How Does Scaling Reshape Skill Structure in Language Model Representations?

Anonymous authors
Paper under double-blind review

Abstract

Scaling language models consistently improves downstream performance, yet the representation-level changes associated with this improvement remain poorly understood. We study this question through skill structure, using skills as an abstraction of latent capabilities reflected in representations and relevant to downstream prediction. To make skill structure comparable across scales, we develop a task- and template-conditioned analysis based on RSS, with common dimensions and matched random-label calibration. Our theory shows that an increase in the positive calibrated between-skill component raises the guaranteed floor on fitted prototype separation, which in turn can tighten the upper bound on held-out skill-recovery error when prototype-estimation error and within-skill variation are controlled. Across Pythia and Qwen2.5, larger models generally exhibit stronger skill structure. Geometric decomposition shows that this trend is driven primarily by increased separation among skill-conditioned representation centers, accompanied by larger held-out prototype margins. We further test whether this geometry contributes functionally to answer formation. Interventions on skill-aligned representation subspaces produce larger behavioral effects than relative-energy-matched random interventions, supporting functional relevance. The same representation diagnostics provide useful signals for selecting adaptation data under limited training budgets. Together, these results show that scaling is accompanied by stronger skill structure, greater between-skill separation, and more reliable held-out skill recoverability, while the associated skill-aligned subspaces are functionally relevant to answer formation.

1 Introduction

Scaling language models is consistently associated with improved performance across a broad range of downstream tasks (Brown et al., 2020; Hoffmann et al., 2022; Wei, 2022). Recent work has begun to show that scaling also reshapes representation geometry (Li et al., 2025b; Liu et al., 2025; Redhardt et al., 2026). However, geometric change alone does not explain which representational changes are relevant to the behavioral improvements associated with scaling. This leaves open what task-relevant structure emerges in representation space as models scale.

A substantial body of interpretability work studies neural representations and model components to understand how language models encode and use information (Dalvi et al., 2018; Voita et al., 2019; Huben et al., 2024; Hong et al., 2026). These studies reveal rich representational and computational structure, but they do not directly characterize downstream-relevant capabilities at a level that can be compared systematically across model scales. We therefore use *skills* to abstract task-relevant latent capabilities reflected in representations, with the resulting labels defining a fixed, model-independent partition rather than a presumed intrinsic skill decomposition. This provides a common unit for studying how *skill structure* changes with model scale. These considerations motivate three questions. First, how does *skill structure* change with model scale? Second, to what extent is skill structure estimated from fitting examples recoverable on unseen examples? Third, do skill-aligned representation subspaces contribute functionally to answer formation?

We quantify *skill structure* using task- and template-conditioned RSS, $RSS(S \mid T, D)$, which measures the representation structure associated with skill labels after conditioning on task T and prompt template D. We calibrate this measure against matched random-label partitions and decompose it

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

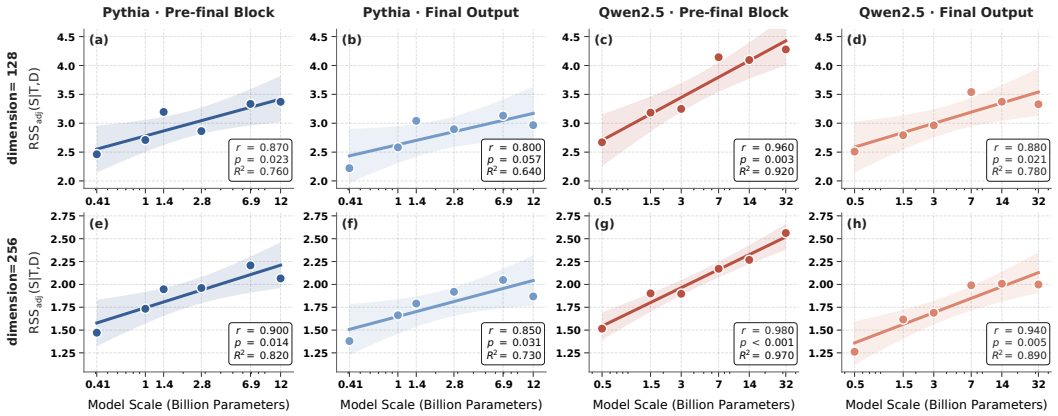

Figure 1: **Skill structure strengthens with model scale.** Adjusted task- and template-conditioned skill structure, $\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D)$, is shown across model scales for Pythia (a,b,e,f) and Qwen2.5 (c,d,g,h). Columns correspond to the pre-final-block and final-output representation locations, while the top and bottom rows correspond to common representation dimensions of $q = 128$ and $q = 256$, respectively. Points denote individual model scales, and solid lines show fitted scale trends. Pearson correlation coefficients (r), corresponding p-values, and R^2 are reported in each panel.

into geometric components. Our theory links separation among skill-conditioned representation centers to fitted prototype separation, and further relates prototype separation to skill recovery on unseen examples. Specifically, it derives an upper bound on held-out skill-identification error that tightens as fitted prototype separation increases relative to prototype-estimation error and within-skill variation.

We examine these relationships across two independently trained model families, Pythia (Biderman et al., 2023) and Qwen2.5 (Yang et al., 2024), spanning a broad range of parameter scales. We use a common skill vocabulary to annotate a fixed set of downstream examples and reuse the same skill annotations across all models. Skill structure is evaluated after conditioning on task and prompt template, and representations from all model scales are mapped to common dimensions before comparison.

Across both model families, larger models generally exhibit stronger skill structure under these controls. Geometric decomposition shows that this scale-dependent trend is driven primarily by increased separation among skill-conditioned representation centers. On held-out examples, prototype margins also generally increase with model scale, indicating that skill structure estimated from fitting examples becomes more reliably recoverable on unseen examples.

We further test whether skill-aligned representation subspaces contribute functionally to answer formation. Interventions on prototype-derived subspaces produce larger behavioral effects than relative-energy-matched random interventions, providing evidence that the identified geometry is functionally relevant to prediction. As a proof of concept, we also use these representation diagnostics to guide data-efficient adaptation under limited training budgets. On Pythia-2.8B and Qwen2.5-3B, the resulting selection strategy improves over a matched variant that uses the same skill-coverage and K-center selection procedure but does not use the representation-based skill-weakness scores. Together, these results show that skill structure estimated from fitting examples remains recoverable on unseen examples, contains components that contribute functionally to answer formation, and can provide actionable signals for adaptation-data selection. Our main contributions are threefold:

- We develop a controlled analysis of *skill structure* across model scales and theoretically connect between-skill separation to held-out skill recoverability. Across Pythia and Qwen2.5, skill structure generally strengthens with model scale.
- We establish the functional relevance of the identified representation geometry through relative-energy-matched interventions, showing that perturbing prototype-derived subspaces affects answer formation more than matched random perturbations.
- We show that these representation diagnostics can guide data-efficient adaptation under limited training budgets, providing useful selection signals beyond coverage and representation diversity alone.

2 Related Work

Scaling and representation structure. Scaling laws characterize how model size, training compute, and data relate to language-modeling performance (Kaplan et al., 2020; Hoffmann et al., 2022; Hernandez et al., 2021). Controlled model suites such as Pythia further enable systematic analysis of how model properties change with scale (Biderman et al., 2023). More recent work has begun to examine how scaling relates to representation geometry and structure (Li et al., 2025b; Liu et al., 2025; Redhardt et al., 2026). These studies suggest that internal representations change systematically with scale, but they do not directly address how downstream-relevant skill structure changes across model scales or whether the resulting structure remains recoverable on unseen examples.

Skill-level analysis of model capabilities. Recent work has used semantically defined skills to obtain more fine-grained views of model capabilities than aggregate benchmark scores (Moayeri et al., 2025; Ye et al., 2023; Yu et al., 2023). For example, prior work infers the skills required by individual evaluation examples and uses skill slices to characterize behavioral differences across foundation models. Such work primarily uses skills for fine-grained capability evaluation and behavioral analysis. In contrast, we study how a fixed skill partition is reflected in representation space and how the resulting skill structure changes across model scales.

Representation geometry, probing, and causal intervention. A growing body of work studies semantic and functional information in neural representation geometry, including causal function vectors (Hewitt & Manning, 2019; Park et al., 2024; Todd et al., 2024). Probing tests whether such information is recoverable, but expressive probes may exploit weak representation structure (Hewitt & Liang, 2019; Belinkov, 2022). Moreover, decodability does not imply functional use (Elazar et al., 2021; Belrose et al., 2023). Recent work therefore intervenes on representation directions, subspaces, or sparse features to test behavioral effects (Arditi et al., 2024; Rimsky et al., 2024; Marks et al., 2024). Accordingly, we separately examine held-out recoverability of skill structure and the functional contribution of skill-aligned subspaces to answer formation.

3 Problem Formulation

Let θ denote the fixed model parameters and $x \in \mathcal{X}$ an input example. For a Transformer layer ℓ, let $z^{(\ell)} = f_\theta^{(\ell)}(x) \in \mathbb{R}^{d_\ell}$ denote the representation at the final token of the prompted input under a given prompt template. For a fixed representation layer, we omit the layer index and write $z = f_\theta(x) \in \mathbb{R}^d$. For each example, let $T \in \mathcal{T}$, $D \in \mathcal{D}$, and $S \in \mathcal{S}$ denote its task, prompt template, and skill label, respectively. We use *skills* as an abstraction of latent capabilities reflected in representations and relevant to downstream tasks.

Because task and prompt template may themselves induce systematic differences in representation space, we characterize *skill structure* after conditioning on T and D. For a representation set $Z \subset \mathbb{R}^d$, let $\mathcal{R}(Z) = \frac{1}{2}\log\det\left(I + \frac{d}{\delta^2}\widehat{\Sigma}\right)$ denote its regularized coding rate (Yu et al., 2020), where $\widehat{\Sigma}$ is the empirical covariance and $\delta > 0$ is a fixed distortion parameter. We define the task- and template-conditioned RSS as $\mathrm{RSS}(S \mid T, D) = \sum_{t,d} p(t,d)\left[\mathcal{R}(Z_{t,d}) - \sum_s p(s \mid t,d)\mathcal{R}(Z_{t,d,s})\right]$, where $Z_{t,d}$ denotes representations under task t and template d, and $Z_{t,d,s}$ the corresponding subset with skill label s. Larger values indicate stronger skill structure within fixed task-template conditions.

For comparison across model scales, representations are mapped to a common dimension before computing RSS, controlling for differences in native representation dimensionality. Because arbitrary label partitions can also produce nontrivial RSS, we construct a matched random-label baseline by shuffling skill labels within each (T, D) cell while preserving empirical skill frequencies. The adjusted score is $\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D) = \mathrm{RSS}(S \mid T, D) - \mathbb{E}_\pi[\mathrm{RSS}_\pi(S \mid T, D)]$. A positive adjusted value therefore indicates skill structure beyond the matched random-label structure expected under the same task, prompt-template, and skill-frequency constraints.

