SKILL CO-ACTIVATION GRAPHS: A FUNCTIONAL TOPOLOGY OF GENERALIZATION IN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) exhibit strong generalization, yet it remains unclear how reusable predictive mechanisms are organized internally to support this behavior. We introduce the *Skill Co-activation Graphs* (SCGs), a functional-topology view in which nodes are skill-indexed components induced from token-level loss-gradient alignment and edges encode their empirical co-activation across inputs. Starting from the autoregressive objective, we derive a loss-induced skill-alignment formulation and establish a task-conditioned generalization bound whose complexity depends on $\mathrm{Tr}(L^{+}\Sigma_T)$, where L^{+} captures SCG resistance and Σ_T captures task-induced skill usage. We further extend this analysis to empirically estimated SCGs through a stability-aware graph-estimation bound. Experiments on Pythia with ARC-Challenge and SciQ reveal systematic relationships between SCG structure and downstream generalization. Larger Pythia models tend to exhibit richer and more stable skill organization, while lower held-out task-conditioned resistance is associated with higher task accuracy. Intervention-based analyses support the functional relevance of the induced skill subspaces. Across both Pythia and Qwen2.5, SCG-guided fine-tuning further shows that high-resistance examples provide a useful structural signal for data selection under limited adaptation budgets.

1 INTRODUCTION

Large language models (LLMs) achieve strong performance across linguistic and reasoning tasks (Grattafiori et al., 2024; Achiam et al., 2023), yet the structural mechanisms underlying their generalization remain insufficiently understood. Existing interpretability studies have revealed important local mechanisms, such as attention heads (Chen et al., 2024), circuits (Marks et al., 2025), and sparse features (Shu et al., 2025). However, these analyses often focus on microscopic units and provide limited tools for characterizing recurring functional patterns at an intermediate scale and relating them to downstream generalization. We refer to such recurring prediction-related patterns as *reusable predictive mechanisms*: not predefined task labels or explicit modules, but post-hoc functional regularities reflected in the model's predictive geometry. In this work, we instantiate this idea as *skill-indexed components*, grouping token predictions according to the directional similarity of their loss gradients with respect to model parameters. This perspective shifts the focus from isolated internal features or externally defined abilities to the model-internal structure of predictive mechanisms, providing a basis for relating their empirical co-activation patterns to downstream generalization.

Recent work on skill composition suggests that complex language tasks can involve combinations of simpler skills (Zhao et al., 2024). For example, predicting the final token in "The city councilmen refused the demonstrators a permit because they feared [MASK]." requires signals from entity roles, coreference, event structure, and causal relations. This motivates modeling not only individual skill-indexed components, but also their empirical co-activation relations across contexts. We represent these relations with a *Skill Co-activation Graph* (SCG), where nodes are induced skill-indexed components and edges encode their co-activation across textual inputs. Under this view, each input induces an activation pattern over multiple reusable components, and the SCG provides a functional topology for characterizing their co-activation structure and its relation to generalization.

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.

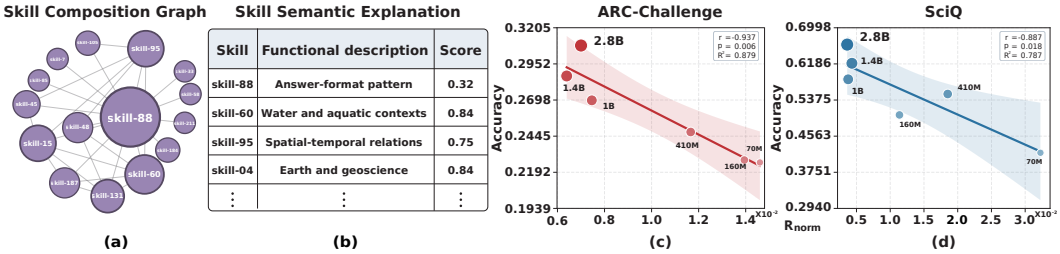

Figure 1: **Empirical SCG structure, skill interpretation, and task-conditioned resistance across Pythia models.** (a) Partial visualization of the empirical SCG constructed on ARC-Challenge for the Pythia analysis, where nodes denote induced skills and edges denote their empirical co-activation relations. Only representative nodes are shown for readability. (b) Post-hoc semantic interpretations of selected skill nodes in (a), generated using GPT-5.6 from their associated token sets. (c) Relationship between normalized task-conditioned resistance $R_{\text{norm}} = \widehat{R}^{\text{norm}}_{\text{task}}$ and held-out accuracy across Pythia model scales on ARC-Challenge. (d) The corresponding relationship across Pythia scales on SciQ. In (c)–(d), each point denotes one model scale, the solid line shows the linear regression fit, and the shaded region denotes the 95% confidence interval. The negative correlations indicate that lower task-conditioned resistance is associated with higher held-out accuracy across Pythia scales.

Theoretically, we connect this functional-topology view to autoregressive learning. Starting from the next-token prediction objective, we derive a skill-alignment term between an input-induced skill attribution pattern and a global skill contribution vector. This links the induced components to the language modeling objective. We then impose graph smoothness over the SCG and obtain a task-conditioned generalization bound involving $\mathrm{Tr}(L^{+}\Sigma_T)$, where L^{+} is the pseudoinverse of the SCG Laplacian and Σ_T is the task-induced skill covariance. This term characterizes how task-induced skill demands align with low-resistance modes of the induced functional topology. Since the SCG is estimated from gradient-based skill induction, we further incorporate a stability-aware graph estimation term, linking the reliability of the estimated-SCG bound to the stability of the induced skill partition.

Empirically, we evaluate our framework on two model families, Pythia (Biderman et al., 2023) and Qwen2.5 (Hui et al., 2024), using two reasoning benchmarks, ARC-Challenge (Clark et al., 2018) and SciQ (Welbl et al., 2017). To prevent frequent tokens from dominating gradient-based skill induction and obscuring functional structure associated with more informative predictions, we stratify tokens into multiple loss-difficulty levels and use the High-loss tokens for skill-indexed component induction in the main experiments. In the main Pythia scaling analysis, larger models generally exhibit richer and more stable skill structure, together with lower held-out task-conditioned resistance. Moreover, lower resistance is strongly associated with higher held-out accuracy, suggesting that scaling reshapes the task-relevant functional topology rather than merely increasing the number of induced components.

To examine whether the induced skills are meaningful and practically useful, we provide semantic, intervention-based, and adaptation evidence. We interpret the induced skills and ablate task-relevant skill subspaces, which causes substantially larger gold-answer NLL degradation than same-rank random-subspace controls in most settings. We further use SCG resistance for fine-tuning data selection, where high-resistance examples yield accuracy gains under limited fine-tuning budgets across the evaluated model–task settings. Our contributions are summarized as follows:

- We propose a functional-topology view of LLM generalization, using empirical skill co-activation structure to characterize relationships among internally induced functional components beyond isolated units.
- We connect autoregressive learning to skill topology by deriving a loss-induced skill-alignment term and task-conditioned generalization bound governed by $\mathrm{Tr}(L^{+}\Sigma_T)$.
- We develop a stability-aware framework for inducing skill-indexed components from token-level gradient geometry and constructing SCGs from their co-activation structure.
- We show on Pythia that SCG structure is associated with held-out task performance, while experiments across Pythia and Qwen2.5 support the functional relevance of skill subspaces and the use of SCG resistance as a data-selection signal under limited fine-tuning budgets.

Overall, our results suggest that scaling is associated not only with richer induced skill structure, but also with changes in functional topology, with lower task-conditioned resistance tending to accompany stronger held-out performance.

2 Related Work

2.1 Mechanistic Interpretability and Feature Discovery

Prior work in mechanistic interpretability has identified a range of internal structures in large language models, including attention heads (Chen et al., 2024; Nanda et al., 2023; Olsson et al., 2022), circuits (Marks et al., 2025), neurons (Bai & Ma, 2025), activation directions (Gao et al., 2025; Rimsky et al., 2024), and sparse features (Shu et al., 2025). These studies provide fine-grained explanations of specific model behaviors, such as induction, factual recall, feature superposition, or localized causal mechanisms. However, many operate at a microscopic level and do not directly characterize how a large collection of reusable predictive mechanisms is organized into a functional topology relevant to generalization.

Our work studies an intermediate level of abstraction. We introduce *skill-indexed components* as coherent predictive patterns induced from token-level gradient geometry. These components lie at a higher level of abstraction than individual neurons or heads, while remaining grounded in the model's internal computation rather than externally defined task labels. In this sense, our framework complements existing feature-discovery methods by moving from isolated interpretable units to the co-activation topology of functional components.

