Psi(\theta)) \geq \mathcal{L}_{\mathrm{odd}}(\Psi_\star). \tag{74}$$

Equality holds throughout exactly when $\mathbb{E}\, g_\theta^2 = 0$ and $\Psi(\theta) = \Psi_\star$, which proves the remaining claims. □

Proof of Theorem 2. Fix a minimizer θ. By Lemma 4, $\Psi(\theta) = \Psi_\star$ and $g_\theta = 0$ almost surely. We write $\Sigma_\Delta := \mathbb{E}[\Delta\hat{m}\,\Delta\hat{m}^\top]$, with the expectation also taken over π. Averaging Equation (65) gives

$$\begin{aligned} \Sigma_\Delta &= \bar{a}\,\Lambda P_U \Lambda + c_\Delta \Lambda, \\ \bar{a} &:= \frac{\mathbb{E}[n_\pi^2 + n_\pi + 2b_\pi]}{M^2}, \qquad c_\Delta := \frac{2\,\mathbb{E}[n_\pi]\,\mathrm{tr}\,\Lambda_{11}}{M^2} > 0, \end{aligned} \tag{75}$$

where $c_\Delta > 0$ because π differs from the identity with positive probability.

Proof of (i): the head. Because h_0 is continuous, g_θ is a continuous function of $(x_{1:M}, \beta, x_q)$. It vanishes almost surely under a Gaussian law with full support, and hence vanishes identically. Setting $\beta = 0$ (so that $y_{1:M} = 0$) and $x_{1:M} = 0$ gives

$$x_q^\top \Omega h_0(x_q) \equiv 0. \tag{76}$$

Setting $\beta = 0$, $x_1 = v$ and $x_2 = \cdots = x_M = 0$ gives

$$\frac{1}{M}(x_q^\top \Omega W_x v)(v^\top \kappa_x) = 0 \quad \text{for all } v, x_q \in \mathbb{R}^d. \tag{77}$$

Suppose $\kappa_x \neq 0$. Then $\Omega W_x v = 0$ on the dense set $\{v : v^\top \kappa_x \neq 0\}$, so $\Omega W_x = 0$. In that case $\Psi_\star = \Psi(\theta) = \Omega u_y \kappa_x^\top$ would have rank at most $1 < r$, a contradiction. Therefore $\kappa_x = 0$. (For $r = 1$ this step fails: $\kappa = (U, 0)$ with $\Omega W_x = 0$ and $\Omega u_y = \Psi_\star U$ also attains the minimum, so the assumption $r \geq 2$ cannot be dropped.) Next, $\Psi_\star = \kappa_y \Omega W_x \neq 0$ forces $\kappa_y \neq 0$. The remaining identity $\kappa_y \overline{y^2}\, x_q^\top \Omega u_y \equiv 0$, evaluated at a prompt with $\overline{y^2} > 0$, forces $\Omega u_y = 0$. With $W := \kappa_y W_x$ and $v_y := \kappa_y u_y$, Equation (71) becomes

$$h(\mathcal{P}) = h_0(x_q) + W\hat{m} + \overline{y^2}\, v_y, \qquad \Omega W = \Psi_\star, \qquad \Omega v_y = 0. \tag{78}$$

By Equation (59),

$$f(\mathcal{P}) = x_q^\top \Psi_\star \hat{m} = (U^\top x_q)^\top \Gamma_{\mathrm{c}} S_{11} \beta, \tag{79}$$

which is the prediction in Theorem 2(i). Moreover, with $W_{\mathrm{c}} := WU\Gamma_{\mathrm{c}}^{-1}$,

$$W\hat{m} = WUS_{11}\beta + WU_\perp S_{21}\beta = W_{\mathrm{c}}\hat{\beta}(\mathcal{P}) + WU_\perp U_\perp^\top \hat{m}.$$

Hence, with $\nu(\mathcal{P}) := WU_\perp U_\perp^\top \hat{m} + \overline{y^2}\, v_y$,

$$h(\mathcal{P}) = h_0(x_q) + W_{\mathrm{c}}\hat{\beta}(\mathcal{P}) + \nu(\mathcal{P}), \tag{80}$$

$$\text{and} \quad \Omega\nu = \Psi_\star U_\perp U_\perp^\top \hat{m} + \overline{y^2}\,\Omega v_y = 0. \tag{81}$$

Proof of (i): the read-out. Let $\mathcal{F} := \mathrm{range}[W, v_y]$, so that $\mathcal{F} \perp \mathcal{H}_0$ by Assumption 2. For the fixed head, Lemma 4 and the argument above show that a read-out minimizes $\mathcal{L}$ if and only if it maps $[W, v_y]$ to $[\Psi_\star, 0]$ and satisfies $x_q^\top \Omega h_0(x_q) \equiv 0$. The first condition constrains only the restriction of the read-out to $\mathcal{F}$. The second holds whenever the read-out vanishes on $\mathcal{F}^\perp \supseteq \mathcal{H}_0$. The squared Frobenius norm is the sum of the squared norms of these two restrictions. The minimum-norm read-out is therefore

$$\Omega = [\Psi_\star, 0][W, v_y]^+, \tag{82}$$

which vanishes on $\mathcal{F}^\perp$. Hence

$$r = \mathrm{rank}(\Omega W) \le \mathrm{rank}\,\Omega \le \mathrm{rank}[\Psi_\star, 0] = r,$$

that is, $\dim \mathcal{R} = r$. If $v_y \perp \mathrm{range}(W)$, the pseudoinverse of $[W, v_y]$ stacks W^+ on top of v_y^+, so $\Omega = \Psi_\star W^+$. If moreover $W^\top W = sI_d$, then

$$\Omega = \frac{1}{s}\Psi_\star W^\top \quad \text{and} \quad \mathcal{R} = W\,\mathrm{span}(U), \quad \text{with} \quad P_\mathcal{R} = \frac{1}{s} W P_U W^\top. \tag{83}$$

Proof of (ii): exact identities. Since $\mathcal{R}$ is the row space of Ω, $\Omega = \Omega P_\mathcal{R}$. The patched prediction is

$$f_V = \tilde{f} + x_q^\top \Omega P_V \big(h(\mathcal{P}) - h(\tilde{\mathcal{P}})\big). \tag{84}$$

If $V \supseteq \mathcal{R}$, then $\Omega P_V = \Omega$, so $f_V = f$. If $V \perp \mathcal{R}$, then $\Omega P_V = \Omega P_\mathcal{R} P_V = 0$, so $f_V = \tilde{f}$. For a swap between prompts $\mathcal{P}_1, \mathcal{P}_2$ with the same query,

$$x_q^\top \Omega\big(h(\mathcal{P}_1) - P_\mathcal{R} h(\mathcal{P}_1) + P_\mathcal{R} h(\mathcal{P}_2)\big) = x_q^\top \Omega h(\mathcal{P}_2) = f(\mathcal{P}_2). \tag{85}$$

These identities use only $\Omega = \Omega P_\mathcal{R}$ and the shared query. They therefore hold for any corruption that keeps the query.

Proof of (ii): the recovery. A label permutation leaves S and $\overline{y^2}$ unchanged. By Equation (78), $h(\mathcal{P}) - h(\tilde{\mathcal{P}}) = W\Delta\hat{m}$, so

$$f_V - \tilde{f} = x_q^\top \Omega P_V W \Delta\hat{m} \quad \text{and} \quad f - \tilde{f} = x_q^\top \Psi_\star \Delta\hat{m}. \tag{86}$$

Because x_q is independent of $\Delta\hat{m}$,

$$\mathrm{Rec}(V) = \frac{\mathrm{tr}\left(\Lambda\,\Omega P_V W \Sigma_\Delta \Psi_\star\right)}{\mathrm{tr}\left(\Lambda \Psi_\star \Sigma_\Delta \Psi_\star\right)}. \tag{87}$$

By Equation (75), the denominator is at least $c_\Delta \|\Lambda^{1/2}\Psi_\star\Lambda^{1/2}\|_F^2 > 0$. The numerator is linear in P_V. Since $P_V + P_{V^\perp} = I_D$ and $\mathrm{Rec}(\mathbb{R}^D) = 1$, we get

$$\mathrm{Rec}(V) + \mathrm{Rec}(V^\perp) = 1. \tag{88}$$

For V uniformly distributed on the k-dimensional subspaces of $\mathbb{R}^D$ and independent of the data, $\mathbb{E}P_V = \frac{k}{D} I_D$, so $\mathbb{E}\,\mathrm{Rec}(V) = k/D$.