Our second object of interest is *held-out skill recoverability*. Given disjoint fitting and evaluation sets, skill prototypes are estimated from $\mathcal{D}_{\mathrm{fit}}$ only and kept fixed when evaluating unseen examples from $\mathcal{D}_{\mathrm{eval}}$. For a task-template condition $c = (t, d)$, a held-out representation is assigned a skill by $\widehat{S}(z; c) = \arg\min_s \|z - \widehat{\mu}_{c,s}\|_2$. For a held-out example with true skill s, its prototype margin is

$M_c(z) = \min_{s' \neq s} \|z - \widehat{\mu}_{c,s'}\|_2 - \|z - \widehat{\mu}_{c,s}\|_2$. A larger positive margin indicates stronger separation from competing skill prototypes and more reliable held-out recovery.

4 Theory: Skill Structure and Held-Out Recoverability

In this section, we develop a representation-level theoretical framework to study what forms of skill structure enable more reliable skill identification on unseen examples. We show that greater separation among skill-conditioned centers can improve held-out skill identification.

4.1 Calibrated Geometric Sources of Skill Structure

To characterize the geometric sources of $\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D)$, consider a fixed task–template condition $c = (t, d)$. For skill s, let $\mu_{c,s}$ and $\Sigma_{c,s}$ denote the empirical mean and covariance of $Z_{c,s}$, and let $p_{c,s} = p(s \mid c)$. Define the condition-level mean $\mu_c = \sum_s p_{c,s}\mu_{c,s}$, the pooled within-skill covariance $\Sigma_{W,c} = \sum_s p_{c,s}\Sigma_{c,s}$, and the between-skill covariance $B_c = \sum_s p_{c,s}(\mu_{c,s} - \mu_c)(\mu_{c,s} - \mu_c)^\top$. By the law of total covariance, $\Sigma_c = \Sigma_{W,c} + B_c$, where Σ_c is the empirical covariance of Z_c.

Let $\lambda_q = \delta^2/q$. We define two geometric components. The between-skill separation component is $\Phi_{B,c} = \frac{1}{2}\log\det\big(I + (\Sigma_{W,c} + \lambda_q I)^{-1/2} B_c (\Sigma_{W,c} + \lambda_q I)^{-1/2}\big)$. The covariance component is $\Phi_{\mathrm{cov},c} = \frac{1}{2}\left[\log\det(\Sigma_{W,c} + \lambda_q I) - \sum_s p(s \mid c)\log\det(\Sigma_{c,s} + \lambda_q I)\right]$. The former measures separation among skill-conditioned representation centers relative to pooled within-skill variation, whereas the latter captures differences among skill-conditioned covariance structures.

Lemma 1 (Raw geometric decomposition)**.** *For every task–template condition c,*

$$\mathrm{RSS}_c(S) = \Phi_{B,c} + \Phi_{\mathrm{cov},c}. \tag{1}$$

For a fixed skill partition, both raw components are non-negative. However, finite-sample partitions can produce nonzero values even when the labels are unrelated to the representation geometry. We therefore calibrate both components using the same matched permutation baseline as $\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D)$. Let π denote a random permutation of skill labels within condition c, with Z_c fixed, and let $\Phi_{B,c}^{(\pi)}$ and $\Phi_{\mathrm{cov},c}^{(\pi)}$ denote the corresponding quantities. Define $\Phi_{B,c}^{\mathrm{adj}} = \Phi_{B,c} - \mathbb{E}_\pi[\Phi_{B,c}^{(\pi)}]$ and $\Phi_{\mathrm{cov},c}^{\mathrm{adj}} = \Phi_{\mathrm{cov},c} - \mathbb{E}_\pi[\Phi_{\mathrm{cov},c}^{(\pi)}]$.

Theorem 1 (Permutation-adjusted geometric decomposition)**.** *For every task–template condition c,*

$$\mathrm{RSS}_{\mathrm{adj},c}(S) = \Phi_{B,c}^{\mathrm{adj}} + \Phi_{\mathrm{cov},c}^{\mathrm{adj}}, \qquad \mathrm{RSS}_{\mathrm{adj}}(S \mid T, D) = \Phi_B^{\mathrm{adj}} + \Phi_{\mathrm{cov}}^{\mathrm{adj}}, \tag{2}$$

where $\mathrm{RSS}_{\mathrm{adj},c}(S) = \mathrm{RSS}_c(S) - \mathbb{E}_\pi[\mathrm{RSS}_c^{(\pi)}(S)]$. The condition-averaged components are $\Phi_B^{\mathrm{adj}} = \sum_c p(c)\Phi_{B,c}^{\mathrm{adj}}$ and $\Phi_{\mathrm{cov}}^{\mathrm{adj}} = \sum_c p(c)\Phi_{\mathrm{cov},c}^{\mathrm{adj}}$.

Theorem 1 separates calibrated skill structure into between-skill separation and skill-conditioned covariance differences. Unlike the raw components, the adjusted components are signed because each is measured relative to its matched permutation baseline. Our subsequent analysis focuses on Φ_B, since prototype-based skill recovery depends directly on the separation among skill-conditioned centers. This provides the geometric link from calibrated skill structure to the prototype geometry analyzed next.

4.2 From Between-Skill Separation to Prototype Geometry

The decomposition above characterizes skill structure at the level of representation distributions. We next relate between-skill separation to the prototype geometry used for held-out skill recovery. In this subsection, all empirical quantities defined in Section 4.1 are evaluated on $\mathcal{D}_{\mathrm{fit}}$.

For a fixed task–template condition c, let $\widehat{\mu}_{c,s}$ denote the fitted prototype of skill s and $\widehat{\mu}_c = \sum_s p(s \mid c)\widehat{\mu}_{c,s}$. Define $A_c = \Sigma_{W,c} + \lambda_q I$ and the normalized prototype displacement $u_{c,s} = A_c^{-1/2}(\widehat{\mu}_{c,s} - \widehat{\mu}_c)$. Because B_c and $\Sigma_{W,c}$ are evaluated on the same fitting examples, the normalized between-skill scatter is $G_c = A_c^{-1/2} B_c A_c^{-1/2} = \sum_s p(s \mid c) u_{c,s} u_{c,s}^\top$, and $\Phi_{B,c} = \frac{1}{2}\log\det(I + G_c)$.

Let $h_c = 1 - \sum_s p(s \mid c)^2$, and define the weighted mean squared distance between distinct normalized prototypes as $\bar{d}_c^2 = h_c^{-1} \sum_{s \neq s'} p(s \mid c) p(s' \mid c) \|u_{c,s} - u_{c,s'}\|_2^2$.

Proposition 1 (Between-skill separation and prototype separation). *For $h_c > 0$ and $r_c = \mathrm{rank}(G_c) \geq 1$,*

$$\bar{d}_c^2 \geq \frac{2r_c}{h_c} \left[\exp\left(\frac{2\Phi_{B,c}}{r_c} \right) - 1 \right]. \tag{3}$$

Proposition 1 shows that stronger between-skill separation, as measured by $\Phi_{B,c}$, raises a lower bound on average normalized prototype separation. Since $\Phi_{B,c}^{(\pi)} \geq 0$ for every matched permutation, $\Phi_{B,c} \geq [\Phi_{B,c}^{\mathrm{adj}}]_+$. Thus, positive adjusted between-skill separation also yields a conservative lower bound on average prototype separation.

Average separation alone does not guarantee that every pair of skills remains well separated. We therefore define $d_{\min,c}^2 = \min_{s \neq s'} \|u_{c,s} - u_{c,s'}\|_2^2$ and $\beta_c = d_{\min,c}^2 / \bar{d}_c^2$ if $\bar{d}_c^2 > 0$, and $\beta_c = 0$ otherwise. Here, $\beta_c \in [0, 1]$ measures how well the minimum pairwise separation is preserved relative to the average. We further define the minimum fitted prototype separation as $\widehat{\Delta}_c = \min_{s \neq s'} \|\widehat{\mu}_{c,s} - \widehat{\mu}_{c,s'}\|_2$.

For cross-scale comparison, let m index model scale and let K_c denote the number of active skills under condition c. Because the labeled population is fixed across scales, $p(s \mid c)$ and hence h_c are scale independent. Moreover, $A_{c,m} \succeq \lambda_q I$, and the centered skill prototypes span at most $K_c - 1$ dimensions, so $r_{c,m} \leq R_c := K_c - 1$.

Assumption 1 (Uniform pairwise non-collapse). *For each condition c, there exists a constant $\underline{\beta}_c > 0$, independent of model scale, such that $\beta_{c,m} \geq \underline{\beta}_c$ for all evaluated scales m.*

Corollary 1 (Scale-comparable prototype-separation bound). *Under Assumption 1,*

$$\widehat{\Delta}_{c,m}^2 \geq \kappa_c \left[\exp\left(\frac{2[\Phi_{B,c,m}^{\mathrm{adj}}]_+}{R_c} \right) - 1 \right], \qquad \kappa_c = \frac{2\lambda_q \underline{\beta}_c R_c}{h_c}. \tag{4}$$

All quantities in κ_c are independent of model scale. Therefore, under the uniform pairwise non-collapse assumption, $[\Phi_{B,c,m}^{\mathrm{adj}}]_+$ is the scale-varying term in the guaranteed lower bound on fitted prototype separation. An increase in $[\Phi_{B,c,m}^{\mathrm{adj}}]_+$ consequently raises this guaranteed floor, without requiring the realized minimum separation to increase monotonically at every scale.

4.3 From Prototype Separation to Held-Out Skill Recoverability

We next characterize when prototype separation estimated from $\mathcal{D}_{\mathrm{fit}}$ supports reliable skill recovery on unseen examples. For a held-out example with skill s under condition c, write $z = \mu_{c,s}^\star + \xi$, where $\mu_{c,s}^\star$ is the population skill center and ξ denotes within-skill variation.

Assumption 2 (Held-out regularity). *For each condition c, the fitted prototypes satisfy $\max_s \|\widehat{\mu}_{c,s} - \mu_{c,s}^\star\|_2 \leq \eta_c$. Moreover, for all $r \geq 0$, $\Pr(\|\xi\|_2 \geq r \mid c, s) \leq C \exp(-r^2/(2\sigma_c^2))$.*

The first condition bounds prototype-estimation error, while the second controls within-skill variation on unseen examples. Recall that $\widehat{\Delta}_c$ denotes the minimum fitted prototype separation defined in Section 4.2.

Lemma 2 (Held-out prototype-margin lower bound). *Under Assumption 2, the prototype margin $M_c(z)$ defined in Section 3 satisfies $M_c(z) \geq \widehat{\Delta}_c - 2\|\xi\|_2 - 2\eta_c$.*

Lemma 2 shows that fitted prototype separation yields a positive held-out margin when it dominates both within-skill variation and prototype-estimation error.