2.2 Skills, Compositionality, and Gradient-Based Functional Units

A growing body of work studies language model capabilities through benchmark taxonomies and emergent ability analyses (Berti et al., 2025; Ni et al., 2025; Srivastava et al., 2023; Bommasani et al., 2023), as well as through skills and reasoning primitives (Lippl et al., 2025; Zhao et al., 2024; Thomm et al., 2024). Other studies investigate complex-skill emergence, skill composition, compositional generalization, and LLM-extracted skill graphs for instruction tuning and data selection (Li et al., 2025a; Kaur et al., 2025; Didolkar et al., 2024; Zhao et al., 2024; Arora & Goyal, 2023). Together, these works motivate analyzing complex behavior in terms of reusable abilities and their interactions across contexts. Most existing approaches, however, define skills behaviorally, using benchmark categories, task labels, or LLM-extracted skill descriptions.

In contrast to behaviorally defined or LLM-extracted skills, we induce skill-indexed components from the model's internal predictive geometry. Inspired by gradient-based analyses of functional structure (Michaud et al., 2023), we use token-level loss gradients as local signatures of parameter-space sensitivity and group tokens with similar gradient directions as relying on related predictive mechanisms. Rather than studying these components in isolation, we construct a SCG over them and analyze how its co-activation topology relates to task-conditioned generalization.

2.3 Theoretical Analyses of Next-Token Prediction

Recent theory has studied next-token prediction (NTP) from several perspectives, including autoregressive expressivity, optimization bias, self-attention mechanisms, memory capacity, and Rademacher-style generalization bounds that depend on data size, sequence length, and model parameters (Li et al., 2025b; Flemings et al., 2024; Madden et al., 2024; Li et al., 2024; Malach, 2023). These works provide important token-, architecture-, and parameter-level perspectives on NTP.

Our analysis differs by introducing topology into the generalization complexity. Rather than depending only on parameter- or data-level quantities, our bound contains the topology-aware term $\mathrm{Tr}(L^{+}\Sigma_T)$, where L^{+} captures the resistance structure of the SCG and Σ_T captures task-induced skill covariance. This makes the bound task-conditioned: it characterizes how task-induced skill demands align with low-resistance modes of the model's internal functional topology.

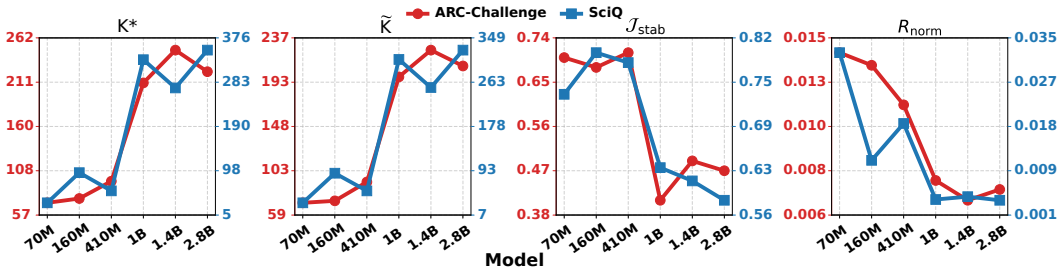

Figure 2: **Held-out task-conditioned SCG diagnostics for Pythia models on ARC-Challenge and SciQ.** For each Pythia model, the SCG is constructed from High-loss training tokens and evaluated on all held-out test examples. The four panels show $K^\star$, $\widetilde{K}$, $\mathcal{J}_{\text{stab}}$, and R_{norm}, respectively.

2.4 Graph-Based Generalization and Functional Topology

Graph-based learning theory uses Laplacian smoothness, spectral complexity, and effective resistance to analyze generalization over structured domains (Zhang & Amini, 2024; Belkin et al., 2004; Smola & Kondor, 2003). In most prior settings, the graph is defined over data points, labels, tasks, or external dependency structures. Our SCG differs: its nodes are internally induced skill-indexed components, and its edges encode their empirical co-activation relations across inputs.

This distinction leads to a task-conditioned view of generalization. Rather than relying on a topology-level quantity such as $\mathrm{Tr}(L^+)$, our analysis involves $\mathrm{Tr}(L^+\Sigma_T)$, where L^+ captures the resistance geometry of the SCG and Σ_T captures task-induced skill covariance. This term characterizes how task-induced skill demands align with low-resistance modes of functional topology.

3 Task-Conditioned Skill Topology and Generalization

This section develops a task-conditioned generalization theory through the SCG framework, linking skill-demand–topology alignment to generalization while accounting for empirically estimated graphs. The theory consists of four parts. First, we derive a centered skill-alignment score from the autoregressive language modeling loss. Second, we show that the generalization complexity of graph-smooth skill-alignment functions is governed by a task-conditioned resistance term. Third, we extend the analysis to the practical setting where the SCG is constructed from gradient-induced skill partitions. Finally, we derive a bound-aware rule for selecting the number of skill clusters.

3.1 Loss-Induced Skill Alignment Decomposition

We begin by connecting autoregressive prediction to skill alignment. Let $x = (x_{<t}, x_t)$ denote a next-token prediction instance, where $x_{<t}$ is the context and x_t is the target token. Consider a finite set of skill-indexed components $\mathcal{S} = \{s_1, \ldots, s_K\}$. For a fixed context $x_{<t}$, let $\pi_k(x_{<t})$ denote the context-dependent attribution weight of the k-th skill component, with $\pi_k(x_{<t}) \geq 0$ and $\sum_{k=1}^K \pi_k(x_{<t}) = 1$. Let $p_k(x_t \mid x_{<t})$ denote the corresponding skill-conditioned predictive likelihood. We use the following analysis-level decomposition: $P_\theta(x_t \mid x_{<t}) = \sum_{k=1}^K \pi_k(x_{<t}) p_k(x_t \mid x_{<t})$, which should be interpreted as an attributional decomposition of predictive likelihood, rather than as an assumption that the model explicitly samples a discrete latent skill. Let $\Phi(x) = (\Phi_1(x), \ldots, \Phi_K(x)) \in \Delta^{K-1}$ be an auxiliary skill-attribution distribution, where $\Phi_k(x) > 0$ and $\sum_{k=1}^K \Phi_k(x) = 1$. This distribution is introduced for variational analysis and need not coincide with $\pi(x_{<t})$. Let $\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^K$, define $\Pi = I_K - K^{-1}\mathbf{1}\mathbf{1}^\top$, and let $z(x) = \Pi\Phi(x)$ be the centered skill-attribution vector.

Theorem 1 (Loss-Induced Skill Alignment Decomposition)**.** *For any prediction instance $x = (x_{<t}, x_t)$, the autoregressive negative log-likelihood $\mathcal{L}_{\text{AR}}(x) = -\log P_\theta(x_t \mid x_{<t})$ satisfies $\mathcal{L}_{\text{AR}}(x) \leq -\sum_{k=1}^K \Phi_k(x) \log p_k(x_t \mid x_{<t}) + D_{\text{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big)$.*

Furthermore, let $\bar{a} \in \mathbb{R}^K$ be a fixed reference coefficient vector and suppose the skill-conditioned log-likelihood is decomposed as $\log p_k(x_t \mid x_{<t}) = c(x) + \beta\bar{a}_k + \bar{r}_k(x)$, where $c(x)$ is a sample-

dependent baseline, $\beta > 0$, and $\bar{r}_k(x)$ is the corresponding residual term. Then

$$\mathcal{L}_{\mathrm{AR}}(x) \le C(x) - \beta \bar{a}^\top \Phi(x) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big) + \bar{\varepsilon}_r(x), \tag{1}$$

where $C(x) = -c(x)$ and $\bar{\varepsilon}_r(x) = -\sum_{k=1}^{K} \Phi_k(x)\bar{r}_k(x)$. If $\bar{a} = \bar{a}_0\mathbf{1} + \bar{a}_\perp$, with $\bar{a}_0 = K^{-1}\mathbf{1}^\top \bar{a}$ and $\bar{a}_\perp \in \mathbf{1}^\perp$, then $\bar{a}^\top \Phi(x) = \bar{a}_0 + \bar{a}_\perp^\top z(x)$. Hence, after absorbing the input-independent constant into the baseline, the leading skill-dependent term in the loss upper bound is $-\beta \bar{a}_\perp^\top z(x)$.

Theorem 1 shows that a centered linear alignment term appears explicitly in an upper-bound decomposition of autoregressive loss, rather than being introduced independently of the prediction objective. Motivated by this form, we consider the skill-alignment functions $h_a(x) = a^\top z(x)$, with $a \in \mathbf{1}^\perp$, and analyze their generalization under the SCG topology.

3.2 Task-Conditioned SCG Generalization Bound

We now place the loss-induced skill-alignment form on a functional topology. The SCG is constructed over model-internal skill-indexed components induced from predictive behavior. For brevity, we refer to these induced components as skills when no ambiguity arises.

Let $\mathcal{C}$ denote the corpus used for graph construction, where each $X \in \mathcal{C}$ is a graph-construction example, typically a token sequence. For each induced skill s_k, define its binary usage indicator as $\psi_k(X) = \mathbb{I}\{X \text{ contains at least one token assigned to } s_k\}$. For two distinct skills s_i and s_j, define their raw co-activation count by $N_{ij} = \sum_{X \in \mathcal{C}} \psi_i(X)\psi_j(X)$ for $i < j$. In the main experiments, we construct a binary co-activation SCG by setting $W_{ij} = W_{ji} = \mathbb{I}\{N_{ij} > 0\}$ for $i \neq j$, and $W_{ii} = 0$. Thus, an edge records that two induced skills are co-activated in at least one graph-construction example. This binary construction captures the existence of empirical co-activation relations rather than their frequencies.