Proof of (ii): the isotropic case. Now let (BD) hold, $W^\top W = sI_d$ and $v_y \perp \mathrm{range}(W)$, so that $\Omega = \frac{1}{s}\Psi_\star W^\top$. Under (BD), $\Lambda U = U\Lambda_{11}$, and Equation (75) gives $\Sigma_\Delta U = U(\bar{a}\Lambda_{11}^2 + c_\Delta \Lambda_{11})$. Consequently $\Sigma_\Delta \Psi_\star \Lambda \Psi_\star = U\Xi U^\top$ with

$$\Xi := (\bar{a}\Lambda_{11}^2 + c_\Delta\Lambda_{11})\Gamma_c \Lambda_{11} \Gamma_c \succ 0,$$

a product of commuting positive definite functions of Λ_{11}. Substituting into Equation (87) and using the cyclicity of the trace,

$$\mathrm{Rec}(V) = \frac{\frac{1}{s}\,\mathrm{tr}\left(W^\top P_V W\, U\Xi U^\top\right)}{\mathrm{tr}\,\Xi} = \mathrm{tr}(P_V N'), \qquad N' := \frac{WU\Xi U^\top W^\top}{s\,\mathrm{tr}\,\Xi}. \tag{89}$$

The matrix N' is positive semidefinite, its range is $W\,\mathrm{span}(U) = \mathcal{R}$, and

$$\mathrm{tr}\,N' = \mathrm{tr}(\Xi U^\top W^\top W U)/(s\,\mathrm{tr}\,\Xi) = 1. \tag{90}$$

Hence $0 \le \mathrm{Rec}(V) \le 1$. If $V \subseteq V'$, then $P_{V'} - P_V$ is an orthogonal projector, and

$$\mathrm{Rec}(V') - \mathrm{Rec}(V) = \|(N')^{1/2}(P_{V'} - P_V)\|_F^2 \ge 0.$$

If moreover $\Lambda_{11} \propto I_r$, then $\Xi \propto I_r$, so $N' = \frac{1}{r} P_{\mathcal{R}}$ and

$$\mathrm{Rec}(V) = \frac{1}{r} \operatorname{tr}(P_V P_{\mathcal{R}}) = \frac{1}{r} \sum_i \cos^2 \phi_i, \tag{91}$$

where the ϕ_i are the principal angles between V and $\mathcal{R}$.

Proof of (iii): the cross-covariance. Since $\mathbb{E}\|Y\|^2 < \infty$ and $\mathbb{E}\|h(\mathcal{P})\|^2 < \infty$, the cross-moment exists. Because $h(\mathcal{P})$ is a function of $(\beta, x_{1:M}, x_q)$, the tower property gives

$$\mathbb{E}[h(\mathcal{P})Y^\top] = \mathbb{E}[h(\mathcal{P})\beta^\top \Theta(x_q)^\top]. \tag{92}$$

Here $\mathbb{E}\|\Theta(x_q)\|_F^2 = \mathbb{E}\|\Theta(x_q)\beta\|^2 \le \mathbb{E}\|Y\|^2$, so every term below is integrable. We substitute Equation (78) term by term. The h_0 term vanishes because β has mean zero and is independent of x_q. The v_y term vanishes because $\beta \mapsto -\beta$ preserves the joint law, leaves $\overline{y^2}$ and x_q unchanged, and changes the sign of β. The remaining term: $(\hat{m}, \beta)$ is independent of x_q, so Lemma 2 gives

$$\mathbb{E}[W\hat{m}\beta^\top \Theta(x_q)^\top] = W\Lambda U\, \mathbb{E}[\Theta(x_q)]^\top. \tag{93}$$

As $\mathbb{E}[\Theta(x_q)]^\top \in \mathbb{R}^{r\times p}$ has rank r, the column space of $\mathbb{E}[h(\mathcal{P})Y^\top]$ is $W\,\mathrm{span}(\Lambda U) = \mathcal{W}$.

Proof of (iii): the other estimators. Centroids. $\mathbb{E}[x_i y_{\pi(i)} \mid \beta, \pi]$ equals $\Lambda U\beta$ if $\pi(i) = i$ and 0 otherwise. Hence

$$\mathbb{E}[h(\mathcal{P}) - h(\tilde{\mathcal{P}}) \mid \beta] = \frac{\mathbb{E}[n_\pi]}{M} W\Lambda U\beta =: c_\pi W\Lambda U\beta, \tag{94}$$

with $\mathbb{E}[n_\pi] > 0$. These centroids span $\mathcal{W}$ whenever the relations' β span $\mathbb{R}^r$. *Principal components.* For $\Lambda = I_d$ and $W^\top W = sI_d$, Equation (75) gives

$$\mathbb{E}\big[(h(\mathcal{P}) - h(\tilde{\mathcal{P}}))(h(\mathcal{P}) - h(\tilde{\mathcal{P}}))^\top\big] = W\Sigma_\Delta W^\top = s\bar{a}\, P_{\mathcal{W}} + c_\Delta WW^\top. \tag{95}$$

Its top-r eigenspace is $\mathcal{W}$. *Independent copy.* If Y' is independent of $(\beta, x_{1:M}, x_q)$ and has the law of Y, then $\mathbb{E}Y' = \mathbb{E}[\Theta(x_q)]\,\mathbb{E}\beta = 0$, so $\mathbb{E}[h(\mathcal{P})Y'^\top] = 0$.

Proof of (iii): identification and threshold. If $W^\top W = sI_d$ and $v_y \perp \mathrm{range}(W)$, then $\mathcal{R} = W\,\mathrm{span}(U)$, as shown in Equation (83). Under (BD), $\Lambda U = U\Lambda_{11}$, so $\mathcal{W} = W\,\mathrm{span}(U) = \mathcal{R}$. For the threshold, let $s_1 \ge \cdots \ge s_{m'} > 0$ be the nonzero singular values of $\mathbb{E}[h(\mathcal{P})Y^\top]$, so that $m' \le r$. The selected rank at threshold $\omega \in (0,1)$

$$\hat{r} := \min\Big\{k : \sum_{i\le k} s_i^2 \ge \omega \sum_i s_i^2\Big\} \tag{96}$$

is at most $m' \le r$. It equals r only if $m' = r$ and $s_r^2 > (1-\omega)\sum_i s_i^2$. Since $rs_r^2 \le \sum_i s_i^2$, this requires $r < 1/(1-\omega)$.

Proof of (iv). Since S is independent of β with $\mathbb{E}[S_{21}] = \Lambda_{21}$, and since $\mathbb{E}[\overline{y^2} \mid \beta] = \beta^\top\Lambda_{11}\beta$,

$$\mathbb{E}[\nu(\mathcal{P}) \mid \beta] = WU_\perp\Lambda_{21}\beta + (\beta^\top\Lambda_{11}\beta)\, v_y. \tag{97}$$

Moreover,

$$\|WU_\perp\Lambda_{21}\beta\| \le \|W\|_{\mathrm{op}}\|\Lambda_{12}\|_{\mathrm{op}}\|\beta\| \le \rho\|W\|_{\mathrm{op}}\|\beta\| \quad \text{under (NBD).} \tag{98}$$

Now assume (BD). Then $t_i = U_\perp^\top x_i \sim \mathcal{N}(0, \Lambda_{22})$ is independent of z_i. Given $(z_{1:M}, \beta)$, the labels are fixed and

$$S_{21}\beta = \frac{1}{M}\sum_i t_i y_i \sim \mathcal{N}\Big(0, \frac{\overline{y^2}}{M}\Lambda_{22}\Big). \tag{99}$$

Hence the conditional law of $\nu(\mathcal{P})$ depends on $(z_{1:M}, \beta)$ only through $\overline{y^2}$. Given β, the variables $z_i^\top\beta$ are i.i.d. $\mathcal{N}(0, \beta^\top\Lambda_{11}\beta)$, so $\overline{y^2}$ has the law of $(\beta^\top\Lambda_{11}\beta)\,\chi^2_M/M$. Therefore the law of $\nu(\mathcal{P})$ given β depends on β only through $\beta^\top\Lambda_{11}\beta$.

Finally, if $W^\top W = sI_d$ and $v_y \perp \mathrm{range}(W)$, then $P_{\mathcal{R}} = \frac{1}{s} WP_U W^\top$. This projector annihilates $h_0(x_q) \in \mathcal{H}_0$, v_y and $\mathrm{range}(WU_\perp)$, while $P_{\mathcal{R}} W\hat{m} = WP_U\hat{m} = W_c\hat{\beta}(\mathcal{P})$. Hence

$$P_{\mathcal{R}^\perp} h(\mathcal{P}) = h_0(x_q) + \nu(\mathcal{P}). \tag{100}$$

Proof of (v). Apply Lemma 4 with U replaced by I_d, so that r becomes d and Λ_{11} becomes I_d. This identifies the minimizer for the unstructured prior as $\Psi(\theta) = \zeta_d I_d$ with $\zeta_d := M/(M+d+1)$ (Zhang et al., 2024). Since $\operatorname{rank}(\zeta_d I_d) = d \geq 2$, the argument for (i) applies. The head therefore satisfies Equation (78) with $\Omega W = \zeta_d I_d$, and its minimum-norm read-out is

$$\Omega = \zeta_d W^+ = \frac{\zeta_d}{s} W^\top. \tag{101}$$

The estimators in (iii) depend only on the states, so on concept-family prompts they return $\mathcal{W} = W \operatorname{span}(U)$. On these prompts, Equation (87) holds with $\Psi_\star$ replaced by $\zeta_d I_d$. Because $W^\top P_{\mathcal{W}} W = s P_U$,

$$\operatorname{Rec}(\mathcal{W}) = \frac{\frac{\zeta_d^2}{s} \operatorname{tr}\left(W^\top P_{\mathcal{W}} W \Sigma_\Delta\right)}{\zeta_d^2 \operatorname{tr}(\Sigma_\Delta)} = \frac{\operatorname{tr}(P_U \Sigma_\Delta)}{\operatorname{tr}(\Sigma_\Delta)}. \tag{102}$$