Theorem 2 (Held-out skill recoverability). *Under Assumption 2, if $\widehat{\Delta}_c > 2\eta_c$, then the nearest-prototype rule $\widehat{S}(z; c)$ defined in Section 3 satisfies*

$$\Pr\Big(\widehat{S}(z; c) \neq s \mid c, s \Big) \leq C \exp\Big[-(\widehat{\Delta}_c - 2\eta_c)^2 / (8\sigma_c^2) \Big]. \tag{5}$$

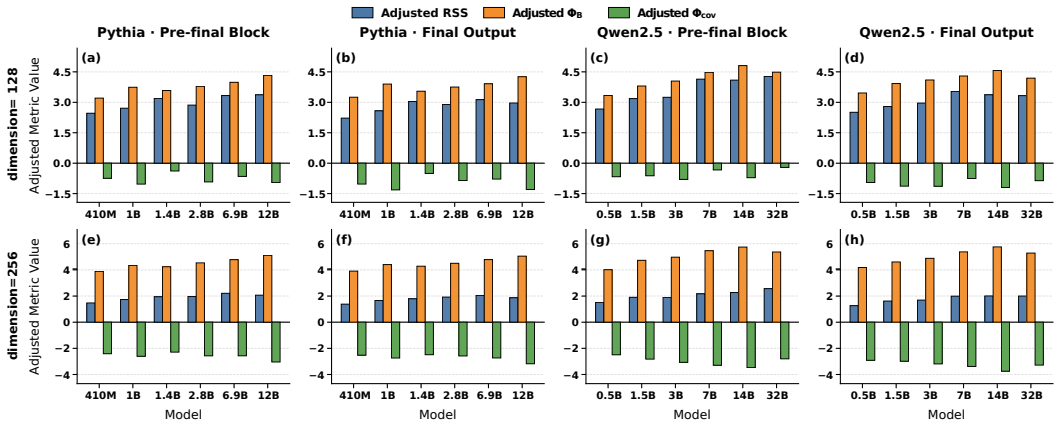

Figure 2: **Decomposition of adjusted skill structure.** Adjusted task- and template-conditioned skill structure, $\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D)$, is decomposed into the adjusted between-skill component Φ_B^{adj} and the adjusted covariance component $\Phi_{\mathrm{cov}}^{\mathrm{adj}}$ across model scales. Columns correspond to the pre-final-block and final-output representation locations for Pythia and Qwen2.5, while the top and bottom rows correspond to common representation dimensions of $q = 128$ and $q = 256$, respectively. Bars show $\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D)$ and its two adjusted geometric components at each model scale.

Theorem 2 shows that held-out skill recoverability depends on fitted prototype separation relative to prototype-estimation error and within-skill variation. Larger $\widehat{\Delta}_c$ tightens the recovery-error bound when η_c and σ_c are controlled. Combined with Corollary 1, this establishes the theoretical link from calibrated between-skill separation to more reliable held-out skill recovery: increasing $[\Phi_{B,c}^{\mathrm{adj}}]_+$ raises the guaranteed prototype-separation floor, which can tighten the recovery bound when the remaining terms do not deteriorate with model scale. Complete proofs are provided in Appendix A.

5 Experimental

5.1 Experimental Setup

Models and tasks. We evaluate six Pythia (Biderman et al., 2023) models ranging from 410M to 12B parameters and six Qwen2.5 (Yang et al., 2024) models ranging from 0.5B to 32B. Our main evaluation uses ARC-Challenge, ARC-Easy (Clark et al., 2018), and SciQ (Welbl et al., 2017). All models are evaluated on the same fixed population of 1,497 examples with shared skill annotations and three semantically equivalent prompt templates. Further details are provided in Appendix B.

Representation extraction. Our primary representation is the residual-stream state at the final token of the prompted input, immediately before the last Transformer block. We also evaluate the output of the final Transformer block as a robustness check. All representations are extracted before answer generation.

Skill structure and held-out recoverability. We measure skill structure using $\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D)$ after conditioning on task and prompt template. Representations from all model scales are mapped to common dimensions before comparison, and matched random-label partitions are used to calibrate structure arising from arbitrary label assignments. For held-out evaluation, skill prototypes are estimated from fitting examples only and kept fixed when identifying skills on unseen examples. We use disjoint 70/30 fitting–evaluation splits and report the held-out prototype margin as our primary measure of skill recoverability.

Behavioral evaluation and intervention. We separately evaluate task performance and test whether skill-aligned representation subspaces contribute functionally to answer formation. At the pre-final-block location, let $\tilde{h} = h - \mu_c$ denote the representation centered within task-template condition c, and let $P_U(\tilde{h}) = UU^\top \tilde{h}$ denote its component in the skill-aligned subspace. We apply the relative-energy intervention $\Delta h_{\mathrm{skill}} = \eta \|\tilde{h}\|_2 P_U(\tilde{h}) / \|P_U(\tilde{h})\|_2$ and set $h' = h - \Delta h_{\mathrm{skill}}$. We compare this intervention with an orthogonal random-subspace perturbation of the same per-example

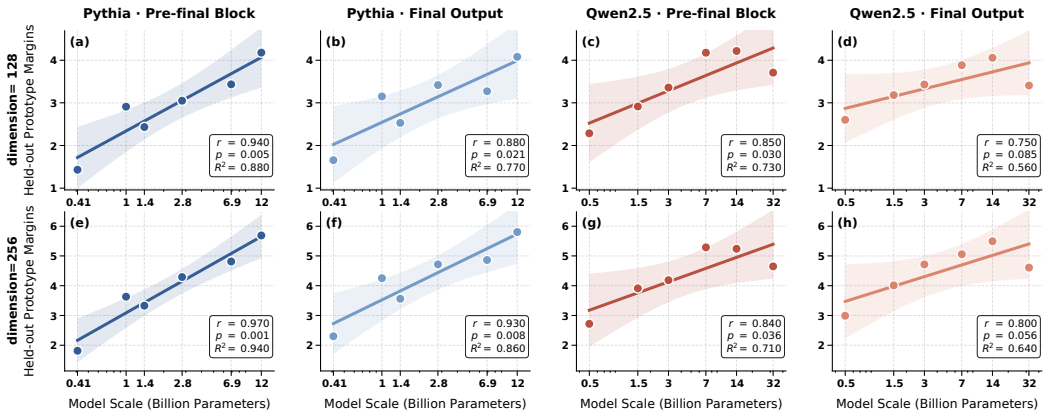

Figure 3: **Held-out prototype margins generally increase with model scale.** Held-out prototype margins are shown across model scales for Pythia (a,b,e,f) and Qwen2.5 (c,d,g,h). Columns correspond to the pre-final-block and final-output representation locations, while the top and bottom rows correspond to common representation dimensions of $q = 128$ and $q = 256$, respectively. Larger margins indicate that unseen examples are more strongly separated from their nearest competing skill prototypes. Points denote individual model scales, and solid lines show fitted scale trends. Pearson correlation coefficients (r), corresponding p-values, and R^2 are reported in each panel.

norm. Geometry-based calibration selects $\eta = 0.04$ for the primary analysis, with $\eta = 0.02$ used as a lower-strength sensitivity analysis. Let ΔG_{skill} and ΔG_{rand} denote the reductions in mean token log-probability of the full gold-choice continuation. We define $U_G = \Delta G_{\text{skill}} - \Delta G_{\text{rand}}$, where $U_G > 0$ indicates a larger behavioral effect from the skill-aligned intervention. Implementation details and robustness analyses are provided in Appendix C.

5.2 Skill Structure Strengthens with Model Scale

We examine the scale dependence of skill structure using $\mathrm{RSS}_{\text{adj}}(S \mid T, D)$ after conditioning on task and prompt template. As shown in Figure 1, larger models generally exhibit stronger skill structure across both Pythia and Qwen2.5. This tendency is observed at both representation locations and under both common dimensions, although Pythia exhibits larger local fluctuations across scales. Because the trend persists after mapping all representations to a common dimension, it cannot be attributed solely to differences in native representation dimensionality.

We next decompose $\mathrm{RSS}_{\text{adj}}(S \mid T, D)$ according to Theorem 1. As shown in Figure 2, the scale-dependent increase in skill structure is driven primarily by the adjusted between-skill separation component Φ_B^{adj}, which remains positive and generally increases with model scale across both model families and representation locations. In contrast, the adjusted covariance component $\Phi_{\text{cov}}^{\text{adj}}$ is consistently negative and partially offsets the between-skill contribution. Thus, the observed increase in $\mathrm{RSS}_{\text{adj}}(S \mid T, D)$ is associated primarily with greater calibrated separation among skill-conditioned representation centers, rather than differences in their covariance structure.

5.3 Skill Recoverability on Unseen Examples

Our geometric analysis indicates that the scale-dependent change in skill structure is driven primarily by greater separation among skill-conditioned representation centers. We next ask whether skill structure estimated from fitting examples remains recoverable on unseen examples. We estimate skill prototypes from the fitting examples and keep them fixed when evaluating previously unseen examples. Held-out skill recoverability is evaluated using the prototype margin defined in Section 3, which measures how much closer an unseen example is to its correct skill prototype than to its nearest competing prototype.

As shown in Figure 3, the held-out prototype margin generally increases with model scale under both common representation dimensions. The trend is particularly consistent in Pythia: at the pre-final-block location, the correlations with scale are $r = 0.94$ and $r = 0.97$ for $q = 128$ and $q = 256$, respectively, with positive trends also observed at the final-output location. Qwen2.5 shows the

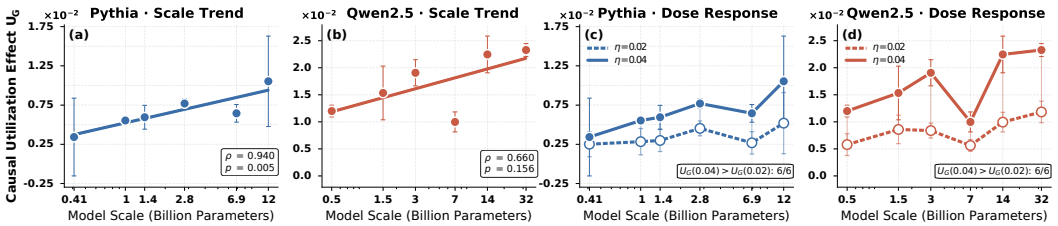

Figure 4: **Skill-aligned representation subspaces contribute functionally to answer formation.** The causal utilization effect U_G is shown for Pythia and Qwen2.5. Panels (a,b) show its relationship with model scale at relative intervention strength $\eta = 0.04$, while panels (c,d) compare $\eta = 0.02$ and $\eta = 0.04$. Positive U_G indicates that the skill-aligned intervention produces a larger reduction in the gold-answer score than a relative-energy-matched random intervention. Points denote means across three split seeds, with error bars indicating standard deviations. In (a,b), solid lines show fitted scale trends, with Spearman coefficients (ρ) and corresponding p-values reported. In (c,d), the stronger intervention yields larger U_G at all six model scales in both families. The dashed horizontal line marks $U_G = 0$.

same overall pattern at the pre-final block, while the relationship is weaker at the final output. Thus, the scale dependence is most consistent at the pre-final-block location used in our primary analysis, while remaining directionally similar at the final-output location.