Let $G = (V, E, W)$ be the resulting unweighted SCG. Let $D = \mathrm{diag}(d_1, \ldots, d_K)$, with $d_i = \sum_{j=1}^{K} W_{ij}$, and define the graph Laplacian $L = D - W$. We assume that G is connected, so $0 = \mu_1 < \mu_2 \le \cdots \le \mu_K$ and $\mathrm{Null}(L) = \mathrm{span}\{\mathbf{1}\}$. On $\mathbf{1}^\perp$, L is positive definite. Its Moore–Penrose pseudoinverse is denoted by L^+; equivalently, if $L = \sum_{r=2}^{K} \mu_r u_r u_r^\top$ on $\mathbf{1}^\perp$, then $L^+ = \sum_{r=2}^{K} \mu_r^{-1} u_r u_r^\top$. Motivated by Theorem 1, we consider alignment functions $h_a(x) = a^\top z(x)$ with $a \in \mathbf{1}^\perp$. The centering constraint removes the constant direction, which is not controlled by the Laplacian since $L\mathbf{1} = 0$. We define the graph-smooth alignment class and the empirical task-conditioned skill second-moment matrix as $\mathcal{H}_G(\lambda) = \{h_a(x) = a^\top z(x) : a \in \mathbf{1}^\perp,\ a^\top L a \le \lambda\}$, $\Sigma_S = \frac{1}{n}\sum_{i=1}^{n} z_i z_i^\top$, where $S = \{x_i\}_{i=1}^{n}$ is an i.i.d. sample and $z_i = z(x_i)$. Since $z_i = \Pi\Phi(x_i)$ and $\Phi(x_i) \in \Delta^{K-1}$, we have $\|z_i\|_2 \le 1$, and hence $\mathrm{Tr}(\Sigma_S) \le 1$. The proof is deferred to Appendix A.3.

Lemma 1 (Rademacher Complexity of the Graph-Smooth Alignment Class)**.** *For $\mathcal{H}_G(\lambda)$, the empirical Rademacher complexity satisfies $\widehat{\mathfrak{R}}_S(\mathcal{H}_G(\lambda)) \le \sqrt{\frac{\lambda}{n}\,\mathrm{Tr}(L^+\Sigma_S)}$.*

Lemma 1 shows that the complexity of graph-smooth alignment functions is governed by the interaction between the SCG resistance operator L^+ and the task-induced skill second-moment matrix Σ_S.

To obtain a bounded loss class, we keep the residual term in Theorem 1 fixed at the reference coefficient $\bar{a}$ and define $q(x) = D_{\mathrm{KL}}(\Phi(x) \,\|\, \pi(x_{<t})) + \bar{\varepsilon}_r(x)$. Thus, $q(x)$ is fixed with respect to the coefficient a varied over the hypothesis class. For $\tau > 0$, $\beta > 0$, and $M > 0$, let $\rho(u) = \min\{M, \max\{0, u\}\}$, and define $\ell_a(x) = \rho\big(\tau - \beta a^\top z(x) + q(x)\big)$, $\mathcal{L}_G(\lambda) = \{\ell_a : a \in \mathbf{1}^\perp,\ a^\top L a \le \lambda\}$. Then $0 \le \ell_a(x) \le M$.

Lemma 2 (Rademacher Complexity of the Skill-Alignment Loss Class)**.** *The empirical Rademacher complexity of $\mathcal{L}_G(\lambda)$ satisfies $\widehat{\mathfrak{R}}_S\big(\mathcal{L}_G(\lambda)\big) \le \beta\sqrt{\frac{\lambda}{n}\,\mathrm{Tr}(L^+\Sigma_S)}$.*

Theorem 2 (Task-Conditioned SCG Generalization Bound)**.** *Assume $0 \le \ell_a(x) \le M$ for all x and all $a \in \mathbf{1}^\perp$ satisfying $a^\top L a \le \lambda$. Then, for any $\delta > 0$, with probability at least $1 - \delta$ over*

$S = \{x_i\}_{i=1}^n$, *the following bound holds uniformly for all* $a \in \mathbf{1}^\perp$ *with* $a^\top L a \le \lambda$*:*

$$\mathbb{E}_{x\sim\mathcal{D}}[\ell_a(x)] \le \frac{1}{n}\sum_{i=1}^{n}\ell_a(x_i) + 2\beta\sqrt{\frac{\lambda}{n}\operatorname{Tr}(L^+\Sigma_S)} + 3M\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{2}$$

Theorem 2 identifies $\operatorname{Tr}(L^+\Sigma_S)$ as the task-conditioned resistance term governing the topology-dependent complexity of the skill-alignment class. Using the spectral decomposition of L, this term can be written as $\operatorname{Tr}(L^+\Sigma_S) = \sum_{r=2}^{K} \frac{u_r^\top \Sigma_S u_r}{\mu_r}$. Here $u_r^\top \Sigma_S u_r$ measures task-induced skill demand along the r-th graph spectral mode, while $1/\mu_r$ gives its inverse-connectivity weight. Thus, the bound is looser when task-induced skill demand concentrates on weakly connected, high-resistance modes, and tighter when it concentrates on well-integrated, low-resistance modes.

Corollary 1 (Topology-Only Bound as a Coarse Upper Bound)**.** *Since* $\operatorname{Tr}(\Sigma_S) \le 1$*, the task-conditioned complexity is upper bounded by the topology-only complexity:* $\operatorname{Tr}(L^+\Sigma_S) \le \operatorname{Tr}(L^+)$.

3.3 Stability-Aware Learned-SCG Generalization Bound

The preceding analysis assumes that the SCG is fixed. In practice, however, the SCG is constructed from gradient-induced skill partitions and empirical co-activation statistics. We now show how uncertainty from empirical graph estimation enters the generalization bound.

Let $\mathcal{K}$ be a finite set of candidate cluster numbers. For each $K \in \mathcal{K}$, we perform B repeated subsampling runs and let $\mathcal{P}_K^{(b)}$ denote the partition obtained from the b-th subsampled token set. For two runs $b \ne c$, let $\mathrm{ARI}_{b,c}(K)$ be the Adjusted Rand Index computed on the intersection of the corresponding subsampled token sets. Define $\overline{\mathrm{ARI}}(K) = \frac{2}{B(B-1)}\sum_{1\le b<c\le B}\mathrm{ARI}_{b,c}(K)$, and $\mathrm{Var}_{\mathrm{ARI}}(K) = \frac{2}{B(B-1)}\sum_{1\le b<c\le B}\left(\mathrm{ARI}_{b,c}(K) - \overline{\mathrm{ARI}}(K)\right)^2$. The stability error score is $s(K) = 1 - \overline{\mathrm{ARI}}(K) + \alpha\sqrt{\mathrm{Var}_{\mathrm{ARI}}(K)} + \eta\Omega(K)$, where $\alpha, \eta > 0$, and $\Omega(K)$ penalizes excessive cluster fragmentation. After computing $s(K)$, we rerun clustering on the full token set under the same candidate K, obtain the full-data partition $\widehat{\mathcal{P}}_K$, and construct the empirical SCG $\widehat{G}_K$ with Laplacian $\widehat{L}_K$. The subsampled partitions are used only to estimate stability, whereas $\widehat{L}_K$ is the final empirical Laplacian entering the bound.

Assumption 1 (Stability-to-Graph Perturbation)**.** *For each* $K \in \mathcal{K}$*, let* $L_K^\star$ *denote the latent population SCG Laplacian corresponding to the ideal* K*-component skill structure, and let* $\widehat{L}_K$ *denote the full-data empirical SCG Laplacian. There exists a constant* $C_L > 0$ *such that, with probability at least* $1 - \delta_G$*,* $\|\widehat{L}_K - L_K^\star\|_{\mathrm{op}} \le C_L s(K), \qquad \forall K \in \mathcal{K}.$

Assumption 1 formalizes clustering stability as an observable proxy for graph estimation error. It does not equate ARI stability with Laplacian perturbation; rather, under a fixed SCG construction pipeline, it links instability of the induced skill partition to instability of the resulting empirical graph. For compactness, define $\varepsilon_K = C_L s(K)$, and let $m_K = \mu_2(L_K^\star)$ be the algebraic connectivity of the latent population SCG. We assume $0 < \varepsilon_K < m_K$.

Lemma 3 (Lower Bound on the Learned Graph Connectivity)**.** *Under Assumption 1, for every* $K \in \mathcal{K}$*,* $\mu_2(\widehat{L}_K) \ge m_K - \varepsilon_K$.