For a uniformly random permutation and $M \geq 2$,

$$\mathbb{E}[n_\pi] = M - 1, \qquad \mathbb{E}[n_\pi^2] = M^2 - 2M + 2 \quad \text{and} \quad \mathbb{E}[b_\pi] = \frac{1}{2}. \tag{103}$$

With $\Lambda = I_d$, Equation (75) becomes

$$M^2 \Sigma_\Delta = (M^2 - M + 2) P_U + 2(M-1) r \, I_d. \tag{104}$$

Hence

$$M^2 \operatorname{tr}(P_U \Sigma_\Delta) = r\big(M^2 - M + 2 + 2(M-1)r\big),$$
$$M^2 \operatorname{tr}(\Sigma_\Delta) = r\big(M^2 - M + 2 + 2(M-1)d\big),$$

which gives the stated value. For the bound, note that $M^2 - M + 2 \leq 2M(M-1)$ for $M \geq 2$. Therefore

$$\operatorname{Rec}(\mathcal{W}) \leq \frac{M^2 - M + 2}{2(M-1)d} + \frac{r}{d} \leq \frac{M + r}{d}. \tag{105}$$

The identity $\operatorname{Rec}(\mathcal{W}^\perp) = 1 - \operatorname{Rec}(\mathcal{W})$ is the additivity of (ii) in Equation (88), whose proof uses only that the read-out is linear in the state. □

Corollary 4 (Written subspace and padding)**.** *Under the assumptions of Theorem 2, let*

$$\mathcal{W} := W \operatorname{span}(\Lambda U) \quad \textit{and} \quad N := W \Sigma_\Delta \Psi_\star \Lambda \Omega. \tag{106}$$

(a) For every subspace $V \subseteq \mathbb{R}^D$,

$$\operatorname{Rec}(V) = \operatorname{tr}(P_V N) / \operatorname{tr}(N), \tag{107}$$

where $\operatorname{tr}(N) > 0$ and $\operatorname{range}(N) \subseteq \mathcal{W}$. Consequently, for every $\Lambda \succ 0$, $\operatorname{Rec}(V) = 1$ if $V \supseteq \mathcal{W}$ and $\operatorname{Rec}(V) = 0$ if $V \perp \mathcal{W}$. Moreover,

$$|\operatorname{Rec}(V^\perp)| \leq \frac{\|N\|_F}{\operatorname{tr}(N)} \|P_{V^\perp} P_{\mathcal{W}}\|_F. \tag{108}$$

(b) Let V_0 be a fixed k_0-dimensional subspace, and let $\mathcal{V}'$ be uniformly distributed on the k_1-dimensional subspaces of $V_0^\perp$, independent of the data. Then

$$\mathbb{E} \operatorname{Rec}(V_0 \oplus \mathcal{V}') = \operatorname{Rec}(V_0) + \frac{k_1}{D - k_0} \operatorname{Rec}(V_0^\perp). \tag{109}$$

Proof. Part (a). By Equation (87) and the cyclicity of the trace, the numerator of $\operatorname{Rec}(V)$ is $\operatorname{tr}(P_V N)$ and the denominator is $\operatorname{tr}(N) > 0$. Since $\Lambda P_U \Lambda U = \Lambda U \Lambda_{11}$, Equation (75) gives

$$\Sigma_\Delta \Psi_\star = \Lambda U (\bar{a} \Lambda_{11} + c_\Delta I_r) \Gamma_c U^\top, \tag{110}$$

so $\operatorname{range}(N) \subseteq W \operatorname{span}(\Lambda U) = \mathcal{W}$. It follows that $P_V N = N$ if $V \supseteq \mathcal{W}$ and $P_V N = 0$ if $V \perp \mathcal{W}$. Finally, $N = P_{\mathcal{W}} N$, and the Cauchy–Schwarz inequality for the Frobenius inner product gives

$$|\operatorname{tr}(P_{V^\perp} N)| = |\operatorname{tr}(P_{V^\perp} P_{\mathcal{W}} N)| \leq \|P_{V^\perp} P_{\mathcal{W}}\|_F \|N\|_F,$$

which is Equation (108).

Part (b). Since $V_0 \perp \mathcal{V}'$, we have $P_{V_0 \oplus \mathcal{V}'} = P_{V_0} + P_{\mathcal{V}'}$. The law of $\mathcal{V}'$ is invariant under every orthogonal map that fixes V_0 and acts on $V_0^\perp$. Hence $\mathbb{E} P_{\mathcal{V}'}$ commutes with all such maps and vanishes on V_0, so $\mathbb{E} P_{\mathcal{V}'} = c P_{V_0^\perp}$ for some scalar c. Taking traces gives $c = k_1/(D - k_0)$. Since Rec is linear in P_V, Equation (109) follows. □

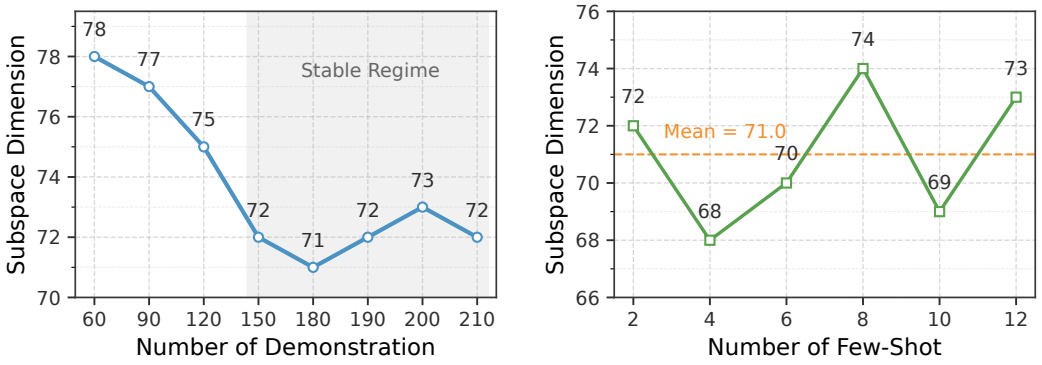

Figure 3: Dimension stability of concept-subspace estimates. The 98% criterion selects a consistently low rank across estimation-set sizes and shot counts.

Remarks. Part (a) links the estimators of (iii) to patching without the isotropy conditions of (ii). It concerns Rec only: $f_{\mathcal{W}} = f$ holds pointwise when $\mathcal{W} = \mathcal{R}$, but not in general. Part (b) with $k_0 = 0$ gives $\mathbb{E}\,\mathrm{Rec}(V) = k/D$. For $k_0 > 0$, directions that behave like random directions in $V_0^{\perp}$ recover only a fraction $k_1/(D - k_0)$ of what V_0 leaves to its complement. The additivity of Theorem 2(ii) in Equation (88) and the padding identity rely on the linearity of Rec in P_V. A recovery rate based on accuracy or risk agrees with Rec in the exact cases of (ii), where $f_V \in \{f, \tilde{f}\}$ pointwise, but it need not be additive otherwise.

B Experimental Details and Concept-Subspace Analyses

B.1 Implementation Details

Models and infrastructure. We run experiments on open-source language models, primarily Llama-3-8B and Qwen2.5-7B. To facilitate precise extraction and manipulation of internal representations, we implement all forward passes and causal interventions (activation patching) using the `transformer_lens` library. All experiments are executed on one NVIDIA V100 GPU. Our software environment is built upon Python 3.10, PyTorch 2.0, and CUDA 11.8.

Dataset and prompt construction. Our task and relation data are derived from the CounterFact dataset. To ensure robust statistical estimation, we filter the dataset to retain only relations with a sufficient number of unique subject-target pairs (e.g., $N \geq 90$ depending on the specific experiment). When constructing n-shot in-context learning prompts, we randomly sample n demonstration pairs (e.g., $n = 12$) from the same relation pool, ensuring that the demonstration subjects are disjoint from the test query subject to prevent data leakage. For tasks requiring different prompt formats (e.g., symbolic, QA, or verbose), the same set of subject-target pairs is injected into distinct string templates to control for semantic equivalence.

Subspace estimation protocol. To estimate the concept subspace $\widehat{U}_{\ell^\star}$ at a target layer $\ell^\star$ (typically late-middle layers such as 26 or 30), we collect the residual stream activations at the final sequence position (the query token). Depending on the experiment, the subspace is extracted either via the cross-covariance between the hidden states $\mathbf{H}$ and the target unembedding vectors Y ($W_U[:, y^*]$ for the first token y^* of the correct answer), or via counterfactual ablation ($\Delta\mathbf{H} = \mathbf{H}_{\text{task}} - \mathbf{H}_{\text{entity}}$). We then perform Singular Value Decomposition (SVD). Instead of fixing a hard-coded dimension, we dynamically determine the rank $\hat{r}$ by selecting the top singular vectors that account for a predefined threshold of the explained variance (set to 98% across all main experiments).