Section 4.2 shows that greater calibrated between-skill separation raises the guaranteed lower bound on fitted prototype separation, while Theorem 2 shows that larger prototype separation can tighten the upper bound on held-out skill-recovery error when prototype-estimation error and within-skill variation are controlled. Consistent with this theoretical link, held-out prototype margins generally increase with model scale, with the clearest trends at the pre-final-block location. These results indicate that skill structure estimated from fitting examples remains recoverable on unseen examples.

5.4 Skill-Aligned Subspaces Contribute Functionally to Answer Formation

The held-out analysis indicates that skill structure estimated from fitting examples remains recoverable on unseen examples, but does not establish whether the identified representation geometry contributes functionally to answer formation. We therefore intervene on skill-aligned representation subspaces at the pre-final-block location and compare their effects with relative-energy-matched random subspaces.

As shown in Figure 4, U_G is positive at every evaluated scale in both model families. At $\eta = 0.04$, Pythia shows a clear positive relationship with model scale, with a Spearman correlation of $\rho = 0.94$ ($p = 0.005$). Qwen2.5 also shows positive effects across all scales, although its relationship with scale is less regular, with $\rho = 0.66$ ($p = 0.156$). Thus, the identified skill-aligned subspaces contribute functionally to answer formation in both model families, whereas evidence for scale-dependent strengthening is substantially clearer in Pythia. The intervention-strength comparison provides an additional consistency check. Increasing the relative perturbation strength from $\eta = 0.02$ to $\eta = 0.04$ increases U_G at all six scales in both Pythia and Qwen2.5. This dose-dependent response further supports the interpretation that the observed behavioral effect is specific to the targeted skill-aligned subspaces rather than reflecting a generic consequence of perturbing the representation.

Together with the preceding skill-structure and held-out recoverability results, these findings argue against interpreting the identified skill structure as merely a decodable geometric correlate. Skill structure generally strengthens with model scale and remains recoverable on unseen examples, while targeted perturbations of skill-aligned subspaces selectively affect answer formation. Together, these results indicate that the identified skill structure contains functionally relevant representation components, without implying that it fully mediates the behavioral effects of scaling.

6 Representation-Guided Data-Efficient Adaptation

The preceding results show that skill structure can be characterized in representation space and that skill-aligned representation components are functionally relevant to answer formation. We next ask

Table 1: Held-out `exam_style` equal-task Macro accuracy (%) under a fixed adaptation budget of $B = 256$ underlying training examples. Values are means over three pipeline replicates. *Ours w/o diagnosis* uses the same skill-coverage and representation-diversity procedure as the full method but removes the representation-derived skill-weakness reweighting.

Model	Base	Random	Skill-bal.	High-loss	K-center	Ours w/o diagnosis	Ours
Pythia-2.8B	35.65	49.19	49.27	46.29	48.78	49.22	**50.43**
Qwen2.5-3B	78.37	89.52	89.19	89.65	90.73	90.46	**90.75**

whether the representation diagnostics developed above can guide adaptation-data selection under a limited training budget.

Representation-guided selection. Using a diagnosis split disjoint from the candidate adaptation pool, we estimate a model-specific weakness score for each skill from two representation diagnostics: relative between-skill separation and leave-one-out prototype recoverability. Larger scores correspond to skills whose representations are less separated and less reliably recoverable.

Our selector combines this diagnosis with skill coverage and representation diversity. We first reserve a minimum number of examples for each skill and then fill the remaining task budget using a global K-center criterion. The complete method applies a bounded, one-sided reweighting that increases the priority of candidates associated with representation-weak skills without downweighting other skills. A matched variant, *Ours w/o diagnosis*, uses the same coverage and diversity procedure but removes the weakness-based reweighting.

All diagnosis and selection scores are computed exclusively from the official training data. The frozen evaluation population is not used for representation diagnosis, candidate selection, or LoRA training, apart from defining the shared task–skill ontology. We compare against Random, Skill-balanced, High-loss, and global K-center selection under the same adaptation budgets and LoRA training protocol. Full experimental details are provided in Appendix D.

Cross-family evaluation. We evaluate whether the representation-derived weakness signal remains useful across model families using Pythia-2.8B and Qwen2.5-3B under the same 256-example adaptation budget. For Pythia-2.8B, our method achieves 50.43% held-out Macro accuracy, exceeding the standard baselines. More importantly, it improves over the matched *Ours w/o diagnosis* variant from 49.22% to 50.43%, a gain of 1.21 percentage points. For Qwen2.5-3B, our method reaches 90.75%, matching the K-center baseline while improving over *Ours w/o diagnosis* from 90.46% to 90.75%; the improvement is observed in all three replicates.

These results provide cross-family evidence that representation-derived skill weakness contains useful information beyond skill coverage and representation diversity alone. In particular, the matched comparison with *Ours w/o diagnosis* indicates that the representation diagnostics can provide an additional signal for prioritizing adaptation examples under a limited training budget.

7 Discussion

Taken together, our findings support a representation-level view in which model scaling is accompanied by stronger skill structure and more reliable skill recoverability on unseen examples. The geometric analysis further links this trend to increased between-skill separation, while the intervention results show that skill-aligned representation subspaces contribute functionally to answer formation. These representation diagnostics can also provide useful signals for data-efficient adaptation under limited training budgets. Our analysis uses a fixed, model-independent semantic skill partition whose representational relevance is evaluated through random-label calibration, held-out recovery, and intervention, rather than assumed to reflect a unique intrinsic decomposition of the model. Moreover, the intervention results establish functional relevance but do not imply that the identified structure fully mediates the behavioral effects of scaling. Extending this framework to model-derived functional units, broader tasks, architectures, and training stages remains an important direction for future work.

AI USE DISCLOSURE

In this work, we used GPT-5.6 to assist with the preparation and refinement of figures included in the paper. We also used GPT-5.6 to assist with drafting, editing, and polishing portions of the manuscript to improve clarity, readability, and presentation. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, March 2022. doi: 10.1162/coli_a_00422. URL https://aclanthology.org/2022.cl-1.7/.

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. Leace: Perfect linear concept erasure in closed form. *ArXiv*, abs/2306.03819, 2023. URL https://api.semanticscholar.org/CorpusID:259088549.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. Pythia: A suite for analyzing large language models across training and scaling. In *International conference on machine learning*, pp. 2397–2430. PMLR, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *ArXiv*, abs/1803.05457, 2018. URL https://api.semanticscholar.org/CorpusID:3922816.

Fahim Dalvi, Nadir Durrani, Hassan Sajjad, Yonatan Belinkov, Anthony Bau, and James R. Glass. What is one grain of sand in the desert? analyzing individual neurons in deep nlp models. In *AAAI Conference on Artificial Intelligence*, 2018. URL https://api.semanticscholar.org/CorpusID:56895415.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL https://aclanthology.org/2021.tacl-1.10/.

Danny Hernandez, Jared Kaplan, Thomas Henighan, and Sam McCandlish. Scaling laws for transfer. *ArXiv*, abs/2102.01293, 2021. URL https://api.semanticscholar.org/CorpusID:231749962.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

John Hewitt and Christopher D. Manning. A structural probe for finding syntax in word representations. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4129–4138, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1419. URL https://aclanthology.org/N19-1419/.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

Guan Zhe Hong, Bhavya Vasudeva, Vatsal Sharan, Cyrus Rashtchian, Prabhakar Raghavan, and Rina Panigrahy. Latent concept disentanglement in transformer-based language models. In *International Conference on Learning Representations*, volume 2026, pp. 27223–27260, 2026.

Robert Huben, Hoagy Cunningham, Logan Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, volume 2024, pp. 7827–7845, 2024.

Jared Kaplan, Sam McCandlish, Thomas Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020. URL https://api.semanticscholar.org/CorpusID:210861095.

Jiazheng Li, Lu Yu, Qing Cui, Zhiqiang Zhang, Jun Zhou, Yanfang Ye, and Chuxu Zhang. Mass: Mathematical data selection via skill graphs for pretraining large language models. *arXiv preprint arXiv:2503.14917*, 2025a.

Melody Li, Kumar Krishna Agrawal, Arna Ghosh, Komal Teru, Adam Santoro, Guillaume Lajoie, and Blake Richards. Tracing the representation geometry of language models from pretraining to post-training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 54691–54724. Curran Associates, Inc., 2025b. doi: 10.52202/085713-1829. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/4eac651e6c06d6b374332cd14ab50f55-Paper-Conference.pdf.

Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. *ArXiv*, abs/2505.10465, 2025. URL https://api.semanticscholar.org/CorpusID:278636265.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. *ArXiv*, abs/2403.19647, 2024. URL https://api.semanticscholar.org/CorpusID:268732732.

Mazda Moayeri, Vidhisha Balachandran, Varun Chandrasekaran, Safoora Yousefi, Thomas Fel, Soheil Feizi, Besmira Nushi, Neel Joshi, and Vibhav Vineet. Unearthing skill-level insights for understanding trade-offs of foundation models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 34449–34476, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/5526c73e3ff4f2a34009e13d15f52fcb-Paper-Conference.pdf.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/park24c.html.

Florian Redhardt, Yassir Akram, and Simon Schug. Scaling can lead to compositional generalization. *Advances in Neural Information Processing Systems*, 38:55593–55633, 2026.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522, 2024.

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

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5797–5808, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1580. URL https://aclanthology.org/P19-1580/.

J Wei. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

Johannes Welbl, Nelson F. Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In Leon Derczynski, Wei Xu, Alan Ritter, and Tim Baldwin (eds.), *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4413. URL https://aclanthology.org/W17-4413/.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bo-Wen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yi-Chao Zhang, Yunyang Wan, Yuqi Liu, Zeyu Cui, Zhen-Ru Zhang, Zihan Qiu, Shanghaoran Quan, and Zekun Wang. Qwen2.5 technical report. *ArXiv*, abs/2412.15115, 2024. URL https://api.semanticscholar.org/CorpusID:274859421.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. Flask: Fine-grained language model evaluation based on alignment skill sets. *ArXiv*, abs/2307.10928, 2023. URL https://api.semanticscholar.org/CorpusID:259991144.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-mix: a flexible and expandable family of evaluations for ai models. *ArXiv*, abs/2310.17567, 2023. URL https://api.semanticscholar.org/CorpusID:264490642.

Yaodong Yu, Kwan Ho Ryan Chan, Chong You, Chaobing Song, and Yi Ma. Learning diverse and discriminative representations via the principle of maximal coding rate reduction. *ArXiv*, abs/2006.08558, 2020. URL https://api.semanticscholar.org/CorpusID:219687643.

A PROOFS OF THEORETICAL RESULTS

This appendix provides complete derivations of the theoretical results in Section 4. The analysis of skill structure is formulated at the empirical level, conditional on a fixed set of representations and empirical skill frequencies, consistent with the definition of $\mathrm{RSS}(S \mid T, D)$ in Section 3. The subsequent held-out analysis relates skill-prototype statistics estimated from $\mathcal{D}_{\mathrm{fit}}$ to representations of unseen examples in $\mathcal{D}_{\mathrm{eval}}$.