Lemma 4 (Pseudoinverse Perturbation)**.** *Suppose* $L_K^\star$ *and* $\widehat{L}_K$ *are connected graph Laplacians with the same nullspace* $\mathrm{span}\{\mathbf{1}\}$*. If* $\|\widehat{L}_K - L_K^\star\|_{\mathrm{op}} \le \varepsilon_K < m_K$*, then* $\|\widehat{L}_K^+ - L_K^{\star+}\|_{\mathrm{op}} \le \frac{\varepsilon_K}{m_K(m_K-\varepsilon_K)}$.

Lemma 5 (Transfer of Empirical Smoothness to Population Smoothness)**.** *Suppose* $a \in \mathbf{1}^\perp$ *satisfies* $a^\top \widehat{L}_K a \le \lambda$*. Then* $a^\top L_K^\star a \le \lambda_K, \qquad \lambda_K := \lambda\frac{m_K}{m_K-\varepsilon_K}$.

Lemmas 3–5 show, under the stability assumption, how graph-estimation error affects the connectivity and pseudoinverse of the empirical SCG. They also quantify the cost of transferring a smoothness constraint from the empirical graph to the latent population graph.

Next, let $\Sigma_{S,K}^\star$ and $\widehat{\Sigma}_{S,K}$ denote the ideal and empirical task-conditioned skill covariance matrices, defined as $\Sigma_{S,K}^\star = \frac{1}{n}\sum_{i=1}^n z_{i,K}^\star z_{i,K}^{\star\top}$ and $\widehat{\Sigma}_{S,K} = \frac{1}{n}\sum_{i=1}^n \widehat{z}_{i,K}\widehat{z}_{i,K}^\top$, respectively. We measure their estimation discrepancy by $\varepsilon_{\Sigma,K}^{(1)} = \|\Sigma_{S,K}^\star - \widehat{\Sigma}_{S,K}\|_*$, where $\|\cdot\|_*$ denotes the nuclear norm.

Lemma 6 (Perturbation of Task-Conditioned Resistance)**.** *Assume* $\mathrm{Tr}(\Sigma^{\star}_{S,K}) \le 1$. *Then*

$$\mathrm{Tr}(L_K^{\star+}\Sigma^{\star}_{S,K}) \le \mathrm{Tr}(\widehat{L}_K^{+}\widehat{\Sigma}_{S,K}) + \Delta_K, \qquad \Delta_K = \frac{\varepsilon_K}{m_K(m_K-\varepsilon_K)} + \frac{\varepsilon^{(1)}_{\Sigma,K}}{m_K-\varepsilon_K}. \tag{3}$$

Theorem 3 (Stability-Aware Learned-SCG Generalization Bound)**.** *Let $\mathcal{K}$ be a finite candidate set. Suppose Assumption 1 holds with probability at least $1-\delta_G$. For each $K \in \mathcal{K}$, let $\widehat{L}_K$ be the full-data empirical SCG Laplacian and let $\widehat{\Sigma}_{S,K}$ be the empirical task-conditioned skill covariance induced by the learned skill representation. Assume $0 \le \ell_a(x) \le M$. Then, with probability at least $1-\delta-\delta_G$, the following bound holds simultaneously for all $K \in \mathcal{K}$ and all $a \in \mathbf{1}^{\perp}$ satisfying $a^{\top}\widehat{L}_K a \le \lambda$:*

$$\mathbb{E}[\ell_a(x)] \le \frac{1}{n}\sum_{i=1}^{n}\ell_a(x_i) + 2\beta\sqrt{\frac{\lambda_K}{n}\left[\mathrm{Tr}(\widehat{L}_K^{+}\widehat{\Sigma}_{S,K}) + \Delta_K\right]} + 3M\sqrt{\frac{\log(2|\mathcal{K}|/\delta)}{2n}}. \tag{4}$$

Here $\lambda_K = \lambda m_K/(m_K-\varepsilon_K)$ and $\Delta_K = \varepsilon_K/[m_K(m_K-\varepsilon_K)] + \varepsilon^{(1)}_{\Sigma,K}/(m_K-\varepsilon_K)$. Equivalently, since $\varepsilon_K = C_L s(K)$, we have $\Delta_K = C_L s(K)/[m_K(m_K - C_L s(K))] + \varepsilon^{(1)}_{\Sigma,K}/(m_K - C_L s(K))$.

Theorem 3 extends the fixed-SCG bound to the empirical setting, where the graph is estimated from finite samples. The empirical term $\mathrm{Tr}(\widehat{L}_K^{+}\widehat{\Sigma}_{S,K})$ is the task-conditioned resistance of the estimated SCG, while Δ_K captures additional uncertainty from graph and skill-covariance estimation. The factor $\lambda_K = \lambda m_K/(m_K - \varepsilon_K)$ reflects that a function that is λ-smooth on the empirical SCG may only be λ_K-smooth on the latent population SCG. The logarithmic factor $\log|\mathcal{K}|$ accounts for selecting the cluster number from a finite candidate set.

These bounds are intended as diagnostic tools rather than objectives for optimizing the underlying language model. Theorem 2 identifies $\mathrm{Tr}(L^{+}\Sigma_S)$ as the fixed-SCG complexity term for graph-smooth skill-alignment functions. Since the SCG is estimated from finite samples, Theorem 3 motivates the empirical diagnostic $\mathrm{Tr}(\widehat{L}_K^{+}\widehat{\Sigma}_{S,K})$, together with the graph-estimation penalty Δ_K. Thus, empirical SCG resistance characterizes the topology-dependent term appearing in the bound, while clustering stability assesses the reliability of the constructed SCG.

3.4 Empirical SCG-Based Generalization Diagnostics

The preceding bounds motivate diagnostics for analyzing the generalization behavior of language models through empirical SCGs. Theorem 2 identifies the task-conditioned resistance $\mathrm{Tr}(L^{+}\Sigma_S)$ as the key topology-dependent complexity term for a fixed SCG. Since the SCG is constructed from finite samples in practice, Theorem 3 shows that its reliability depends on graph-estimation stability. Accordingly, we first select the cluster number $K^{\star}$ using the stability criterion and fix the SCG before held-out evaluation. We then compute $\widehat{R}_{\mathrm{task}}(M,T;K^{\star}) = \mathrm{Tr}\left(\widehat{L}^{+}_{M,K^{\star}}\widehat{\Sigma}_{T,M,K^{\star}}\right)$ as the empirical task-conditioned diagnostic. Here $\widehat{L}_{M,K^{\star}}$ is the selected empirical SCG Laplacian of model M, and $\widehat{\Sigma}_{T,M,K^{\star}}$ is the empirical task-induced skill second-moment matrix for task T. For the fixed SCG, a smaller value indicates that task-induced skill demands concentrate more on low-resistance modes and reduces the corresponding empirical complexity term in the bound. Thus, clustering stability is used to select and assess the reliability of the empirical SCG, while task-conditioned resistance serves as a post-selection diagnostic of its task-dependent topology.

4 Experiments

In this section, we empirically examine whether the proposed SCG framework provides meaningful diagnostics of task-relevant functional topology in LLMs. Detailed token-level loss analyses that motivate loss-stratified skill induction are deferred to Appendix C.1; the main experiments focus on scale-dependent SCG diagnostics, intervention-based functional validation, and SCG-guided data selection under limited fine-tuning budgets.

Table 1: **Functional validation of induced skill subspaces across Pythia model scales on ARC-Challenge and SciQ.** We report the increase in per-token gold-answer NLL (ΔNLL$\uparrow$) after ablating task-relevant skill subspaces (Task) or same-rank random subspaces (RandSub).

Dataset	Intervention	70M	160M	410M	1B	1.4B	2.8B
ARC-Challenge	Task	22.822	23.773	1.936	1.664	1.806	1.541
	RandSub	1.238	0.224	0.070	−0.001	0.004	0.049
SciQ	Task	18.999	17.503	1.158	0.535	0.376	−0.352
	RandSub	1.697	−0.033	0.091	0.084	−0.013	0.084

4.1 Experimental Setup

We evaluate the proposed SCG framework on two model families, Pythia (Biderman et al., 2023) and Qwen2.5 (Hui et al., 2024), using two reasoning benchmarks: ARC-Challenge (Clark et al., 2018) and SciQ (Welbl et al., 2017). We consider multiple model scales within each family to examine how the induced skill topology varies with model size. Unless otherwise specified, token-level losses and gradients are computed autoregressively, with gradients taken with respect to the last Transformer block. We stratify tokens by token-level loss and construct the main SCGs from High-loss tokens in the training split. The induced skill partition, skill vocabulary, and SCG are fixed before evaluation; held-out test examples are mapped onto this frozen train-induced skill vocabulary without reconstructing the graph. For structural analysis, we report the stability-selected cluster number $K^\star$, the number of retained SCG nodes $\widetilde{K}$, clustering stability J_{stab}, and normalized task-conditioned resistance R_{norm}. For task performance, we use length-normalized multiple-choice accuracy, while intervention effects are evaluated using gold-answer NLL. Detailed construction, intervention, and fine-tuning protocols are provided in Appendices B and C.