B.2 Stability of the Selected Rank

To test whether the low rank reflects a stable task-aligned representation rather than an extraction artifact, we conduct the subspace estimation procedure at $\ell^\star = 30$ under the same CounterFact setup

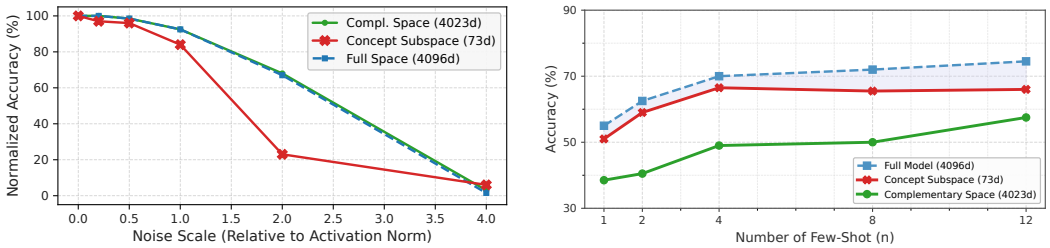

Figure 4: Subspace-selective sensitivity and few-shot scaling. *Left:* concept-aligned noise degrades accuracy more sharply than matched-norm full-space or complementary-space noise. *Right:* concept-subspace interventions recover most of the few-shot gains across shot counts.

while varying (i) the number of examples used for subspace estimation and (ii) the number of in-context demonstrations. In all cases, the estimated rank is the smallest $\hat{r}$ satisfying $\sum_{i=1}^{\hat{r}} s_i^2 / \sum_i s_i^2 \times 100\% \geq 98\%$. Figure 3 (left) shows that the selected rank decreases from 69 to 61 as the estimation set grows and then stabilizes at 61–62 for dataset sizes 150–210. This is consistent with small-sample overestimation: under a conservative 98% cutoff, limited data retain noisy tail directions, whereas more data suppress them without revealing a larger latent task manifold. Figure 3 (right) shows a similarly stable pattern across prompt lengths: varying the few-shot count from 2 to 12 keeps the rank within 68–74 (mean 71.0), with no monotonic trend. Thus, more demonstrations strengthen evidence for the same relation but do not create new task directions.

Taken together, these robustness checks strengthen the main mechanistic claim. They do not imply that there is a single exact "true" rank recoverable from data, and the absolute number depends on the estimator and on the 98% threshold. Rather, they establish a more important and defensible conclusion: across a reasonable range of estimation sizes and prompt configurations, the model consistently exhibits a sharply low-dimensional, task-aligned operating subspace, whose dimension stays in a narrow band from 60 to 70. This is the level of stability needed for the causal interpretation in the main experiments. The concept-only patching results are therefore not a brittle consequence of a single hand-picked extraction setting, but rather evidence for a stable geometric bottleneck in the model's task representation.

B.3 Noise Sensitivity, Few-Shot Scaling and Layerwise Emergence

Subspace-selective noise sensitivity. We next test whether the model is selectively sensitive to perturbations in the learned concept subspace. In the same CounterFact multi-relation setting, we estimate a 73-dimensional concept projector $P_{\widehat{U}_{\ell^\star}} = \widehat{U}_{\ell^\star}\widehat{U}_{\ell^\star}^\top$ at the intervention layer using the 98% explained-variance threshold. For correctly answered few-shot examples, we sample Gaussian noise ξ, decompose it into a concept-aligned component $P_{\widehat{U}_{\ell^\star}}\xi$ and a complementary component $(I - P_{\widehat{U}_{\ell^\star}})\xi$, rescale each to the same norm ratio relative to the activation norm, and patch it into the residual stream; isotropic full-space noise with matched norm serves as a baseline. This isolates the effect of perturbation direction rather than magnitude.

Figure 4 (left) shows a clear directional asymmetry. At small noise scales, all perturbations have a limited effect, but in the moderate-noise regime, concept-subspace noise degrades accuracy much more rapidly than complementary-space or full-space noise: around $2\times$ the activation norm, concept-aligned perturbations reduce accuracy to the low-20% range, whereas the other two still retain roughly two-thirds accuracy. Only at the largest scale do all perturbations collapse performance. This pronounced sensitivity confirms that task-critical information is geometrically concentrated. Because the model can tolerate significant noise in the complementary space but rapidly degrades when the concept subspace is disrupted, we conclude that this low-dimensional subspace acts as the central causal bottleneck controlling the model's logical inference.

Few-shot scaling of subspace-mediated ICL gains. Using the same multi-relation CounterFact setup as above, we vary the number of few-shot examples $n \in \{1, 2, 4, 8, 12\}$ and ask whether the learned concept subspace remains the dominant carrier of recoverable few-shot gains as the prompt

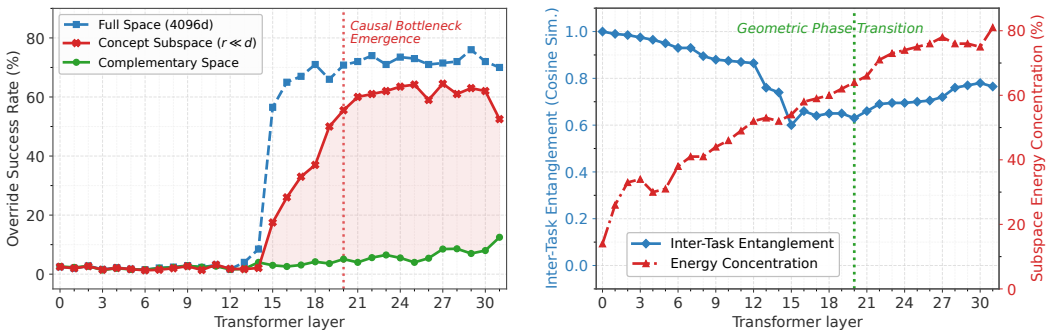

Figure 5: Layerwise emergence of a causal concept bottleneck. *Left:* override success for full-state, concept-subspace, and complementary-space interventions across layers. *Right:* inter-task entanglement decreases while concept-subspace energy concentration increases with depth.

grows. For each query, we compute the activation change between the n-shot run and its zero-shot counterpart at the selected layer $\ell^\star$, decompose this change into the learned concept subspace $\widehat{U}_{\ell^\star}$ (73d) and its orthogonal complement, and inject each component into the zero-shot run. We then compare the resulting accuracy with the original full n-shot model.

Figure 4 (right) shows a stable ordering across all shot counts: full-model performance is highest, concept-subspace recovery is consistently second, and complementary-space recovery is substantially weaker. Notably, the 73-dimensional concept subspace already recovers most of the few-shot benefit, increasing from 51% at 1-shot to about 66% from 4 shots onward, while the full model rises from 55% to 75%. This suggests that the core relation-level rule is quickly compressed into a low-dimensional state: once a few examples identify the task, most of the transferable gain is already available within the learned bottleneck.

The complementary space is not uninformative: its accuracy increases from 39% to 58% as n grows, yet it remains substantially less predictive than the learned subspace. The result supports a *concentrated-mediator* claim rather than an exclusivity claim: the concept subspace carries the dominant recoverable task signal, while off-subspace directions likely encode auxiliary information such as calibration, entity-specific disambiguation, or other distributed contextual effects that become more useful at larger n. The widening gap between the full model and concept-only intervention at higher shot counts is therefore informative rather than contradictory, indicating that additional demonstrations refine the computation beyond the current linear subspace estimate.

Layerwise emergence of a causal concept bottleneck. Using the same CounterFact multi-relation setting, we estimate a layer-specific subspace $\widehat{U}_\ell$ at each layer with the same 98% explained-variance criterion, and repeat the relation-swap intervention with three replacements: the full residual state, only the projection onto $\widehat{U}_\ell$, or only its orthogonal complement. We report the resulting layerwise override success rate.

Figure 5 (left) shows a clear stagewise pattern. In shallow layers, all three interventions remain near the floor, indicating that relational identity is not yet represented in a causally usable form. In the middle of the network, full-state replacement rises first while concept-subspace replacement lags behind, suggesting that task-relevant information is present before it is compressed into a compact low-dimensional representation. In later layers, concept-subspace replacement becomes consistently strong and approaches full-state performance, whereas complementary-space replacement remains near zero. This separation supports a late-layer causal bottleneck: most manipulable task identity is concentrated in a small set of task-aligned directions rather than distributed across the ambient 4096-dimensional residual stream. Since full-state replacement remains somewhat stronger, we interpret the learned subspace as a dominant but not exhaustive mediator.

We also track two geometric diagnostics. Let $P_{\widehat{U}_\ell} = \widehat{U}_\ell \widehat{U}_\ell^\top$ and let $h_\ell^{(R,s)}$ be the query-token residual state for relation R and subject s. For subject-matched relation pairs, define inter-task entanglement $\mathrm{Ent}_\ell = \mathbb{E}_{s,R_1 \neq R_2}[\cos(P_{\widehat{U}_\ell} h_\ell^{(R_1,s)}, P_{\widehat{U}_\ell} h_\ell^{(R_2,s)})]$ and energy concentration $C_\ell =$

Table 2: Task–format separability across representation spaces. Concept coordinates improve relation separability while suppressing format separability.