A.1 PROOF OF THE PERMUTATION-ADJUSTED GEOMETRIC DECOMPOSITION

We prove Theorem 1. Fix a task–template condition $c = (t, d)$ containing n_c examples. For skill s, let $n_{c,s}$ denote the corresponding number of examples and $p_{c,s} = n_{c,s}/n_c$. Let $\mu_{c,s}$ and $\Sigma_{c,s}$ denote the empirical mean and covariance of $Z_{c,s}$, respectively, where the covariance uses the $1/n_{c,s}$ normalization. The condition-level mean is

$$\mu_c = \sum_s p_{c,s}\mu_{c,s}. \tag{6}$$

Define the pooled within-skill covariance and the between-skill covariance as

$$\Sigma_{W,c} = \sum_s p_{c,s}\Sigma_{c,s}, \qquad B_c = \sum_s p_{c,s}(\mu_{c,s} - \mu_c)(\mu_{c,s} - \mu_c)^\top. \tag{7}$$

The finite-sample law of total covariance gives

$$\Sigma_c = \Sigma_{W,c} + B_c. \tag{8}$$

Using the coding rate defined in Section 3, and recalling that $\lambda_q = \delta^2/q$, we have

$$\mathcal{R}(Z) = \frac{1}{2}\log\det\Big(I + \frac{q}{\delta^2}\widehat{\Sigma}\Big) = \frac{1}{2}\log\det\Big(\frac{q}{\delta^2}(\widehat{\Sigma} + \lambda_q I)\Big).$$

Hence,

$$\mathcal{R}(Z) = \frac{q}{2}\log\Big(\frac{q}{\delta^2}\Big) + \frac{1}{2}\log\det(\widehat{\Sigma} + \lambda_q I).$$

Because the additive term $\frac{q}{2}\log(q/\delta^2)$ is the same for the pooled and skill-conditioned coding rates and $\sum_s p_{c,s} = 1$, it cancels in their difference. Therefore, the contribution of condition c to the unadjusted skill-structure measure is

$$\mathrm{RSS}_c(S) = \frac{1}{2}\log\det(\Sigma_c + \lambda_q I) - \frac{1}{2}\sum_s p_{c,s}\log\det(\Sigma_{c,s} + \lambda_q I). \tag{9}$$

Substituting Equation 8 and adding and subtracting $\frac{1}{2}\log\det(\Sigma_{W,c} + \lambda_q I)$ yields

$$\begin{aligned}\mathrm{RSS}_c(S) = \frac{1}{2}\Big[\log\det(\Sigma_{W,c} + B_c + \lambda_q I) - \log\det(\Sigma_{W,c} + \lambda_q I)\Big] \\ + \frac{1}{2}\Big[\log\det(\Sigma_{W,c} + \lambda_q I) - \sum_s p_{c,s}\log\det(\Sigma_{c,s} + \lambda_q I)\Big].\end{aligned} \tag{10}$$

Let

$$A_c = \Sigma_{W,c} + \lambda_q I. \tag{11}$$

Since $\lambda_q > 0$, we have $A_c \succ 0$, and

$$A_c + B_c = A_c^{1/2}\left(I + A_c^{-1/2}B_cA_c^{-1/2}\right)A_c^{1/2}. \tag{12}$$

Therefore,

$$\det(A_c + B_c) = \det(A_c)\det\left(I + A_c^{-1/2}B_cA_c^{-1/2}\right). \tag{13}$$

The first term in Equation 10 is thus

$$\Phi_{B,c} = \frac{1}{2}\log\det\left(I + A_c^{-1/2}B_cA_c^{-1/2}\right), \tag{14}$$

whereas the second is

$$\Phi_{\mathrm{cov},c} = \frac{1}{2}\Big[\log\det(\Sigma_{W,c} + \lambda_q I) - \sum_s p_{c,s}\log\det(\Sigma_{c,s} + \lambda_q I)\Big]. \tag{15}$$

Hence

$$\mathrm{RSS}_c(S) = \Phi_{B,c} + \Phi_{\mathrm{cov},c}. \tag{16}$$

Both raw components are non-negative. Since $B_c \succeq 0$,

$$G_c = A_c^{-1/2} B_c A_c^{-1/2} \succeq 0. \tag{17}$$

If $\gamma_1, \ldots, \gamma_q \geq 0$ are the eigenvalues of G_c, then

$$\Phi_{B,c} = \frac{1}{2}\sum_{j=1}^{q}\log(1+\gamma_j) \geq 0. \tag{18}$$

Moreover,

$$\Sigma_{W,c} + \lambda_q I = \sum_s p_{c,s}(\Sigma_{c,s} + \lambda_q I). \tag{19}$$

By concavity of $\log\det(\cdot)$ on the positive-definite cone,

$$\log\det(\Sigma_{W,c} + \lambda_q I) \geq \sum_s p_{c,s}\log\det(\Sigma_{c,s} + \lambda_q I), \tag{20}$$

which implies

$$\Phi_{\mathrm{cov},c} \geq 0. \tag{21}$$

We next apply the matched permutation adjustment defined in Section 3. Let π denote a random permutation of the skill labels within condition c, keeping Z_c fixed and preserving the empirical skill frequencies. Applying Equation 16 to the permuted partition gives

$$\mathrm{RSS}_c^{(\pi)}(S) = \Phi_{B,c}^{(\pi)} + \Phi_{\mathrm{cov},c}^{(\pi)}. \tag{22}$$

Subtracting the permutation expectation from the observed decomposition gives

$$\begin{aligned}
\mathrm{RSS}_{\mathrm{adj},c}(S) &= \mathrm{RSS}_c(S) - \mathbb{E}_\pi\left[\mathrm{RSS}_c^{(\pi)}(S)\right] \\
&= \left(\Phi_{B,c} - \mathbb{E}_\pi\left[\Phi_{B,c}^{(\pi)}\right]\right) + \left(\Phi_{\mathrm{cov},c} - \mathbb{E}_\pi\left[\Phi_{\mathrm{cov},c}^{(\pi)}\right]\right) \\
&= \Phi_{B,c}^{\mathrm{adj}} + \Phi_{\mathrm{cov},c}^{\mathrm{adj}}.
\end{aligned} \tag{23}$$

Averaging over task–template conditions,

$$\begin{aligned}
\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D) &= \sum_c p(c)\,\mathrm{RSS}_{\mathrm{adj},c}(S) \\
&= \Phi_B^{\mathrm{adj}} + \Phi_{\mathrm{cov}}^{\mathrm{adj}},
\end{aligned} \tag{24}$$

where

$$\Phi_B^{\mathrm{adj}} = \sum_c p(c)\Phi_{B,c}^{\mathrm{adj}}, \qquad \Phi_{\mathrm{cov}}^{\mathrm{adj}} = \sum_c p(c)\Phi_{\mathrm{cov},c}^{\mathrm{adj}}. \tag{25}$$

This proves Theorem 1.

The adjusted components need not be non-negative because they measure deviations from their matched permutation expectations. In particular,

$$\Phi_{\mathrm{cov},c}^{\mathrm{adj}} < 0 \quad \Longleftrightarrow \quad \Phi_{\mathrm{cov},c} < \mathbb{E}_\pi\left[\Phi_{\mathrm{cov},c}^{(\pi)}\right], \tag{26}$$

which is fully consistent with the raw quantity $\Phi_{\mathrm{cov},c} \geq 0$.

A.2 Proof of the Prototype-Separation Bridge

We prove Proposition 1 and Corollary 1. All quantities in this subsection are evaluated on $\mathcal{D}_{\text{fit}}$. Hence, the empirical skill-conditioned centers defining B_c coincide with the fitted skill prototypes, so that $\mu_{c,s} = \widehat{\mu}_{c,s}$. For brevity, write $p_s = p(s \mid c)$.

Recall that

$$u_{c,s} = A_c^{-1/2}(\widehat{\mu}_{c,s} - \widehat{\mu}_c), \qquad \widehat{\mu}_c = \sum_s p_s \widehat{\mu}_{c,s}. \tag{27}$$

Since

$$\sum_s p_s(\widehat{\mu}_{c,s} - \widehat{\mu}_c) = 0,$$

we have

$$\sum_s p_s u_{c,s} = 0.$$

Moreover,

$$G_c = \sum_s p_s u_{c,s} u_{c,s}^\top, \qquad \operatorname{tr}(G_c) = \sum_s p_s \|u_{c,s}\|_2^2. \tag{28}$$

Expanding the weighted pairwise squared distances,

$$\begin{aligned}\sum_{s,s'} p_s p_{s'} \|u_{c,s} - u_{c,s'}\|_2^2 &= 2\sum_s p_s \|u_{c,s}\|_2^2 - 2\left\|\sum_s p_s u_{c,s}\right\|_2^2 \\ &= 2\operatorname{tr}(G_c).\end{aligned} \tag{29}$$

Because the terms with $s = s'$ vanish and

$$h_c = \sum_{s \neq s'} p_s p_{s'},$$

the weighted average pairwise separation satisfies

$$\bar{d}_c^2 = \frac{2\operatorname{tr}(G_c)}{h_c}. \tag{30}$$

Let $\gamma_1, \ldots, \gamma_{r_c} > 0$ denote the nonzero eigenvalues of G_c, where $r_c = \operatorname{rank}(G_c)$. Since

$$2\Phi_{B,c} = \sum_{j=1}^{r_c} \log(1 + \gamma_j), \tag{31}$$

the arithmetic–geometric mean inequality gives

$$\begin{aligned}\exp\left(\frac{2\Phi_{B,c}}{r_c}\right) &= \left[\prod_{j=1}^{r_c}(1+\gamma_j)\right]^{1/r_c} \\ &\leq 1 + \frac{1}{r_c}\sum_{j=1}^{r_c}\gamma_j = 1 + \frac{\operatorname{tr}(G_c)}{r_c}.\end{aligned} \tag{32}$$

Therefore,

$$\operatorname{tr}(G_c) \geq r_c\left[\exp\left(\frac{2\Phi_{B,c}}{r_c}\right) - 1\right]. \tag{33}$$

Combining this inequality with Equation 30 yields

$$\bar{d}_c^2 \geq \frac{2r_c}{h_c}\left[\exp\left(\frac{2\Phi_{B,c}}{r_c}\right) - 1\right], \tag{34}$$

which proves Proposition 1.