4.2 Empirical SCG Diagnostics and Task Generalization

We next test whether empirical SCGs exhibit the task-conditioned properties motivated by our theory under held-out evaluation. For each model and dataset, we select $K^\star$ using the stability criterion on High-loss training tokens, filter semantically uninformative clusters, and construct the SCG over the remaining skills. The resulting skill vocabulary and graph Laplacian are then fixed. Held-out examples are mapped onto this train-induced vocabulary, without reconstructing the graph, to compute task-conditioned resistance.

Figure 1(a) shows a partial empirical SCG, whose nodes denote induced skills and edges their empirical co-activation relations. Figure 1(b) provides semantic interpretations of selected skills using the procedure in Section B.3. Figure 2 summarizes held-out diagnostics across Pythia scales on ARC-Challenge and SciQ. Across both datasets, $K^\star$ and $\widetilde{K}$ generally increase toward billion-scale models, while $\mathcal{J}_{\mathrm{stab}}$ decreases and R_{norm} is lower than in smaller models. Although not strictly monotonic, these trends indicate richer and more stable induced skill structure with lower held-out task-conditioned resistance. Corresponding SCG diagnostics for Qwen2.5 are reported in Appendix C.2.

Figure 1(c)–(d) relates this resistance to task performance. Across six Pythia scales, R_{norm} is strongly negatively correlated with held-out accuracy on ARC-Challenge ($r = -0.937$) and SciQ ($r = -0.887$). Thus, lower-resistance SCGs tend to co-occur with stronger held-out performance, consistent with Theorem 3, where lower task-conditioned resistance yields a tighter bound.

4.3 Functional Validation of Induced Skills

The preceding SCG diagnostics are correlational. To test whether the induced skills are functionally relevant, we perform subspace interventions and compare task-relevant skill directions with same-rank random representation subspaces. Table 1 reports the increase in per-token gold-answer NLL after intervention. On ARC-Challenge, task-relevant skill ablation causes substantially larger degradation than RandSub at every Pythia scale. A similar pattern holds on SciQ for five of the six models. Overall, the expected separation is observed in 11 of 12 model–task settings, supporting the interpretation of the induced skills as functionally relevant rather than arbitrary representation directions. The unusually large effects in the smallest models should not be interpreted as stronger skill

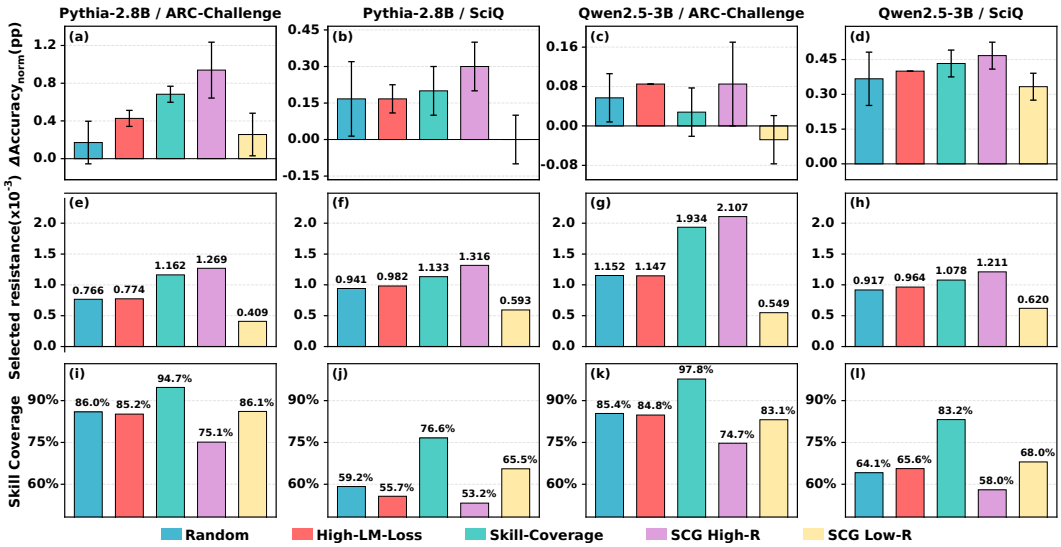

Figure 3: **SCG-guided fine-tuning on ARC-Challenge and SciQ.** We compare five data-selection strategies under the same subset size and fine-tuning budget for Pythia-2.8B and Qwen2.5-3B. Panels (a)–(d) report the improvement in normalized accuracy ($\Delta\text{Accuracy}_{\text{norm}}$, in percentage points), with means and standard deviations over three seeds. Panels (e)–(h) show the mean resistance of the selected examples, and panels (i)–(l) show their skill coverage. Resistance is computed on the fixed train-induced SCG, while skill coverage denotes the fraction of train-induced skills activated by the selected examples.

localization, but may instead reflect greater entanglement among functional directions. Additional Qwen2.5 results are provided in Appendix C.3.

4.4 SCG-Guided Fine-Tuning Data Selection

We further evaluate whether SCG structure can provide a data-selection signal for fine-tuning. We consider Pythia-2.8B and Qwen2.5-3B on ARC-Challenge and SciQ, and compare five selection strategies: Random, High-LM-Loss, Skill-Coverage, SCG High-R, and SCG Low-R. For each model–task setting, a fixed-size subset is selected from a candidate pool disjoint from the data used to construct the SCG, and all methods use the same fine-tuning budget within that setting. We report the change in normalized held-out accuracy as the primary metric, averaged over three random seeds. Detailed experimental design is provided in Appendix C.4.

Figure 3 shows that SCG resistance provides a useful selection signal across both model families and datasets. SCG High-R yields positive accuracy gains in all four model–task settings and consistently exceeds the SCG Low-R reverse control. The selection diagnostics confirm that High-R selects examples with larger mean resistance, whereas Low-R selects lower-resistance examples. Skill-Coverage selects subsets with broader skill coverage, but its accuracy gains do not consistently exceed those of High-R. High-R also yields gains comparable to or larger than High-LM-Loss across the evaluated settings. Together, these results suggest that SCG resistance provides a structural signal complementary to conventional difficulty and skill coverage under limited fine-tuning budgets.

5 Discussion

Our results support a functional-topology view in which generalization is related not only to individual skill-indexed components, but also to their organization within the SCG. Across model scales, richer induced skill structure and lower task-conditioned resistance are associated with stronger held-out performance, while subspace ablations support the functional relevance of the induced skill directions. For adaptation, selecting high-resistance examples yields consistent gains over the low-resistance control under limited fine-tuning budgets, suggesting that SCG resistance can serve as a structural data-selection signal. Future work should examine the robustness of these findings across broader model families, tasks, and alternative definitions of skills.

AI USE DISCLOSURE

In this work, we used GPT-5.6 to assist with the post-hoc semantic interpretation of the induced skill nodes presented in Figure 1, using the associated token sets as input. We also used GPT-5.6 to assist with drafting and editing portions of the manuscript to improve clarity and readability. All AI-assisted content was reviewed and verified by the authors. We take responsibility for the final content of this work, including text and artifacts produced with the aid of generative AI.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

Hao Bai and Yi Ma. Improving neuron-level interpretability with white-box language models. In Beidi Chen, Shijia Liu, Mert Pilanci, Weijie Su, Jeremias Sulam, Yuxiang Wang, and Zhihui Zhu (eds.), *Conference on Parsimony and Learning*, volume 280 of *Proceedings of Machine Learning Research*, pp. 810–836. PMLR, 24–27 Mar 2025.

Mikhail Belkin, Irina Matveeva, and Partha Niyogi. Regularization and semi-supervised learning on large graphs. In *International conference on computational learning theory*, pp. 624–638. Springer, 2004.

Leonardo Berti, Flavio Giorgi, and Gjergji Kasneci. Emergent abilities in large language models: A survey. *arXiv preprint arXiv:2503.05788*, 2025.

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.

Rishi Bommasani, Percy Liang, and Tony Lee. Holistic evaluation of language models. *Annals of the New York Academy of Sciences*, 1525(1):140–146, 2023.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. *Advances in Neural Information Processing Systems*, 37:66479–66567, 2024.

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 preprint arXiv:1803.05457*, 2018.

Aniket Didolkar, Anirudh Goyal, Nan R Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Rezende, Yoshua Bengio, Michael Mozer, and Sanjeev Arora. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812, 2024.

James Flemings, Meisam Razaviyayn, and Murali Annavaram. Differentially private next-token prediction of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4390–4404, 2024.

Lang Gao, Jiahui Geng, Xiangliang Zhang, Preslav Nakov, and Xiuying Chen. Shaping the safety boundaries: Understanding and defending against jailbreaks in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25378–25398, 2025.

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

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, et al. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*, 2024.

Simran Kaur, Simon Park, Anirudh Goyal, and Sanjeev Arora. Instruct-skillmix: A powerful pipeline for llm instruction tuning. In *International Conference on Learning Representations*, volume 2025, pp. 28842–28983, 2025.

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.

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

Zhihao Li, Xue Jiang, Liyuan Liu, Xuelin Zhang, Hong Chen, and Feng Zheng. On the generalization ability of next-token-prediction pretraining. In *Forty-second International Conference on Machine Learning*, 2025b.