Space	Task silhouette	Format silhouette
Full space	0.112	0.129
Concept subspace	0.411	-0.007
Complementary space	0.022	0.173

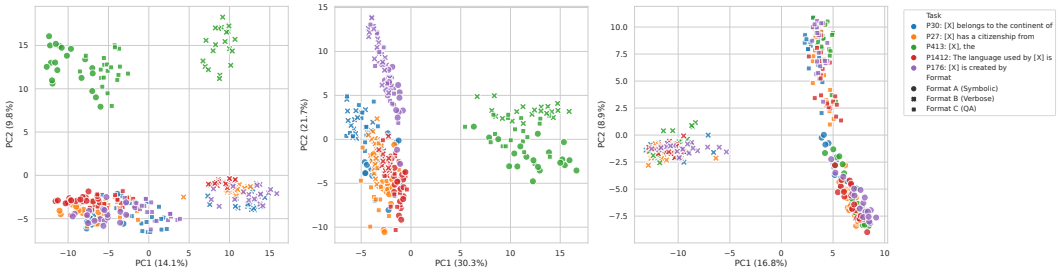

Figure 6: Task–format geometry after concept-subspace projection. PCA of task vectors shows that concept coordinates cluster primarily by relation, whereas complementary directions retain more prompt-format structure. *Left:* full space; *Middle:* concept subspace; *Right:* complementary space.

$100 \cdot \mathbb{E}_{s,R}[\|P_{\widehat{U}_\ell} h_\ell^{(R,s)}\|_2 / \|h_\ell^{(R,s)}\|_2]$. Here Ent_ℓ measures inter-task entanglement in the learned subspace, while C_ℓ measures the fraction of residual-state norm captured by that subspace. Figure 5 (right) shows the corresponding geometric picture. In shallow layers, subject-matched prompts for different relations have high inter-task entanglement, consistent with representations that are still dominated by shared entity or surface-context information. As depth increases, this entanglement drops from its shallow-layer value, while the concept-subspace concentration increases. Thus, the same layer range in which concept-subspace interventions become effective is also where relation-specific directions become more separated and residual variation is increasingly captured by the learned subspace. We interpret this as geometric consolidation of task identity rather than a strict phase transition: the trend supports the emergence of a compact causal bottleneck, but does not imply that all task-relevant information is confined to $\widehat{U}_\ell$.

B.4 Format Invariance and Label-Corruption Stability

Task-format disentanglement in concept coordinates. To test whether the learned subspace captures task identity rather than superficial prompt format, we construct a factorial dataset with five relation tasks and three prompt formats: symbolic, verbose natural language, and question answering. For each relation R, format f, and query instance i, we extract an incremental task vector at the selected layer $v_{\ell^\star}^{(R,f,i)} = h_{\ell^\star,\mathrm{few}}^{(R,f,i)} - h_{\ell^\star,\mathrm{zero}}^{(f,i)}$, and decompose it using the learned orthonormal concept basis $\widehat{U}_{\ell^\star}$: $\beta^{(R,f,i)} = \widehat{U}_{\ell^\star}^\top v_{\ell^\star}^{(R,f,i)}$ and $v_\perp^{(R,f,i)} = (I - \widehat{U}_{\ell^\star}\widehat{U}_{\ell^\star}^\top) v_{\ell^\star}^{(R,f,i)}$. We compare the full residual vector v, the concept coordinates β, and the complementary component $v_\perp$. For each representation, we compute silhouette scores using either relation labels or prompt-format labels. A format-invariant task representation should increase task separability while suppressing format separability.

Table 2 and Figure 6 reveal a striking geometric disentanglement. Compared to the ambiguous full space, the low-dimensional concept subspace extracts a format-invariant task manifold: task representations form dense, well-separated clusters (0.411), while format noise is erased (-0.007), forcing identical tasks from disparate templates to perfectly overlap. Conversely, in the complementary space, task identity collapses (0.022), leaving the logical rule drowned out by dominant surface-format residues (0.173). This diagnostic supports interpreting $\widehat{U}_{\ell^\star}$ as a task-aligned concept subspace rather than a generic low-rank projection. It also explains why complementary-subspace patching is ineffective in our intervention experiments: the complement preserves format-related information but contains little of the relation-level signal needed to restore in-context prediction.

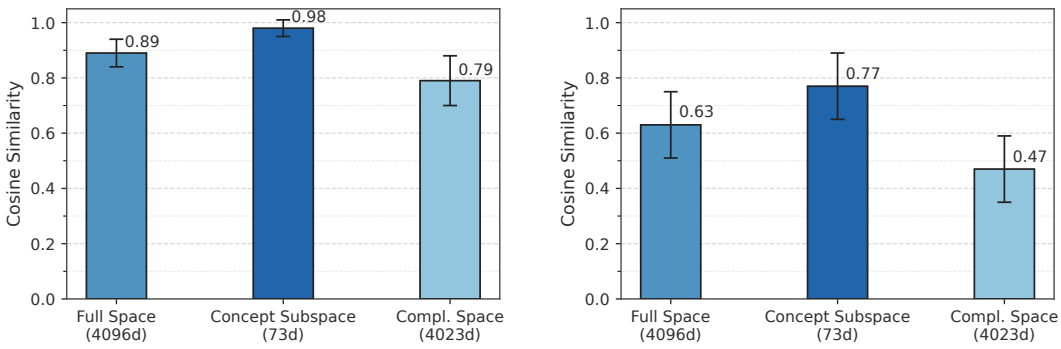

Figure 7: Format invariance and corruption stability of concept representations. *Left:* cross-format cosine similarity. *Right:* clean–corrupted representation similarity under label corruption.

Cross-format alignment of task representations. We test whether the learned subspace captures task semantics rather than the prompt template. Fixing the same underlying relation and few-shot exemplars, we rewrite each prompt in three formats (symbolic, QA, and natural language), extract the query-token residual at layer $\ell^\star$, and compare cross-format cosine similarity in the full residual stream, the learned concept subspace $\widehat{U}_{\ell^\star}$ (73d), and its orthogonal complement.

Figure 7 (left) shows a consistent ordering: similarity is highest in the concept subspace (0.98), lower in the full state (0.89), and lowest in the complementary space (0.79). This suggests that the full residual mixes shared task information with format-specific variation, while much of the surface-form noise lies outside $\widehat{U}_{\ell^\star}$. The near-perfect alignment inside the learned subspace indicates that it preserves the same relation-level computation across prompt realizations.

This also validates the subspace estimator: if $\widehat{U}_{\ell^\star}$ primarily captured lexical or stylistic regularities, projecting onto it would not improve cross-format agreement over the full state. Instead, projection removes format-dependent variation and aligns representations. We therefore interpret $\widehat{U}_{\ell^\star}$ as a format-stable carrier of task information under controlled prompt rewrites while avoiding a stronger claim of robustness under arbitrary distribution shifts.

Stability under label corruption. We test whether the learned concept subspace remains stable when demonstrations are severely corrupted. Starting from a clean few-shot prompt, we construct a noisy version by replacing demonstration labels with random unrelated tokens while keeping the query and prompt structure fixed. At the selected layer, we extract the query-token residual state from both runs and compare clean–noisy cosine similarity in three spaces: the full residual stream, the learned 73-dimensional concept subspace, and its orthogonal complement.

Figure 7 (right) shows a consistent ordering: similarity is highest in the concept subspace (0.77), lower in the full state (0.63), and lowest in the complementary space (0.47). This suggests that the full residual mixes task information with corruption-induced noise, while much of that noise is diverted into off-subspace directions. The higher similarity inside the learned subspace indicates that relation-level information is more stable than the ambient representation under label corruption.

This supports interpreting the learned subspace as a denoised carrier of task information rather than a prompt-specific artifact: projecting onto it increases agreement between clean and corrupted representations instead of reducing it. We view this as representation-level robustness under controlled corruption, not a general guarantee under arbitrary adversarial shifts. Still, alongside the patching and swapping results, this further strengthens the claim that the learned concept subspace captures task-level structure rather than prompt-local statistics.

B.5 Prompt Contamination and Concept-Vector Alignment

Prompt Contamination and Concept-Vector Alignment. We next test whether failure under prompt contamination can be explained geometrically inside the learned concept subspace. Let $P_{\widehat{U}_{\ell^\star}} = \widehat{U}_{\ell^\star}\widehat{U}_{\ell^\star}^\top$ denote the projector onto the learned concept subspace, and let $W_U[:, y^*]$ be the

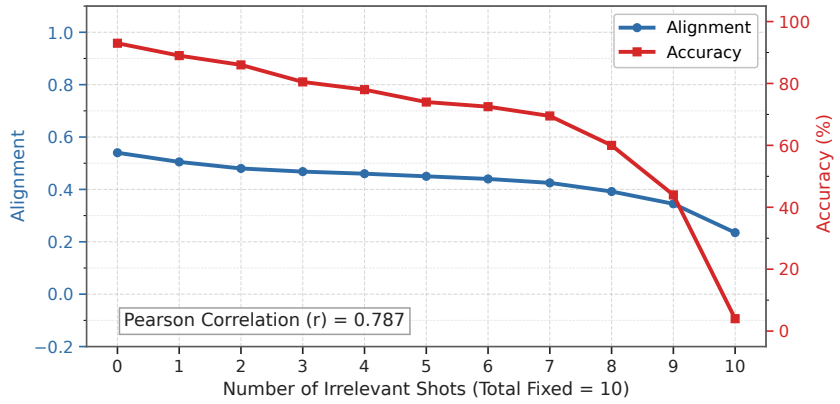

Figure 8: Prompt contamination rotates the inferred concept direction. Alignment in concept space and prediction accuracy jointly decline as irrelevant demonstrations replace relevant shots.