Permutation calibration. For every matched permutation π, $\Phi_{B,c}^{(\pi)} \geq 0$. Hence,

$$\Phi_{B,c} = \Phi_{B,c}^{\mathrm{adj}} + \mathbb{E}_\pi \left[\Phi_{B,c}^{(\pi)}\right] \geq [\Phi_{B,c}^{\mathrm{adj}}]_+, \tag{35}$$

where $[x]_+ = \max(x, 0)$. Since the lower bound in Proposition 1 is monotonically increasing in $\Phi_{B,c}$,

$$\bar{d}_c^2 \geq \frac{2r_c}{h_c}\left[\exp\left(\frac{2[\Phi_{B,c}^{\mathrm{adj}}]_+}{r_c}\right) - 1\right]. \tag{36}$$

Thus, the positive permutation-adjusted excess yields a conservative lower bound on the average normalized separation among fitted skill prototypes.

From average to minimum separation. Define

$$d_{\min,c}^2 = \min_{s \neq s'} \|u_{c,s} - u_{c,s'}\|_2^2, \qquad \beta_c = \begin{cases} \dfrac{d_{\min,c}^2}{\bar{d}_c^2}, & \bar{d}_c^2 > 0, \\ 0, & \bar{d}_c^2 = 0. \end{cases} \tag{37}$$

Then $0 \leq \beta_c \leq 1$ and $d_{\min,c}^2 = \beta_c \bar{d}_c^2$. For every $s \neq s'$,

$$\begin{aligned} \|\widehat{\mu}_{c,s} - \widehat{\mu}_{c,s'}\|_2^2 &= (u_{c,s} - u_{c,s'})^\top A_c (u_{c,s} - u_{c,s'}) \\ &\geq \lambda_{\min}(A_c)\|u_{c,s} - u_{c,s'}\|_2^2. \end{aligned} \tag{38}$$

Taking the minimum over all skill pairs therefore gives

$$\widehat{\Delta}_c^2 \geq \lambda_{\min}(A_c)\beta_c \bar{d}_c^2. \tag{39}$$

Scale-comparable bound. Let m index model scale and let K_c denote the number of active skills under condition c. Because the same labeled fitting population is used across model scales, $p(s \mid c)$ and therefore h_c are independent of m. Moreover,

$$A_{c,m} = \Sigma_{W,c,m} + \lambda_q I \succeq \lambda_q I, \tag{40}$$

which implies

$$\lambda_{\min}(A_{c,m}) \geq \lambda_q. \tag{41}$$

Since the centered normalized prototype displacements satisfy

$$\sum_s p_s u_{c,s} = 0,$$

they span at most $K_c - 1$ dimensions. Thus,

$$r_{c,m} \leq R_c, \qquad R_c := K_c - 1. \tag{42}$$

If $[\Phi_{B,c,m}^{\mathrm{adj}}]_+ = 0$, the lower bound in Corollary 1 is zero, and the result follows immediately from $\widehat{\Delta}_{c,m}^2 \geq 0$.

Now suppose $[\Phi_{B,c,m}^{\mathrm{adj}}]_+ > 0$. Then $\Phi_{B,c,m} > 0$, and hence $r_{c,m} \geq 1$. For $\phi > 0$, define

$$g_\phi(r) = r\left[\exp\left(\frac{2\phi}{r}\right) - 1\right]. \tag{43}$$

Its derivative is

$$g_\phi'(r) = \exp\left(\frac{2\phi}{r}\right)\left(1 - \frac{2\phi}{r}\right) - 1 \leq 0, \tag{44}$$

so $g_\phi(r)$ is non-increasing in r. Since $r_{c,m} \leq R_c$,

$$r_{c,m}\left[\exp\left(\frac{2[\Phi_{B,c,m}^{\mathrm{adj}}]_+}{r_{c,m}}\right) - 1\right] \geq R_c\left[\exp\left(\frac{2[\Phi_{B,c,m}^{\mathrm{adj}}]_+}{R_c}\right) - 1\right]. \tag{45}$$

Under Assumption 1, $\beta_{c,m} \geq \underline{\beta}_c > 0$. Combining Equation 39, the calibrated average-separation bound, $\lambda_{\min}(A_{c,m}) \geq \lambda_q$, and $r_{c,m} \leq R_c$ yields

$$\widehat{\Delta}_{c,m}^2 \geq \frac{2\lambda_q \underline{\beta}_c R_c}{h_c}\left[\exp\left(\frac{2[\Phi_{B,c,m}^{\mathrm{adj}}]_+}{R_c}\right) - 1\right]. \tag{46}$$

This proves Corollary 1.

A.3 Proof of the Held-Out Prototype-Margin Bound

We prove Lemma 2. Consider a held-out example of true skill s under condition c, with

$$z = \mu^\star_{c,s} + \xi. \tag{47}$$

By the prototype-estimation assumption,

$$\|\widehat{\mu}_{c,s} - \mu^\star_{c,s}\|_2 \le \eta_c. \tag{48}$$

Hence, by the triangle inequality,

$$\begin{aligned} \|z - \widehat{\mu}_{c,s}\|_2 &\le \|z - \mu^\star_{c,s}\|_2 + \|\mu^\star_{c,s} - \widehat{\mu}_{c,s}\|_2 \\ &\le \|\xi\|_2 + \eta_c. \end{aligned} \tag{49}$$

For any competing skill $s' \neq s$, the triangle inequality gives

$$\begin{aligned} \|z - \widehat{\mu}_{c,s'}\|_2 &\ge \|\widehat{\mu}_{c,s} - \widehat{\mu}_{c,s'}\|_2 - \|z - \widehat{\mu}_{c,s}\|_2 \\ &\ge \widehat{\Delta}_c - \|z - \widehat{\mu}_{c,s}\|_2. \end{aligned} \tag{50}$$

Taking the minimum over all competing skills and subtracting the distance to the correct prototype yields

$$\begin{aligned} M_c(z) &= \min_{s' \neq s} \|z - \widehat{\mu}_{c,s'}\|_2 - \|z - \widehat{\mu}_{c,s}\|_2 \\ &\ge \widehat{\Delta}_c - 2\|z - \widehat{\mu}_{c,s}\|_2. \end{aligned} \tag{51}$$

Applying Equation 49, we obtain

$$M_c(z) \ge \widehat{\Delta}_c - 2\|\xi\|_2 - 2\eta_c. \tag{52}$$

This proves Lemma 2.

A.4 Proof of the Held-Out Skill-Recoverability Bound

We prove Theorem 2. Under the nearest-prototype rule, a misclassification can occur only if the correct prototype does not have strictly positive margin. Hence,

$$\widehat{S}(z; c) \neq s \quad \Longrightarrow \quad M_c(z) \le 0. \tag{53}$$

By Lemma 2,

$$M_c(z) \ge \widehat{\Delta}_c - 2\|\xi\|_2 - 2\eta_c. \tag{54}$$

Therefore, every misclassified example must satisfy

$$\|\xi\|_2 \ge \frac{\widehat{\Delta}_c - 2\eta_c}{2}. \tag{55}$$

When $\widehat{\Delta}_c > 2\eta_c$, this threshold is positive, so

$$\Pr\left(\widehat{S}(z; c) \neq s \mid c, s\right) \le \Pr\left(\|\xi\|_2 \ge \frac{\widehat{\Delta}_c - 2\eta_c}{2} \,\middle|\, c, s\right). \tag{56}$$

Under the held-out tail assumption,

$$\Pr\left(\|\xi\|_2 \ge r \mid c, s\right) \le C \exp\left(-\frac{r^2}{2\sigma_c^2}\right). \tag{57}$$

Substituting

$$r = \frac{\widehat{\Delta}_c - 2\eta_c}{2}$$

into Equation 56 yields

$$\Pr\left(\widehat{S}(z; c) \neq s \mid c, s\right) \le C \exp\left[-\frac{(\widehat{\Delta}_c - 2\eta_c)^2}{8\sigma_c^2}\right]. \tag{58}$$

This proves Theorem 2.

A.5 Effective Prototype Signal-to-Variation Ratio

The held-out recoverability bound depends on prototype separation only relative to prototype-estimation error and within-skill variation. When $\widehat{\Delta}_c > 2\eta_c$, define the effective prototype signal-to-variation ratio as

$$\Gamma_c = \frac{\widehat{\Delta}_c - 2\eta_c}{\sigma_c} > 0. \tag{59}$$

Theorem 2 can then be written as

$$\Pr\left(\widehat{S}(z;c) \neq s \mid c, s\right) \leq C \exp\left(-\frac{\Gamma_c^2}{8}\right). \tag{60}$$

Thus, the guarantee is governed by fitted prototype separation relative to prototype-estimation error and within-skill variation, rather than by absolute Euclidean separation alone.

This form also makes the bound invariant to a uniform rescaling of the representation space. If all representations are transformed as $z \mapsto az$ for some $a \neq 0$, then

$$\widehat{\Delta}_c \mapsto |a|\widehat{\Delta}_c, \qquad \eta_c \mapsto |a|\eta_c, \qquad \sigma_c \mapsto |a|\sigma_c. \tag{61}$$

Consequently,

$$\Gamma_c \mapsto \frac{|a|\widehat{\Delta}_c - 2|a|\eta_c}{|a|\sigma_c} = \Gamma_c. \tag{62}$$

Therefore, uniformly enlarging all Euclidean distances in representation space does not, by itself, strengthen the held-out recoverability guarantee.

Scale-comparable recovery consequence. Suppose that the prototype-estimation error and within-skill variation are uniformly bounded across model scales:

$$\eta_{c,m} \leq \overline{\eta}_c, \qquad \sigma_{c,m} \leq \overline{\sigma}_c. \tag{63}$$

Define

$$L_c(\phi) = \sqrt{\kappa_c \left[\exp\left(\frac{2[\phi]_+}{R_c}\right) - 1\right]}. \tag{64}$$

By Corollary 1,

$$\widehat{\Delta}_{c,m} \geq L_c\left(\Phi^{\mathrm{adj}}_{B,c,m}\right). \tag{65}$$

Whenever

$$L_c\left(\Phi^{\mathrm{adj}}_{B,c,m}\right) > 2\overline{\eta}_c, \tag{66}$$

we also have $\widehat{\Delta}_{c,m} > 2\eta_{c,m}$, so Theorem 2 applies. Using $\eta_{c,m} \leq \overline{\eta}_c$ and $\sigma_{c,m} \leq \overline{\sigma}_c$ then gives

$$\Pr\left(\widehat{S}(z;c) \neq s \mid c, s, m\right) \leq C \exp\left[-\frac{\left(L_c(\Phi^{\mathrm{adj}}_{B,c,m}) - 2\overline{\eta}_c\right)^2}{8\overline{\sigma}_c^2}\right]. \tag{67}$$

For $\phi > 0$, $L_c(\phi)$ is strictly increasing. Therefore, under the uniform regularity conditions above, a larger positive calibrated between-skill component raises the guaranteed prototype-separation floor and monotonically tightens the corresponding held-out recovery bound.

A.6 Interpretation of the Skill-Structure–Recoverability Chain

The theoretical results above establish a three-stage chain from calibrated skill structure to skill-prototype separation and, subsequently, to held-out recoverability.