Samuel Lippl, Thomas McGee, Kimberly Lopez, Ziwen Pan, Pierce Zhang, Salma Ziadi, Oliver Eberle, and Ida Momennejad. Algorithmic primitives and compositional geometry of reasoning in language models. *arXiv preprint arXiv:2510.15987*, 2025.

Liam Madden, Curtis Fox, and Christos Thrampoulidis. Upper and lower memory capacity bounds of transformers for next-token prediction. *arXiv preprint arXiv:2405.13718*, 2024.

Eran Malach. Auto-regressive next-token predictors are universal learners. *arXiv preprint arXiv:2309.06979*, 2023.

Samuel Marks, Can Rager, Eric Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations*, volume 2025, pp. 23888–23923, 2025.

Eric Michaud, Ziming Liu, Uzay Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability. *arXiv preprint arXiv:2301.05217*, 2023.

Shiwen Ni, Guhong Chen, Shuaimin Li, Xuanang Chen, Siyi Li, Bingli Wang, Qiyao Wang, Xingjian Wang, Yifan Zhang, Liyang Fan, et al. A survey on large language model benchmarks. *arXiv preprint arXiv:2508.15361*, 2025.

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.

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.

Dong Shu, Xuansheng Wu, Haiyan Zhao, Daking Rai, Ziyu Yao, Ninghao Liu, and Mengnan Du. A survey on sparse autoencoders: Interpreting the internal mechanisms of large language models. *arXiv preprint arXiv:2503.05613*, 2025.

Alexander J Smola and Risi Kondor. Kernels and regularization on graphs. In *Learning theory and kernel machines: 16th annual conference on learning theory and 7th kernel workshop, COLT/kernel 2003, Washington, DC, USA, august 24-27, 2003. Proceedings*, pp. 144–158. Springer, 2003.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on machine learning research*, 2023.

Jonathan Thomm, Giacomo Camposampiero, Aleksandar Terzic, Michael Hersche, Bernhard Schölkopf, and Abbas Rahimi. Limits of transformer language models on learning to compose algorithms. *Advances in Neural Information Processing Systems*, 37:7631–7674, 2024.

Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.

Rui-Ray Zhang and Massih-Reza Amini. Generalization bounds for learning under graph-dependence: A survey. *Machine Learning*, 113(7):3929–3959, 2024.

Haoyu Zhao, Simran Kaur, Dingli Yu, Anirudh Goyal, and Sanjeev Arora. Can models learn skill composition from examples? *Advances in Neural Information Processing Systems*, 37:102393–102427, 2024.

A Complete Theoretical Derivations

This appendix provides the complete derivations of the theoretical results used in the main text. We first derive the skill-alignment formulation from the autoregressive negative log-likelihood. We then establish a task-conditioned generalization bound for graph-smooth skill-alignment functions. Finally, we extend the result to the practical setting where the SCG is constructed from gradient-induced skill partitions and empirical co-activation statistics.

A.1 Notation and Setup

Let $x = (x_{<t}, x_t)$ denote an autoregressive prediction instance, where $x_{<t}$ is the preceding context and x_t is the target token to be predicted. Let

$$S = \{x_1, \ldots, x_n\}$$

be an i.i.d. sample drawn from an unknown data distribution $\mathcal{D}$. We consider a finite set of skill-indexed components

$$\mathcal{S} = \{s_1, \ldots, s_K\}.$$

For a fixed context $x_{<t}$, let

$$\pi_k(x_{<t})$$

denote the context-dependent attribution weight assigned to the k-th skill-indexed component. These weights satisfy

$$\pi_k(x_{<t}) \geq 0, \qquad \sum_{k=1}^{K} \pi_k(x_{<t}) = 1.$$

When the context is clear, we write π_k for $\pi_k(x_{<t})$. Let

$$p_k(x_t \mid x_{<t})$$

denote the corresponding skill-conditioned predictive likelihood. We use the following analysis-level mixture decomposition:

$$P_\theta(x_t \mid x_{<t}) = \sum_{k=1}^{K} \pi_k(x_{<t})\, p_k(x_t \mid x_{<t}).$$

This decomposition should be understood as an analytical representation of predictive attribution over skill-indexed components, rather than as an assumption that the model explicitly samples a discrete latent skill.

Let

$$\Phi(x) = \big(\Phi_1(x), \ldots, \Phi_K(x)\big) \in \Delta^{K-1}$$

be an auxiliary skill-attribution distribution, where

$$\Phi_k(x) > 0, \qquad \sum_{k=1}^{K} \Phi_k(x) = 1.$$

The auxiliary distribution $\Phi(x)$ is introduced for variational analysis and need not coincide with the context-dependent attribution weights $\pi(x_{<t})$. Let

$$\mathbf{1} = (1, \ldots, 1)^\top \in \mathbb{R}^K$$

and define the centering projection

$$\Pi = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top.$$

The centered skill-attribution vector is

$$z(x) = \Pi\Phi(x).$$

Since Π is an orthogonal projection and $\Phi(x)$ lies in the probability simplex, we have

$$\|z(x)\|_2 = \|\Pi\Phi(x)\|_2 \leq \|\Phi(x)\|_2 \leq \|\Phi(x)\|_1 = 1.$$

Therefore,

$$\|z(x)\|_2 \le 1.$$

Let $G = (V, E, W)$ be an SCG with $|V| = K$, where each node corresponds to a skill-indexed component and $W \in \mathbb{R}^{K\times K}$ is a symmetric nonnegative edge-weight matrix. Let

$$D = \mathrm{diag}(d_1, \dots, d_K), \qquad d_i = \sum_{j=1}^{K} W_{ij},$$

and define the unnormalized graph Laplacian

$$L = D - W.$$

We assume that G is connected. Hence the eigenvalues of L satisfy

$$0 = \mu_1 < \mu_2 \le \cdots \le \mu_K,$$

and

$$\mathrm{Null}(L) = \mathrm{span}\{\mathbf{1}\}.$$

On the centered subspace $\mathbf{1}^\perp$, the Laplacian is positive definite. Its Moore–Penrose pseudoinverse is denoted by L^+. If

$$L = \sum_{r=2}^{K} \mu_r u_r u_r^\top$$

on $\mathbf{1}^\perp$, then

$$L^+ = \sum_{r=2}^{K} \frac{1}{\mu_r} u_r u_r^\top.$$

We restrict skill coefficient vectors to the centered subspace:

$$a \in \mathbf{1}^\perp, \qquad \mathbf{1}^\top a = 0.$$

This restriction removes the constant direction that is not controlled by the graph Laplacian. Indeed, $L\mathbf{1} = 0$, and thus the graph-smoothness term cannot penalize shifts of the form $a + c\mathbf{1}$. Since such constant components induce only input-independent offsets in the alignment score, they can be absorbed into the threshold or baseline terms. The graph-smooth skill-alignment hypothesis class is defined as

$$\mathcal{H}_G(\lambda) = \left\{ h_a(x) = a^\top z(x) \ : \ a \in \mathbf{1}^\perp,\ a^\top L a \le \lambda \right\}.$$

For a sample $S = \{x_1, \dots, x_n\}$, write

$$z_i = z(x_i),$$

and define the empirical task-conditioned second-moment matrix

$$\Sigma_S = \frac{1}{n} \sum_{i=1}^{n} z_i z_i^\top.$$

Since $\|z_i\|_2 \le 1$, we have

$$\mathrm{Tr}(\Sigma_S) = \frac{1}{n} \sum_{i=1}^{n} \|z_i\|_2^2 \le 1.$$

A.2 Loss-Induced Skill Alignment Decomposition

We first show that the autoregressive negative log-likelihood admits a centered skill-alignment decomposition through a variational upper bound.

Theorem 4 (Loss-Induced Skill Alignment Decomposition)**.** *For any autoregressive prediction instance $x = (x_{<t}, x_t)$, suppose*

$$P_\theta(x_t \mid x_{<t}) = \sum_{k=1}^{K} \pi_k(x_{<t})\, p_k(x_t \mid x_{<t}).$$

Let $\Phi(x) \in \Delta^{K-1}$ be any auxiliary skill-attribution distribution with $\Phi_k(x) > 0$. Then the autoregressive negative log-likelihood

$$\mathcal{L}_{\mathrm{AR}}(x) := -\log P_\theta(x_t \mid x_{<t})$$

satisfies

$$\mathcal{L}_{\mathrm{AR}}(x) \le -\sum_{k=1}^{K} \Phi_k(x) \log p_k(x_t \mid x_{<t}) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big).$$

Furthermore, let $\bar{a} \in \mathbb{R}^K$ be a fixed reference coefficient vector. For any $\beta > 0$ and any scalar baseline $c(x)$, define

$$\bar{r}_k(x) = \log p_k(x_t \mid x_{<t}) - c(x) - \beta \bar{a}_k.$$

Then

$$\mathcal{L}_{\mathrm{AR}}(x) \le C(x) - \beta \bar{a}^\top \Phi(x) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big) + \bar{\varepsilon}_r(x),$$

where

$$C(x) = -c(x), \qquad \bar{\varepsilon}_r(x) = -\sum_{k=1}^{K} \Phi_k(x) \bar{r}_k(x).$$

If $\bar{a} = \bar{a}_0 \mathbf{1} + \bar{a}_\perp$, where

$$\bar{a}_0 = \frac{1}{K} \mathbf{1}^\top \bar{a}, \qquad \bar{a}_\perp \in \mathbf{1}^\perp,$$

then

$$\bar{a}^\top \Phi(x) = \bar{a}_0 + \bar{a}_\perp^\top z(x), \qquad z(x) = \Pi \Phi(x).$$

Consequently, after absorbing the input-independent constant $-\beta \bar{a}_0$ into the baseline, the explicit centered skill-dependent term in the loss upper bound is

$$-\beta \bar{a}_\perp^\top z(x).$$

Thus, the autoregressive loss upper bound gives rise to a centered linear skill-alignment form.