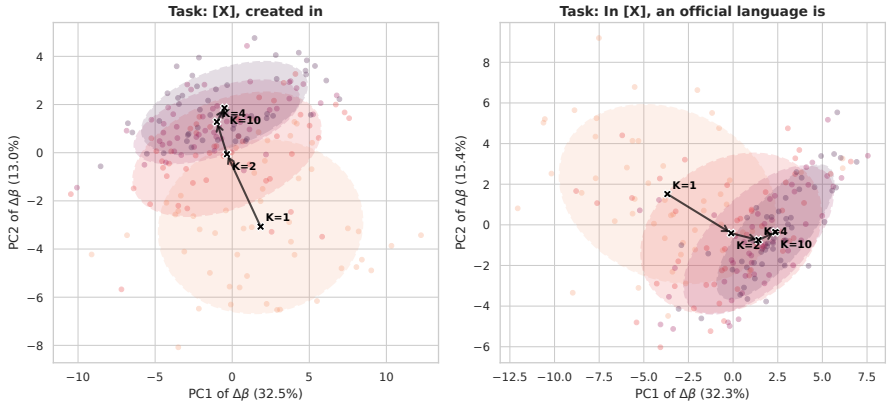

Figure 9: Few-shot contraction of debiased concept-coordinate estimates. Query-centered concept-coordinate displacements concentrate as the number of demonstrations increases.

unembedding vector of the correct answer token y^*. We define an answer-anchored reference vector $u_{\text{ideal}} = \frac{P_{\widehat{U}_{\ell^\star}} W_U[:,y^*]}{\|P_{\widehat{U}_{\ell^\star}} W_U[:,y^*]\|_2}$ and fix the total prompt budget at 10 shots while progressively replacing relevant demonstrations with cross-domain irrelevant examples. For each contamination level, we run autoregressive inference, extract the induced task direction u_{ctx} in the concept subspace, and measure its absolute alignment to the reference $a = \left|\cos(u_{\text{ctx}}, u_{\text{ideal}})\right|$, together with the final prediction accuracy.

Figure 8 shows that internal geometric drift and external performance degradation closely co-vary. As the number of irrelevant shots increases, alignment to the target answer direction decreases steadily, and accuracy declines in parallel, with a Pearson correlation of 0.787. The degradation is initially gradual but becomes much sharper in the high-contamination regime, where both curves bend downward. This pattern suggests that irrelevant demonstrations do not merely add generic noise; they systematically rotate the inferred task vector away from the answer-consistent region of the learned concept subspace. The result strengthens the mechanistic interpretation of $\widehat{U}_{\ell^\star}$: the same subspace that supports recovery and concept swapping also predicts robustness limits under prompt contamination. At the same time, the alignment–accuracy relation is strong but not perfect, so this quantity is best viewed as a compact diagnostic of task fidelity rather than a complete account of all failure modes.

B.6 Few-Shot Contraction of Debiased Concept Coordinates

To examine whether additional demonstrations reduce task-coordinate uncertainty, we isolate context-induced updates from query priors via debiased displacements. Specifically, at layer $\ell^\star$, the update for relation R, query q, and K-shot context c is defined as $\Delta\beta_K^{(R,q,c)} = \beta_K^{(R,q,c)} - \beta_0^{(R,q)}$, where $\beta_0^{(R,q)}$ is the zero-shot baseline. Evaluating eight semantic relations (50 parallel contexts each, $K \in \{1, 2, 4, 10\}$), we project $\Delta\beta_K$ onto its first two principal components, visualizing uncertainty dynamics through K-specific 1.5-standard-deviation covariance ellipses. Figure 9 shows two representative relations. At $K = 1$, the displacement clouds have large covariance ellipses, indicating that a single demonstration yields a noisy, exemplar-dependent estimate. As K increases, the clouds contract and their centroids stabilize, and the most visible contraction occurs from $K = 1$ to $K = 4$, with smaller changes from $K = 4$ to $K = 10$. The contraction pattern is relation-dependent: the "created in" relation exhibits a larger centroid shift, whereas the "official language" relation starts with a broader $K = 1$ cloud and rapidly concentrates near the higher-shot centroid. These trajectories support interpreting β-coordinates as few-shot estimates of task rules: after removing query-specific bias, the inferred coordinate becomes less sensitive to the sampled context as more demonstrations are provided.

C Scope Checks, Rank-Threshold Sensitivity and Probes

This section first checks the scope of the intervention pattern of Section 4 and then examines the estimated subspace. Appendices C.1 and C.2 run the patching and swap experiments on Qwen2.5-7B and on an implicit cross-lingual translation task with Llama-3-8B. Appendices C.3 and C.4 turn to context-defined nonce mappings and, on Qwen2.5-7B, two-step compositional rules, both of which lie outside Theorem 2. Appendix C.5 tests whether a subspace estimated from 24 CounterFact relations transfers to 10 unseen ones. On 990 Qwen2.5-7B layer-27 states, Appendices C.6 and C.7 vary the estimator threshold and compare linear and nonlinear probes on the concept subspace and its complement. Appendices C.5–C.7 compare their results with predictions of the single-head model. Recovery is defined as in Section 4 and, from Appendix C.3 on, computed from unrounded accuracies.

C.1 CounterFact Experiments on Qwen2.5-7B

We further test whether the observed concept-subspace effect is specific to Llama-3-8B Llama Team (2024) by repeating the corresponding experiments on Qwen2.5-7B Qwen Team (2024). Unless otherwise stated, we use the same CounterFact-derived multi-relation prompt construction, clean–corrupted intervention protocol, subspace extraction criterion, and evaluation metrics as in the main experiments. For Qwen2.5-7B, the intervention layer is the 27th transformer layer, and all subspace estimates and patching interventions in this subsection are performed at that layer.

Figure 10 first shows the layerwise localization and low-rank extraction results. The clean-to-corrupted full-state patching curve again exhibits a late-layer pattern: early layers provide little recovery, while recovery becomes substantial only in the later part of the network and approaches the clean reference level near the final layers. This indicates that, as in Llama-3-8B, task-level information is not uniformly available throughout the stack but becomes causally accessible after sufficient contextual processing. At the selected 27th layer, the explained-cross-variance spectrum is sharply concentrated. The same 98% threshold selects a 45-dimensional subspace out of the 3584-dimensional residual stream, suggesting that the task-aligned signal remains highly compressed in this model as well.

Figure 11 tests whether this low-dimensional subspace is causally relevant. In the patching experiment, the baseline accuracy is 32.5%, while the few-shot upper bound is 57.0%. Full-space patching reaches 57.0%, corresponding to 100.0% recovery of the clean–baseline gap. Patching only the 45-dimensional concept subspace reaches 53.0%, corresponding to 83.7% recovery, despite modifying only about 1.1% of the residual dimensions. In contrast, patching the 3539-dimensional complementary space reduces accuracy to 21.0%, yielding a negative recovery rate. Therefore, we do not interpret the complementary intervention as carrying useful task information in this setting. It provides a strong control showing that the recovery is not explained by the number of patched dimensions.

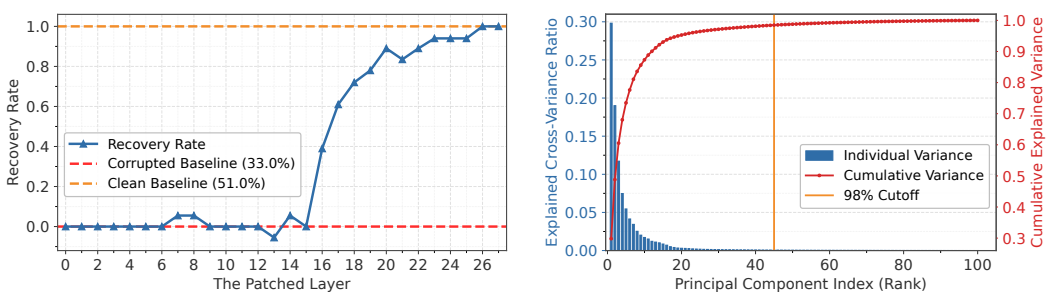

Figure 10: Layerwise localization and low-rank extraction of task-aligned residual directions. *Left*: clean-to-corrupted patching localizes layers where task information is causally accessible. *Right*: the 98% explained-cross-variance threshold selects a low-dimensional concept subspace at layer 27.

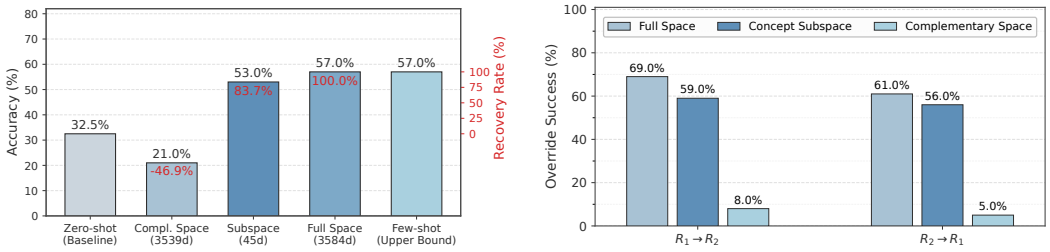

Figure 11: Concept-subspace patching and swapping. *Left:* the learned subspace recovers most of the clean-corrupted gap. *Right:* subspace swaps redirect predictions toward the injected relation.