First, $\Phi^{\mathrm{adj}}_{B,c}$ measures between-skill separation beyond the matched permutation expectation. The corresponding raw component $\Phi_{B,c}$ is linked to the normalized fitted-prototype geometry through Proposition 1, while Corollary 1 uses only the positive calibrated excess to obtain a conservative

Table 2: Composition of the frozen evaluation population.

Dataset	# Examples	# Skills
ARC-Challenge	278	8
ARC-Easy	885	27
SciQ	334	4
Total	1,497	39

lower bound on minimum fitted-prototype separation. Under the uniform non-collapse assumption, this bound remains comparable across model scales.

Second, greater fitted-prototype separation improves the held-out recovery guarantee only when it is sufficiently large relative to prototype-estimation error and within-skill variation. Under uniform bounds on these quantities, Equation 67 makes this dependence explicit: a larger positive adjusted between-skill component raises the guaranteed prototype-separation floor and tightens the corresponding held-out recovery bound.

Why the covariance component does not enter this chain. No analogous prototype-separation guarantee follows from $\Phi_{\mathrm{cov},c}$ alone. Consider two equally weighted skills, $p(1 \mid c) = p(2 \mid c) = 1/2$, with identical centers, $\mu_{c,1} = \mu_{c,2}$, but distinct covariance matrices $\Sigma_{c,1} \neq \Sigma_{c,2}$. In this case, $B_c = 0$, so the prototype separation is zero. However,

$$\Sigma_{W,c} = \frac{1}{2}\left(\Sigma_{c,1} + \Sigma_{c,2}\right),$$

and strict concavity of $\log\det(\cdot)$ gives

$$\log\det(\Sigma_{W,c} + \epsilon I) > \frac{1}{2}\log\det(\Sigma_{c,1} + \epsilon I) + \frac{1}{2}\log\det(\Sigma_{c,2} + \epsilon I),$$

so that $\Phi_{\mathrm{cov},c} > 0$. Covariance heterogeneity can therefore remain positive even when the skill prototypes completely coincide.

This does not imply that covariance structure is uninformative for held-out prediction. Rather, exploiting it would require a covariance-sensitive decision rule together with additional assumptions on covariance estimation. The present theory focuses on the between-skill component because it is directly connected to the prototype-based recoverability analysis used in this work.

B Data Construction and Skill Annotation

B.1 Evaluation Tasks and Frozen Population

Our main analysis uses three multiple-choice datasets: ARC-Challenge, ARC-Easy, and SciQ. Before any cross-scale representation analysis, we construct a single frozen evaluation population and reuse the same underlying examples, skill annotations, and prompt variants across all Pythia and Qwen2.5 models. This ensures that differences across model scales are not confounded by changes in example composition.

After applying the frequency requirement needed to provide sufficient support for skill-level estimation, the frozen population contains 1,497 underlying examples spanning 39 task–skill groups, as summarized in Table 2. ARC-Challenge contributes 278 examples across 8 skills, ARC-Easy contributes 885 examples across 27 skills, and SciQ contributes 334 examples across 4 skills. Each underlying example is rendered with three prompt templates, yielding 4,491 prompt instances per model.

B.2 Skill Annotation

We treat skills as an abstraction of latent capability variations reflected in representations and relevant to solving downstream examples. Each example is assigned a single primary skill that captures its main functional requirement. The skill vocabulary is constructed in a separate annotation stage before the cross-scale analysis and is then frozen for all subsequent experiments.

Inspired by the skill-annotation procedure of Li et al. (2025a), we use an external instruction-tuned language model to construct semantically interpretable skill labels through a multi-stage annotation pipeline. Candidate skills are first proposed from representative examples and then consolidated into a canonical vocabulary by merging synonymous, redundant, or overly specific labels. The resulting frozen vocabulary is subsequently used to annotate the full dataset, followed by an additional consistency check to improve annotation reliability.

The evaluated Pythia and Qwen2.5 models are not involved in the annotation process. The same frozen vocabulary and example-level skill annotations are reused across all model scales, ensuring that the observed cross-scale trends are not induced by model-specific changes in the skill partition.

B.3 Prompt Templates

To distinguish skill structure from variation induced by prompt templates, each example is rendered using three semantically equivalent prompt templates. The templates preserve the question and candidate answers while changing the surrounding instruction format:

```
plain_qa:
Question:
{question}

Options:
{choices}

Answer:

instruction:
Please solve the following multiple-choice problem.

Problem:
{question}

Candidate answers:
{choices}

The correct answer is

exam_style:
This is an exam question. Choose the best option.

{question}

{choices}

Final answer:
```

The same three textual templates are used for every model. Each model is tokenized using its native tokenizer. In analyses involving fitting–evaluation splits, all prompt variants derived from the same underlying example are always assigned to the same split, preventing different templates of an identical question from leaking across fitting and evaluation sets.

C Experimental Details

C.1 Model Families

We evaluate two independently trained model families spanning comparable scale ranges. For Pythia (Biderman et al., 2023), we use models with 410M, 1B, 1.4B, 2.8B, 6.9B, and 12B parameters. For Qwen2.5 (Yang et al., 2024), we use 0.5B, 1.5B, 3B, 7B, 14B, and 32B models. The two families

provide distinct scaling sequences with different model designs and training pipelines, allowing us to examine whether the observed representation trends extend beyond a single model family.

All models are evaluated on the same frozen population described in Appendix B. Model parameters remain fixed throughout the representation analyses and causal interventions.

C.2 Representation Extraction

For each prompt, we extract the representation at the final prompt position before any answer token is generated. Our primary representation location is the residual-stream state immediately before the final Transformer block. This state incorporates the computation of all preceding blocks while leaving the final block and output computation available for downstream intervention.

We also extract the representation after the final Transformer block at the same prompt position. This second location allows us to examine whether the observed skill-structure trends are specific to the primary pre-final-block location or persist at the final-output location.

All representations are initially stored in their native model dimensions and are transformed only in subsequent analyses. No answer continuation is included during representation extraction.

C.3 Common-Dimensional Representation Space

Because RSS depends on representation dimensionality, directly comparing native representations across model scales could confound structural differences with differences in hidden dimension. We therefore map representations from each model to common-dimensional spaces using PCA followed by whitening.

We consider two target dimensions, $q \in \{128, 256\}$. For each model, PCA projects the native representation into the target dimension, after which whitening rescales the retained components according to their estimated variation. All cross-scale RSS comparisons are performed in these common-dimensional spaces.

For analyses involving held-out examples, PCA and whitening are estimated using $\mathcal{D}_{\mathrm{fit}}$ only. The resulting transformation is then kept fixed when mapping examples from $\mathcal{D}_{\mathrm{eval}}$.

C.4 Skill Structure After Conditioning on Task and Prompt Template

For each model, representation location, and common dimension, we compute $\mathrm{RSS}(S \mid T, D)$ as defined in Section 3. Skill structure is measured separately within each fixed task–template condition, so that variation across tasks or prompt templates does not directly contribute to the measured skill structure.

Because high-dimensional representations can exhibit nontrivial structure even under arbitrary finite-sample label partitions, we calibrate $\mathrm{RSS}(S \mid T, D)$ using the matched permutation baseline introduced in Section 3. Specifically, skill labels are permuted independently within each fixed (T, D) condition while preserving the empirical skill frequencies. The adjusted measure is

$$\mathrm{RSS}_{\mathrm{adj}}(S \mid T, D) = \mathrm{RSS}(S \mid T, D) - \mathbb{E}_{\pi}\left[\mathrm{RSS}_{\pi}(S \mid T, D)\right]. \tag{68}$$

All main cross-scale comparisons use this permutation-adjusted skill-structure measure.

C.5 Geometric Decomposition of Skill Structure

To identify the geometric source of the observed scale-dependent change in skill structure, we decompose the permutation-adjusted measure using Theorem 1. The between-skill component Φ_B captures separation among skill-conditioned representation centers relative to pooled within-skill variation, whereas Φ_{cov} captures differences in skill-conditioned covariance structure.

To account for structure induced by arbitrary finite-sample label partitions, we apply the same matched permutation calibration used for $\mathrm{RSS}(S \mid T, D)$ to both components. We therefore analyze Φ_B^{adj} and $\Phi_{\mathrm{cov}}^{\mathrm{adj}}$, defined relative to their corresponding matched permutation expectations.

C.6 Held-Out Skill Recoverability

We evaluate whether skill structure estimated from fitting examples remains recoverable on unseen examples. The underlying examples are split into 70% fitting and 30% evaluation subsets using task- and skill-stratified sampling. All prompt variants derived from the same underlying example are assigned to the same subset to prevent overlap between the fitting and evaluation sets.

The fitting subset is used to estimate PCA, whitening, and skill prototypes. These fitted quantities are then kept fixed when transforming and evaluating examples in the evaluation subset. Within each task–template condition, a held-out representation is assigned to the skill whose fitted prototype is nearest under Euclidean distance.

For an evaluation example with true skill s under condition c, we compute the prototype margin

$$M_c(z) = \min_{s' \neq s} \|z - \widehat{\mu}_{c,s'}\|_2 - \|z - \widehat{\mu}_{c,s}\|_2 . \tag{69}$$

A positive margin indicates that the correct fitted prototype is closer than every competing prototype, while a larger margin indicates stronger separation from the nearest competitor.

C.7 Causal Intervention on Skill-Aligned Representation Subspaces

Recoverability on unseen examples does not by itself establish whether the identified representation geometry contributes functionally to answer formation. We therefore directly intervene on skill-aligned representation subspaces at the pre-final-block location.

For each fitting split and task–template condition $c = (t, d)$, we estimate Euclidean skill prototypes in the same $q = 128$ PCA-whitened space used in the primary held-out recoverability analysis. Let $\widehat{\mu}_{c,s}$ denote the fitted prototype of skill s. We form condition-centered prototype displacement vectors and map their corresponding directions back to the native representation space using the fitted PCA and whitening transformation. The resulting directions are orthonormalized to obtain a basis U_c for the skill-aligned representation subspace. All quantities used to construct U_c are estimated exclusively from fitting examples.

The intervention is applied only to the representation at the final prompt position, immediately before the final Transformer block. Representations at all other token positions and all model parameters remain unchanged. The modified representation is then propagated through the remaining model computation, after which the answer candidates are rescored.

Our behavioral endpoint is the mean token log-probability of the complete gold-choice continuation. Let G_{base}, G_{skill}, and G_{rand} denote the corresponding gold-answer scores for the unmodified model, the skill-aligned intervention, and the matched random intervention, respectively. We define

$$\Delta G_{\text{skill}} = G_{\text{base}} - G_{\text{skill}}, \qquad \Delta G_{\text{rand}} = G_{\text{base}} - G_{\text{rand}}. \tag{70}$$

The causal utilization effect is

$$U_G = \Delta G_{\text{skill}} - \Delta G_{\text{rand}} = G_{\text{rand}} - G_{\text{skill}}. \tag{71}$$

Thus, $U_G > 0$ indicates that the skill-aligned intervention produces a larger reduction in the gold-answer score than a relative-energy-matched random intervention, providing evidence that the targeted skill-aligned subspace contributes functionally to answer formation.