Proof. By the analysis-level skill decomposition,

$$P_\theta(x_t \mid x_{<t}) = \sum_{k=1}^{K} \pi_k(x_{<t})\, p_k(x_t \mid x_{<t}).$$

For notational simplicity, write $\pi_k = \pi_k(x_{<t})$. Since $\Phi_k(x) > 0$ and $\sum_{k=1}^{K} \Phi_k(x) = 1$, we can rewrite the predictive likelihood as

$$P_\theta(x_t \mid x_{<t}) = \sum_{k=1}^{K} \Phi_k(x) \frac{\pi_k p_k(x_t \mid x_{<t})}{\Phi_k(x)}.$$

Define

$$u_k(x) = \frac{\pi_k p_k(x_t \mid x_{<t})}{\Phi_k(x)}.$$

Then

$$P_\theta(x_t \mid x_{<t}) = \sum_{k=1}^{K} \Phi_k(x) u_k(x).$$

The autoregressive loss is therefore

$$\mathcal{L}_{\mathrm{AR}}(x) = -\log \left(\sum_{k=1}^{K} \Phi_k(x) u_k(x) \right).$$

Since $f(u) = -\log u$ is convex on $u > 0$, Jensen's inequality gives

$$-\log \left(\sum_{k=1}^{K} \Phi_k(x) u_k(x) \right) \le \sum_{k=1}^{K} \Phi_k(x) \big(-\log u_k(x) \big).$$

Substituting the definition of $u_k(x)$, we obtain

$$\mathcal{L}_{\mathrm{AR}}(x) \le -\sum_{k=1}^{K} \Phi_k(x) \log\left(\frac{\pi_k p_k(x_t \mid x_{<t})}{\Phi_k(x)}\right).$$

Expanding the logarithm yields

$$\log\left(\frac{\pi_k p_k(x_t \mid x_{<t})}{\Phi_k(x)}\right) = \log \pi_k + \log p_k(x_t \mid x_{<t}) - \log \Phi_k(x).$$

Hence

$$\begin{aligned}\mathcal{L}_{\mathrm{AR}}(x) &\le -\sum_{k=1}^{K} \Phi_k(x) \log p_k(x_t \mid x_{<t}) - \sum_{k=1}^{K} \Phi_k(x) \log \pi_k + \sum_{k=1}^{K} \Phi_k(x) \log \Phi_k(x) \\ &= -\sum_{k=1}^{K} \Phi_k(x) \log p_k(x_t \mid x_{<t}) + \sum_{k=1}^{K} \Phi_k(x) \log \frac{\Phi_k(x)}{\pi_k}.\end{aligned}$$

The second term is the KL divergence

$$D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big) = \sum_{k=1}^{K} \Phi_k(x) \log \frac{\Phi_k(x)}{\pi_k}.$$

Therefore,

$$\mathcal{L}_{\mathrm{AR}}(x) \le -\sum_{k=1}^{K} \Phi_k(x) \log p_k(x_t \mid x_{<t}) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big).$$

We next introduce the fixed reference decomposition of the skill-conditioned log-likelihood. By definition,

$$\bar{r}_k(x) = \log p_k(x_t \mid x_{<t}) - c(x) - \beta \bar{a}_k,$$

which is equivalent to

$$\log p_k(x_t \mid x_{<t}) = c(x) + \beta \bar{a}_k + \bar{r}_k(x).$$

Substituting this identity into the preceding upper bound gives

$$\begin{aligned}\mathcal{L}_{\mathrm{AR}}(x) &\le -\sum_{k=1}^{K} \Phi_k(x)\big(c(x) + \beta \bar{a}_k + \bar{r}_k(x)\big) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big) \\ &= -c(x) \sum_{k=1}^{K} \Phi_k(x) - \beta \sum_{k=1}^{K} \Phi_k(x) \bar{a}_k - \sum_{k=1}^{K} \Phi_k(x) \bar{r}_k(x) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big).\end{aligned}$$

Since $\sum_{k=1}^{K} \Phi_k(x) = 1$, this becomes

$$\mathcal{L}_{\mathrm{AR}}(x) \le -c(x) - \beta \bar{a}^{\top} \Phi(x) - \sum_{k=1}^{K} \Phi_k(x) \bar{r}_k(x) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big).$$

Let

$$C(x) = -c(x), \qquad \bar{\varepsilon}_r(x) = -\sum_{k=1}^{K} \Phi_k(x) \bar{r}_k(x).$$

Then

$$\mathcal{L}_{\mathrm{AR}}(x) \le C(x) - \beta \bar{a}^{\top} \Phi(x) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big) + \bar{\varepsilon}_r(x).$$

It remains to remove the constant direction. Decompose $\bar{a}$ as

$$\bar{a} = \bar{a}_0 \mathbf{1} + \bar{a}_{\perp}, \qquad \bar{a}_0 = \frac{1}{K} \mathbf{1}^{\top} \bar{a}, \qquad \bar{a}_{\perp} \in \mathbf{1}^{\perp}.$$

Then

$$\bar{a}^{\top} \Phi(x) = (\bar{a}_0 \mathbf{1} + \bar{a}_{\perp})^{\top} \Phi(x) = \bar{a}_0 \mathbf{1}^{\top} \Phi(x) + \bar{a}_{\perp}^{\top} \Phi(x).$$

Since $\mathbf{1}^\top \Phi(x) = 1$,

$$\bar{a}^\top \Phi(x) = \bar{a}_0 + \bar{a}_\perp^\top \Phi(x).$$

Moreover, because $\bar{a}_\perp^\top \mathbf{1} = 0$,

$$\begin{aligned}\bar{a}_\perp^\top \Pi\Phi(x) &= \bar{a}_\perp^\top \left(I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right)\Phi(x) \\ &= \bar{a}_\perp^\top \Phi(x) - \frac{1}{K}\bar{a}_\perp^\top \mathbf{1}\mathbf{1}^\top \Phi(x) \\ &= \bar{a}_\perp^\top \Phi(x).\end{aligned}$$

Since $z(x) = \Pi\Phi(x)$, we have

$$\bar{a}_\perp^\top \Phi(x) = \bar{a}_\perp^\top z(x).$$

Therefore,

$$\bar{a}^\top \Phi(x) = \bar{a}_0 + \bar{a}_\perp^\top z(x).$$

Substituting this back into the loss upper bound gives

$$\mathcal{L}_{\mathrm{AR}}(x) \le C(x) - \beta\bar{a}_0 - \beta\bar{a}_\perp^\top z(x) + D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big) + \bar{\varepsilon}_r(x).$$

Absorbing the constant term $-\beta\bar{a}_0$ into the baseline yields the claimed centered linear skill-alignment form. This completes the proof. □

A.3 Task-Conditioned SCG Generalization Bound

The preceding result establishes a centered linear skill-alignment form. Motivated by this form, we consider $h_a(x) = a^\top z(x)$, with $a \in \mathbf{1}^\perp$, and analyze the generalization of graph-smooth skill-alignment functions.