The concept-swap results show the same qualitative specificity. Full-space replacement achieves 69.0% override success for $R_1 \to R_2$ and 61.0% for $R_2 \to R_1$. Replacing only the learned concept subspace remains close to this level, with 59.0% and 56.0% override success, respectively. By contrast, complementary-space replacement is largely ineffective, reaching only 8.0% and 5.0%. Thus, on Qwen2.5-7B, relation identity is again concentrated in the learned task-aligned subspace rather than in the high-dimensional orthogonal complement.

Overall, these results support the same restrained interpretation as the main experiments. The exact selected rank and intervention layer differ across models, and the experiment does not imply a universal concept dimension or a complete circuit-level explanation. Nevertheless, Qwen2.5-7B exhibits the same qualitative pattern: a small task-aligned residual subspace is sufficient to recover most of the clean–corrupted gap and to redirect predictions under concept swaps, whereas the complementary space has little corresponding effect.

C.2 Implicit Cross-Lingual Translation

We use a controlled cross-lingual task to test whether concept-subspace mediation extends beyond factual-relation ICL to a more abstract rule: selecting the output language. From MUSE English–French and English–Spanish dictionaries Lample et al. (2018), we construct word triples and retain only examples whose French and Spanish targets are single Llama-3 tokens, lexically non-overlapping, and not very short or ambiguous. Prompts contain no explicit language tags such as `Fr:` or `Es:`; all demonstrations use the same format, e.g., `water -> eau`, so the model must infer the target-language rule from context. Let R_1 denote French and R_2 Spanish. The "corrupted" run below is a counterfactual competing-rule run, not a malformed prompt. All subspace estimation and interventions are performed on Llama-3-8B at layer 30, at the final query token.

Figure 12 localizes the cross-lingual rule representation. Full-state patching has little effect in early layers and becomes effective only in later layers, matching the late emergence observed in the CounterFact experiments. At layer 30, the cross-variance spectrum is highly concentrated: the same 98% explained-cross-variance criterion selects a 92-dimensional subspace out of the 4096-dimensional

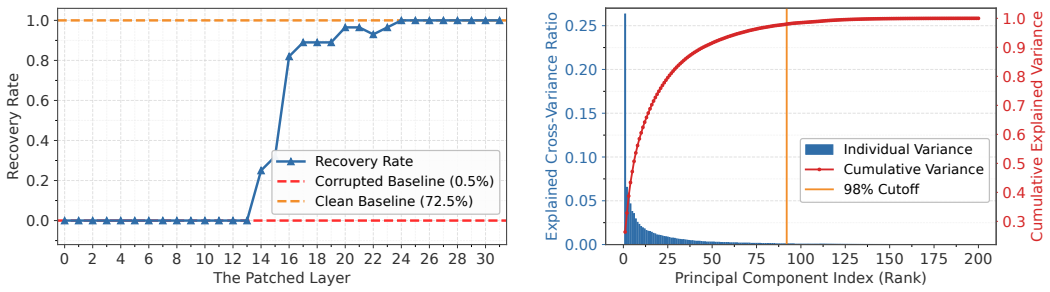

Figure 12: Layerwise localization and low-rank extraction of task-aligned residual directions. *Left*: clean-to-corrupted patching localizes layers where task information is causally accessible. *Right*: the 98% explained-cross-variance threshold selects a low-dimensional concept subspace at layer 30.

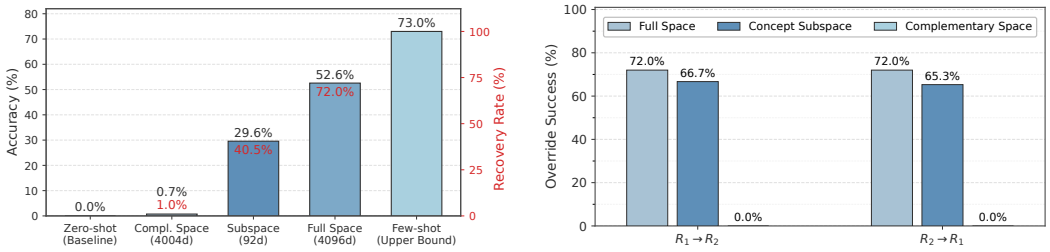

Figure 13: Concept-subspace patching and swapping. *Left:* the learned subspace recovers most of the clean-corrupted gap. *Right:* subspace swaps redirect predictions toward the injected relation.

residual stream. Thus, the target-language rule appears in a compact set of activation directions, although this does not imply that the full translation computation is low-dimensional.

Figure 13 tests the causal role of this subspace. When evaluated against the French target, the competing-rule baseline is essentially at floor, while native few-shot French and Spanish contexts reach 73.0% and 72.0% accuracy, respectively. Full-state patching raises accuracy to 52.6%, recovering 72.0% of the clean–competing gap. Patching only the 92-dimensional concept subspace reaches 29.6% accuracy, or 40.5% recovery, despite modifying only $92/4096$ residual dimensions. In contrast, patching the 4004-dimensional orthogonal complement gives only 0.7% accuracy and 1.0% recovery. The effect is therefore tied to task-aligned directions rather than intervention size.

The swap experiment gives a sharper test of directional control. Here $R_1 \to R_2$ denotes overriding a French-context run toward Spanish output, and $R_2 \to R_1$ the reverse. Full-state replacement achieves 72.0% override success in both directions. Replacing only the learned cross-lingual subspace remains close to this level, with 66.7% for $R_1 \to R_2$ and 65.3% for $R_2 \to R_1$, whereas complementary-space replacement yields 0.0% in both directions. Thus, the learned subspace carries most of the target-language identity needed to redirect generation. The remaining gap in patching accuracy suggests that lexical, calibration, or decoding-related factors outside this single-layer linear subspace also contribute. Overall, the cross-lingual results extend the main finding beyond CounterFact-style relations: a low-dimensional, task-aligned residual subspace acts as a major causal mediator of language-rule selection, but not as a complete circuit-level explanation of cross-lingual generation.

C.3 Context-Defined Nonce Mappings

CounterFact relations are tied to pretrained factual associations. This task removes that dependence and checks the scope of the intervention pattern of Section 4. It does not test Theorem 2, which concerns an idealized single-layer head on the task family of Assumption 1, because a random symbol mapping is not an instance of that family. Each episode draws a fresh random mapping between nonce symbols (e.g., `dax` $\to$ `vop`, `mip` $\to$ `lurn`) and defines it only through the demonstrations.

Table 3: Context-defined nonce mappings, redrawn in every episode. Target and distractor are the fractions of swapped predictions that follow the injected mapping and the distractor mapping.

Intervention	Accuracy (%)	Recovery (%)	Target (%)	Distractor (%)
Clean	72.0	–	–	–
Corrupted	22.0	–	–	–
Zero-shot	19.5	–	–	–
Full state	69.8	95.6	64.5	16.0
Concept subspace	60.9	77.8	55.8	18.5
Complement	31.4	18.8	24.2	22.0
Matched-rank random	25.1	6.2	20.6	21.5

The symbols have no fixed meaning and the mapping is redrawn in every episode, so neither pretraining nor other episodes determine the answer. Zero-shot accuracy, with no demonstrations, measures how often the model predicts the target from the query alone. As in Sections 3.4 and 4, the corrupted context permutes the demonstration labels and keeps the symbols and the prompt format. Here the permuted pairs form another valid mapping, so the corrupted context is as well formed as the clean one. We report top-1 accuracy and the recovery rate of Section 4 for full-state, concept-subspace, complement and matched-rank random patching. Swaps inject the corresponding component from an episode with a different mapping. We report override success, the fraction of predictions that follow the injected mapping, and the fraction that follow a distractor mapping.

On this task, the intervention pattern of Section 4 appears although the answer is defined only by the context: the concept subspace carries most, but not all, of the patchable effect. The answer comes from the demonstrations: accuracy is 72.0% in the clean context, against 19.5% zero-shot and 22.0% in the corrupted context (Table 3). Full-state patching recovers 95.6% of the 50.0-point clean–corrupted gap. The concept subspace recovers 77.8%, about four-fifths of the full-state value, compared with 18.8% for its complement and 6.2% for a matched-rank random subspace. Under the concept-subspace swap, 55.8% of predictions follow the injected mapping, against 64.5% under the full-state swap, and 18.5% follow the distractor, whereas complement and random swaps show no comparable preference.

The result does not show that the model estimates a concept parameter, as the head of Theorem 2 does: copying the output of the matching demonstration, as induction heads do (Olsson et al., 2022), solves the task without one. Two results have the form of predictions of Section 3.4 but do not test them. First, beating the matched-rank control does not establish selectivity, prediction (a). In the setting of Theorem 2(v), a head that is optimal for an unstructured prior has $\mathrm{Rec}(V) = E(V)$, so any subspace with large difference energy $E(V)$ outperforms random directions without structured reading. Prediction (a) requires recovery close to one and well above $E(\widehat{U}_{\ell^\star})$; we do not report $E(\widehat{U}_{\ell^\star})$, and recovery here is partial. Second, if the clean and corrupted contexts, or a swap donor and its recipient, share the query, a subspace that holds the answer assigned to that query would produce the same patching and swap results as a mapping-level task state. The swap therefore bears on query independence, prediction (b), only if donors use different queries. Concept and complement recoveries sum to 96.6%, close to the full-state 95.6%. This near-additivity is only an approximate linearity check, since additivity holds for any linear read-out (Theorem 2(ii)) and top-1 recovery is not linear in the patched state. We do not infer from one task that other context-defined computations share this pattern.