C.8 Relative-Energy-Matched Intervention

The magnitude of a direct skill-aligned perturbation can vary substantially across examples and model scales. To control for this variation in the cross-scale causal analysis, we use a relative-energy-matched intervention whose norm is defined relative to the magnitude of each centered representation.

For a representation h under task–template condition c, define

$$\widetilde{h} = h - \mu_c, \tag{72}$$

where μ_c is estimated from fitting examples only. Its projection onto the skill-aligned representation subspace is

$$P_U(\widetilde{h}) = U_c U_c^\top \widetilde{h}. \tag{73}$$

Rather than removing a fixed fraction of the projected component, we set the perturbation norm relative to the local representation magnitude:

$$\|\Delta h_{\text{skill}}\|_2 = \eta \|\widetilde{h}\|_2. \tag{74}$$

The targeted perturbation is therefore

$$\Delta h_{\text{skill}} = \eta \|\widetilde{h}\|_2 \frac{P_U(\widetilde{h})}{\|P_U(\widetilde{h})\|_2}, \qquad h'_{\text{skill}} = h - \Delta h_{\text{skill}}. \tag{75}$$

For each targeted intervention, we construct a random control subspace R_c with the same rank as U_c and orthogonal to the skill-aligned subspace. Let

$$P_R(\widetilde{h}) = R_c R_c^\top \widetilde{h}. \tag{76}$$

The corresponding random perturbation is

$$\Delta h_{\text{rand}} = \eta \|\widetilde{h}\|_2 \frac{P_R(\widetilde{h})}{\|P_R(\widetilde{h})\|_2}, \qquad h'_{\text{rand}} = h - \Delta h_{\text{rand}}. \tag{77}$$

For retained examples satisfying the geometric support requirement, the two interventions therefore have the same per-example perturbation norm:

$$\|\Delta h_{\text{skill}}\|_2 = \|\Delta h_{\text{rand}}\|_2 = \eta \|h - \mu_c\|_2. \tag{78}$$

This matching controls for perturbation magnitude, so that the comparison between the two interventions reflects their different directions in representation space.

For cross-scale comparison, we construct a common evaluation population separately within each model family. An example–template instance is retained only if the unmodified model answers it correctly at all six evaluated scales, the prompt is not truncated at any scale, and the instance satisfies the geometric support requirement for the relative-energy intervention. This ensures that the causal effect is compared on the same set of valid examples across model scales. The retained population is fixed across intervention strengths and comparison conditions and is not conditioned on whether the fitted prototype readout correctly recovers the skill.

Finally, causal effects are averaged equally across prompt templates within each task and then equally across tasks to obtain the macro causal utilization effect. This aggregation is used for the cross-scale results in Figure 4.

D Representation-Guided Data-Efficient Adaptation Details

D.1 Training Data and Separation from the Evaluation Population

The adaptation experiments use the official training splits of ARC-Challenge, ARC-Easy, and SciQ, annotated with the same frozen skill vocabulary used throughout the main analysis. To maintain a common task–skill space, we retain only task–skill cells represented in the frozen evaluation ontology. The resulting training population is partitioned once into disjoint diagnosis and candidate subsets, and this partition is fixed across models, selection methods, adaptation budgets, and downstream evaluations.

Within each sufficiently supported task–skill cell, approximately 20% of the examples are assigned to the diagnosis subset. Whenever possible, we retain at least three diagnosis examples and five candidate examples per cell. Cells that cannot support both subsets are assigned entirely to the candidate pool and do not receive a representation-derived weakness estimate.

All diagnosis, candidate selection, and LoRA training are performed exclusively on the official training data. No example from the frozen evaluation population is used in these stages. The evaluation population is used only to define the shared task–skill ontology and for held-out evaluation. In addition, the held-out `exam_style` prompt template is excluded from both adaptation-data selection and LoRA training.

D.2 Representation-Based Skill Diagnosis

For each model, we extract pre-final-block representations from the diagnosis subset and map them to a $q = 128$ PCA-whitened space. The PCA projection and whitening transformation are fitted exclusively on the diagnosis examples.

For each skill s within task t, let $\widehat{\mu}_s$ denote its centroid in the diagnosis representation space. We define the mean within-skill squared radius per representation dimension as

$$W_s = \frac{1}{n_s q} \sum_{i:S_i=s} \|z_i - \widehat{\mu}_s\|_2^2 , \tag{79}$$

and the squared distance per dimension to the nearest competing skill centroid as

$$\Delta_s^2 = \frac{1}{q} \min_{s' \neq s} \|\widehat{\mu}_s - \widehat{\mu}_{s'}\|_2^2 . \tag{80}$$

Their ratio defines the relative separation score

$$J_s = \frac{\Delta_s^2}{W_s + \epsilon}. \tag{81}$$

A larger J_s indicates greater centroid separation relative to within-skill dispersion.

We additionally measure how reliably diagnosis examples can be recovered from their skill prototypes using a leave-one-out normalized prototype margin:

$$M_s = \frac{1}{n_s} \sum_{i:S_i=s} \frac{d_{\text{wrong}}(z_i) - d_{\text{correct}}^{\text{LOO}}(z_i)}{d_{\text{wrong}}(z_i) + d_{\text{correct}}^{\text{LOO}}(z_i) + \epsilon}. \tag{82}$$

Here, $d_{\text{correct}}^{\text{LOO}}(z_i)$ is the distance from z_i to the prototype of its correct skill estimated without example i, and $d_{\text{wrong}}(z_i)$ is the distance to the nearest competing skill prototype. The leave-one-out construction prevents an example from artificially reducing its own correct-prototype distance. Larger M_s therefore indicates more reliable prototype-based recovery.

We combine these two diagnostics into a representation-weakness score. Let $z_t(\cdot)$ denote standardization across skills within task t. We first define

$$D_s^{\text{raw}} = -\frac{1}{2} \left[z_t(\log(1 + J_s)) + z_t(M_s) \right], \tag{83}$$

and then standardize the resulting scores once more within the task:

$$D_s = z_t(D_s^{\text{raw}}) . \tag{84}$$

Consequently, larger D_s identifies skills whose representations are comparatively less separated and less reliably recoverable than other skills within the same task. Skills without sufficient diagnosis support are assigned the neutral value $D_s = 0$.

D.3 Coverage-Preserving Diversity Selection

Given a total adaptation budget B, we allocate the budget approximately equally across the three tasks. Let B_t denote the budget assigned to task t, and let K_t be the number of candidate skills represented in its candidate pool. We assign each skill a minimum reserved quota

$$m_t = \max\left(1, \left\lfloor \rho \frac{B_t}{K_t} \right\rfloor\right), \qquad \rho = 0.5, \tag{85}$$

subject to the number of available candidate examples for that skill. If a skill contains fewer than m_t candidates, all of its available candidates are retained.

The reserved set is constructed separately within each skill. When a diagnosis prototype is available, the first reserved example is chosen as the candidate nearest to that prototype. Otherwise, we use the centroid of the candidate representations for that skill as a fallback reference. Any additional reserved examples are selected by farthest-first K-center selection within the same skill, initialized from the first reserved example.

After satisfying the reserved quotas, the remaining budget for task t is allocated by global K-center selection over the remaining candidates. The reserved examples form the initial selected set, so each subsequent candidate is prioritized according to its distance from the examples already selected. This procedure preserves minimum skill coverage while using the remaining budget to increase representation diversity across the task.

D.4 Representation-Guided Acquisition and Matched Ablation

To isolate the contribution of the representation-based skill diagnosis, we construct two selectors that share the same task budgets, minimum skill-coverage constraints, reserved-example construction, and representation-diversity criterion. They differ only in whether the skill-weakness score D_s is used to reweight the remaining candidate selection.

For *Ours w/o diagnosis*, the remaining candidates are selected using the standard task-level global K-center criterion,

$$A_i^{\mathrm{w/o\,diag}} = d(z_i, \mathcal{S}), \tag{86}$$

where $\mathcal{S}$ denotes the set of examples already selected.

Our full method reweights the same diversity criterion according to the representation weakness of the candidate's skill:

$$A_i^{\mathrm{ours}} = w_{s(i)} d(z_i, \mathcal{S}), \tag{87}$$

where

$$w_s = 1 + \beta \max\{0, \tanh(D_s)\}, \qquad \beta = 0.35. \tag{88}$$

By construction,

$$1 \le w_s \le 1.35. \tag{89}$$

The reweighting is therefore bounded and one-sided. Candidates associated with representation-weak skills can receive higher priority, but only when they also provide additional representation-space coverage through $d(z_i, \mathcal{S})$. Skills with non-positive weakness scores are not downweighted. Importantly, D_s does not directly assign a fixed number of examples to each skill beyond the minimum coverage constraint; instead, it modulates candidate priority during the global K-center stage.

D.5 Baselines

We compare the full representation-guided selector with four standard data-selection baselines under the same total adaptation budget and task-level budget allocation:

- **Random**: examples are sampled uniformly from the candidate pool within each task.
- **Skill-balanced**: the task budget is distributed as evenly as possible across the available skills, followed by sampling within each skill.
- **High-loss**: candidates with the largest base-model loss are prioritized within each task.
- **K-center**: candidates are selected using task-level global K-center in the same representation space, without an explicit skill-coverage constraint or representation-based weakness diagnosis.

These baselines capture complementary selection principles: random sampling, explicit skill balancing, example difficulty, and global representation diversity. In addition, *Ours w/o diagnosis* serves as a matched ablation: it uses the same skill-coverage and representation-diversity procedure as the full method but removes the representation-derived weakness reweighting. Comparing the two therefore isolates the additional contribution of the skill-diagnosis signal beyond coverage and diversity alone.

D.6 Adaptation and Evaluation Protocol

The main-text comparison evaluates Pythia-2.8B and Qwen2.5-3B under a common adaptation budget of $B = 256$ unique underlying training examples. The broader supplementary evaluation additionally includes Pythia-1B and Qwen2.5-1.5B and considers both $B = 128$ and $B = 256$.

For each selected underlying training example, we construct adaptation instances using the `plain_qa` and `instruction` prompt templates. The `exam_style` template is excluded from adaptation and reserved for held-out evaluation. All selection methods use the same LoRA optimization settings: three epochs, per-device batch size 4, gradient accumulation over 8 steps, learning rate 2×10^{-4}, LoRA rank $r = 16$, and scaling parameter $\alpha = 32$.

After adaptation, models are evaluated on the frozen 1,497-example population using only the held-out `exam_style` template. We report equal-task Macro multiple-choice accuracy, assigning equal weight to ARC-Challenge, ARC-Easy, and SciQ regardless of differences in dataset size.

The main text focuses on the common $B = 256$ setting for Pythia-2.8B and Qwen2.5-3B to provide a compact cross-family evaluation. The goal of this experiment is not to establish a universally optimal data-selection algorithm, but to test whether the representation-derived skill diagnostics developed in the main analysis provide useful information for prioritizing adaptation examples under a limited training budget.