C.4 Two-Step Compositional Rules

Each context demonstrates two rules, $x \to f(x)$ and $f(x) \to g(f(x))$, and the query asks for the composed output $g(f(x))$. The corrupted context shuffles the demonstrations of one step while preserving the prompt format. On Qwen2.5-7B, we estimate and patch a layer-27 concept subspace (the layer of Appendix C.1) and a two-layer union that adds a layer-23 concept subspace in the same forward pass. Swaps inject the corresponding component from a context with a different composed rule. As illustrated in Figure 14, the layer-27 concept subspace recovers 63.6% of the clean–corrupted gap, about two-thirds of the full-state 94.3%, compared with 24.2% for its complement and 10.1% for a matched-rank random subspace. Its swap makes 50.5% of predictions follow the injected composed rule and 18.0% the distractor, whereas complement and random swaps show no preference.

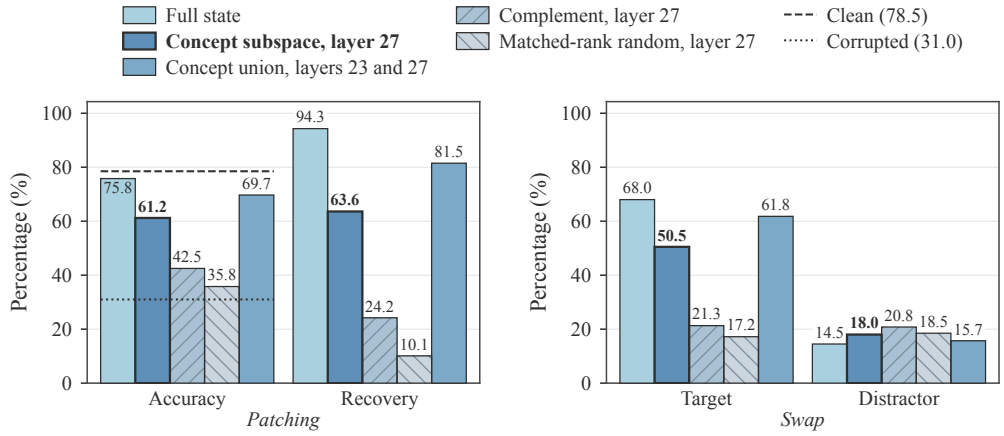

Figure 14: Two-step compositional rules, Qwen2.5-7B. Target and distractor are the fractions of swapped predictions that follow the injected composed rule and the distractor rule.

Table 4: Subspace transfer to unseen relations. Within each evaluation set, the two subspaces share the clean and corrupted accuracies.

Evaluated on	Estimated from	Clean (%)	Corrupted (%)	Concept rec. (%)	Complement rec. (%)
Seen (24)	Seen (24)	60.9	30.4	64.3	24.6
Seen (24)	All (34)	60.9	30.4	80.4	23.2
Unseen (10)	Seen (24)	73.8	22.5	48.8	18.0
Unseen (10)	All (34)	73.8	22.5	56.1	24.4

The patchable effect of the shuffled step on the composed answer thus remains concentrated in task-aligned directions. The layer-27 subspace is, however, an incomplete mediator. Adding the layer-23 subspace raises recovery to 81.5% and the target rate to 61.8%, about 91% of the full-state 68.0%. This is consistent with mediation by subspaces at more than one layer or in more than one component. We cannot yet separate it from a richer layer-27 subspace. If both patches act at the final position, the layer-23 patch affects the prediction only through the layer-27 state. We also report no standalone layer-23 intervention or controls matched to the union. Both the compositional task and the union lie outside Theorem 2. We do not infer from one task and one model that compositional computation is generally low-dimensional.

C.5 Subspace Transfer to Unseen Relations

The main protocol is example-disjoint but not relation-disjoint: the concept subspace is estimated from 12-shot episodes of all 34 relations and then frozen, and evaluation uses 90 newly sampled episodes per relation whose query subjects are disjoint from those used for estimation. To test transfer across relations, we fix a split of the 34 relations into 24 seen and 10 unseen relations (about $70\%/30\%$). We estimate one subspace from examples of the seen relations only and, as a reference, one from all 34 relations. Both are frozen and evaluated by concept and complement patching on examples of each set that do not overlap the estimation examples. In Table 4, recovery is computed against the clean–corrupted gap of each set. These gaps differ (30.5 and 51.3 points), so we compare the two subspaces within a set rather than recoveries across sets.

The seen-relation subspace transfers partially to the unseen relations. Estimated without any of their examples, it recovers 48.8% of their gap, compared with 18.0% for its complement and 56.1% for the reference. In the single-head model, the task centroids of Theorem 2(iii) computed from the estimation relations span $W\Lambda U \operatorname{span}\{\beta_\tau\}$. This is a subspace of the written subspace $\mathcal{W}$, which every task of the family shares, and equals $\mathcal{W}$ once the β_τ span $\mathbb{R}^r$. Proposition 1 is the input-space analogue. The model therefore predicts full transfer if the seen relations' β_τ span $\mathbb{R}^r$, and selective but incomplete recovery otherwise.

Table 5: Rank-threshold sensitivity.

Threshold	Rank	Conc. rec. (%)	Compl. rec. (%)	Conc. swap (%)	Compl. swap (%)
90%	15	42.5	38.2	32.0	11.0
95%	26	63.1	31.4	38.0	10.0
98%	62	74.5	29.4	43.0	8.0
99%	101	75.4	29.2	44.0	7.0
99.5%	143	76.1	27.5	44.0	7.0

Table 6: Relation decoding accuracy (%) from residual-state features.

Feature space	Dim.	Linear	MLP
Learned concept coordinates	62	88.3	83.1
Random rank-matched projection	62	38.4	52.3
Full-residual PCA	256	94.4	91.1
Leading complement PCA components	256	35.9	53.6

These results give no evidence that the shortfall relative to the reference is specific to unseen relations. Within each set, the reference exceeds the seen-relation subspace by 16.1 points on the seen relations and by 7.3 points on the unseen ones (about 4.9 and 3.7 points of accuracy). If the seen relations missed directions that the unseen ones use ($\operatorname{rank} B < r$ in Proposition 1), adding the unseen relations to the estimation would be expected to help them more than the seen relations. In the single-head model, such a gap also lowers recovery on the seen relations, through the permutation term $c_{\triangle}\Lambda$ in Equation (75), so only the ordering of the gains is diagnostic. The gain is instead larger on the seen relations, although without intervals this difference is not conclusive. Fewer estimation examples or a different selected rank could also produce the shortfall, and this experiment cannot separate these causes. Theorem 2 concerns the pooled, logit-level Rec of a single-layer head. The recoveries here are top-1, and concept and complement recoveries sum to 66.8%–103.6% rather than to one as in Theorem 2(ii), so we use the complement only as a control. They are reported without intervals and without full-state or matched-rank random controls. We do not infer a relation-independent subspace, nor that the shortfall stems from relation-specific directions.

C.6 Rank-Threshold Sensitivity

On a set of 990 Qwen2.5-7B layer-27 residual states, we vary the explained-cross-variance threshold while keeping the clean and corrupted baselines fixed (Table 5). Below 95%, the selected subspace omits task-relevant directions and complement recovery is high. From 98% on, recovery and swap success saturate even though the rank more than doubles. This is the shape predicted by Theorem 2(ii)–(iii): recovery depends on how much of the read subspace is captured, not on the rank, and the selected rank is an operational quantity. We therefore use 98% as a fixed operating point, not as an estimate of r. Corollary 4(b) quantifies the plateau. Suppose the directions added beyond rank $k_0 = 62$ behave like random directions in the complement of the 62-dimensional subspace. Then they raise recovery by $k_1/(D - k_0)$ of what that subspace leaves to its complement. With $D = 3584$, this predicts 74.8% and 75.2% at ranks 101 and 143 (observed 75.4% and 76.1%), and complement recoveries of 29.1% and 28.7% (observed 29.2% and 27.5%). The complement plateau near 28% is thus the part of the written subspace that the 98% subspace misses.

C.7 Linear and Nonlinear Probes

We train linear and one-hidden-layer MLP probes to predict the relation from the 990 Qwen2.5-7B layer-27 residual states of Appendix C.6, using a fixed 742/248 train/test split. As shown in Table 6, the MLP decodes part of the relation from the complement (53.6%), but 62 learned coordinates remain more informative (83.1%) than 256 complement components, and complement patching recovers much less. Theorem 2(iv) gives a mechanism. Under (BD), the unread component carries task information only through the scalar $\beta^\top \Lambda_{11} \beta$, which a probe can read but the read-out does not use. This scalar shifts the mean of ν along v_y, so even a linear classifier can separate relations that differ in it, and above-chance linear accuracy on the complement (35.9%) is compatible with (BD). Two further sources lie outside the theorem. One is cross-subspace coupling, as in (NBD). The other is that query entities depend on the relation, which makes $h_0(x_q)$ itself informative.