Skill-alignment surrogate loss. To keep the sample-dependent offset fixed over the hypothesis class, we retain the fixed reference decomposition from Theorem 4 and define

$$q(x) = D_{\mathrm{KL}}\big(\Phi(x) \,\|\, \pi(x_{<t})\big) + \bar{\varepsilon}_r(x),$$

where

$$\bar{\varepsilon}_r(x) = -\sum_{k=1}^{K} \Phi_k(x)\left[\log p_k(x_t \mid x_{<t}) - c(x) - \beta\bar{a}_k\right],$$

and $\bar{a} \in \mathbb{R}^K$ is the fixed reference coefficient vector. Thus, $q(x)$ is fixed with respect to the coefficient a varied over the hypothesis class. Let $\tau > 0$, $\beta > 0$, and define the clipped hinge function

$$\rho(u) = \min\{M, \max\{0, u\}\},$$

where $M > 0$. The skill-alignment surrogate loss is

$$\ell_a(x) = \rho\left(\tau - \beta h_a(x) + q(x)\right) = \rho\left(\tau - \beta a^\top z(x) + q(x)\right).$$

By construction,

$$0 \le \ell_a(x) \le M.$$

Let

$$\mathcal{L}_G(\lambda) = \left\{\ell_a : a \in \mathbf{1}^\perp,\ a^\top L a \le \lambda\right\}.$$

Lemma 7 (Rademacher Complexity of the Graph-Smooth Alignment Class)**.** *For the hypothesis class*

$$\mathcal{H}_G(\lambda) = \left\{h_a(x) = a^\top z(x) : a \in \mathbf{1}^\perp,\ a^\top L a \le \lambda\right\},$$

the empirical Rademacher complexity satisfies

$$\widehat{\mathfrak{R}}_S\big(\mathcal{H}_G(\lambda)\big) \le \sqrt{\frac{\lambda}{n}\,\mathrm{Tr}(L^+\Sigma_S)}.$$

Proof. By definition,

$$\widehat{\mathfrak{R}}_S\big(\mathcal{H}_G(\lambda)\big) = \mathbb{E}_\sigma\left[\sup_{a\in\mathbf{1}^\perp,\, a^\top La\le\lambda} \frac{1}{n}\sum_{i=1}^n \sigma_i h_a(x_i)\right],$$

where $\sigma_1,\dots,\sigma_n$ are independent Rademacher variables. Since $h_a(x_i) = a^\top z_i$,

$$\sum_{i=1}^n \sigma_i h_a(x_i) = a^\top\left(\sum_{i=1}^n \sigma_i z_i\right).$$

Let

$$Z = [z_1,\dots,z_n] \in \mathbb{R}^{K\times n}, \qquad \sigma = (\sigma_1,\dots,\sigma_n)^\top.$$

Then

$$\sum_{i=1}^n \sigma_i z_i = Z\sigma.$$

Thus

$$\widehat{\mathfrak{R}}_S\big(\mathcal{H}_G(\lambda)\big) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{a\in\mathbf{1}^\perp,\, a^\top La\le\lambda} a^\top Z\sigma\right].$$

Because L is positive definite on $\mathbf{1}^\perp$, define

$$b = L^{1/2}a.$$

Then

$$\|b\|_2^2 = a^\top La \le \lambda,$$

so $\|b\|_2 \le \sqrt{\lambda}$. Conversely, on $\mathbf{1}^\perp$,

$$a = L^{+1/2}b.$$

Hence

$$a^\top Z\sigma = b^\top L^{+1/2}Z\sigma.$$

Therefore,

$$\sup_{a\in\mathbf{1}^\perp,\, a^\top La\le\lambda} a^\top Z\sigma = \sup_{\substack{b\in\mathbf{1}^\perp\\ \|b\|_2\le\sqrt{\lambda}}} b^\top L^{+1/2}Z\sigma.$$

By the Cauchy–Schwarz inequality,

$$b^\top L^{+1/2}Z\sigma \le \|b\|_2\,\|L^{+1/2}Z\sigma\|_2 \le \sqrt{\lambda}\,\|L^{+1/2}Z\sigma\|_2.$$

Since $Z\sigma \in \mathbf{1}^\perp$, the supremum is attained by choosing b aligned with $L^{+1/2}Z\sigma$. Hence

$$\sup_{\substack{b\in\mathbf{1}^\perp\\ \|b\|_2\le\sqrt{\lambda}}} b^\top L^{+1/2}Z\sigma = \sqrt{\lambda}\,\|L^{+1/2}Z\sigma\|_2.$$

Consequently,

$$\widehat{\mathfrak{R}}_S\big(\mathcal{H}_G(\lambda)\big) = \frac{\sqrt{\lambda}}{n}\mathbb{E}_\sigma\left[\|L^{+1/2}Z\sigma\|_2\right].$$

Applying Jensen's inequality gives

$$\mathbb{E}_\sigma\left[\|L^{+1/2}Z\sigma\|_2\right] \le \sqrt{\mathbb{E}_\sigma\left[\|L^{+1/2}Z\sigma\|_2^2\right]}.$$

Thus

$$\widehat{\mathfrak{R}}_S\big(\mathcal{H}_G(\lambda)\big) \le \frac{\sqrt{\lambda}}{n}\sqrt{\mathbb{E}_\sigma\left[\|L^{+1/2}Z\sigma\|_2^2\right]}.$$

We now compute the expectation inside the square root:

$$\begin{aligned}\|L^{+1/2}Z\sigma\|_2^2 &= (L^{+1/2}Z\sigma)^\top(L^{+1/2}Z\sigma)\\ &= \sigma^\top Z^\top L^+ Z\sigma.\end{aligned}$$

Let

$$M_Z = Z^\top L^+ Z.$$

Then

$$\sigma^\top M_Z \sigma = \sum_{i=1}^{n}\sum_{j=1}^{n} \sigma_i (M_Z)_{ij} \sigma_j.$$

Since $\mathbb{E}[\sigma_i \sigma_j] = 0$ for $i \neq j$ and $\mathbb{E}[\sigma_i^2] = 1$,

$$\mathbb{E}_\sigma[\sigma^\top M_Z \sigma] = \sum_{i=1}^{n} (M_Z)_{ii} = \mathrm{Tr}(M_Z).$$

Therefore,

$$\mathbb{E}_\sigma \left[\|L^{+1/2} Z\sigma\|_2^2 \right] = \mathrm{Tr}(Z^\top L^+ Z).$$

Using cyclic invariance of the trace,

$$\mathrm{Tr}(Z^\top L^+ Z) = \mathrm{Tr}(L^+ Z Z^\top).$$

Moreover,

$$ZZ^\top = \sum_{i=1}^{n} z_i z_i^\top = n\Sigma_S.$$

Hence

$$\mathrm{Tr}(L^+ ZZ^\top) = n\,\mathrm{Tr}(L^+ \Sigma_S).$$

Substituting back,

$$\widehat{\mathfrak{R}}_S\big(\mathcal{H}_G(\lambda)\big) \leq \sqrt{\frac{\lambda}{n}\,\mathrm{Tr}(L^+\Sigma_S)}.$$

This completes the proof. □

Lemma 8 (Rademacher Complexity of the Skill-Alignment Loss Class)**.** *The empirical Rademacher complexity of the loss class $\mathcal{L}_G(\lambda)$ satisfies*

$$\widehat{\mathfrak{R}}_S\big(\mathcal{L}_G(\lambda)\big) \leq \beta\sqrt{\frac{\lambda}{n}\,\mathrm{Tr}(L^+\Sigma_S)}.$$

Proof. For each sample x_i, define

$$g_i(u) = \rho\left(\tau - \beta u + q(x_i)\right).$$

Then

$$\ell_a(x_i) = g_i(a^\top z_i).$$

Since $q(x_i)$ is fixed with respect to a, ρ is 1-Lipschitz, and $u \mapsto \tau - \beta u + q(x_i)$ is β-Lipschitz, the composition g_i is β-Lipschitz:

$$|g_i(u) - g_i(v)| \leq \beta |u - v|.$$

The empirical Rademacher complexity of the loss class is

$$\widehat{\mathfrak{R}}_S\big(\mathcal{L}_G(\lambda)\big) = \mathbb{E}_\sigma \left[\sup_{a \in \mathbf{1}^\perp,\, a^\top L a \leq \lambda} \frac{1}{n} \sum_{i=1}^{n} \sigma_i g_i(a^\top z_i) \right].$$

Define

$$\widetilde{g}_i(u) = g_i(u) - g_i(0).$$

Then $\widetilde{g}_i(0) = 0$, and $\widetilde{g}_i$ remains β-Lipschitz. Moreover,

$$g_i(a^\top z_i) = \widetilde{g}_i(a^\top z_i) + g_i(0).$$

Thus

$$\sum_{i=1}^{n} \sigma_i g_i(a^\top z_i) = \sum_{i=1}^{n} \sigma_i \widetilde{g}_i(a^\top z_i) + \sum_{i=1}^{n} \sigma_i g_i(0).$$

The second term is independent of a, and its expectation over the Rademacher variables is zero:

$$\mathbb{E}_\sigma\left[\sum_{i=1}^{n}\sigma_i g_i(0)\right]=0.$$

Therefore,

$$\widehat{\mathfrak{R}}_S\big(\mathcal{L}_G(\lambda)\big)=\mathbb{E}_\sigma\left[\sup_{a\in\mathbf{1}^\perp,\ a^\top La\le\lambda}\frac{1}{n}\sum_{i=1}^{n}\sigma_i\widetilde{g}_i(a^\top z_i)\right].$$

By the Rademacher contraction inequality,

$$\widehat{\mathfrak{R}}_S\big(\mathcal{L}_G(\lambda)\big)\le\beta\mathbb{E}_\sigma\left[\sup_{a\in\mathbf{1}^\perp,\ a^\top La\le\lambda}\frac{1}{n}\sum_{i=1}^{n}\sigma_i a^\top z_i\right].$$

The right-hand side is

$$\beta\,\widehat{\mathfrak{R}}_S\big(\mathcal{H}_G(\lambda)\big).$$

Applying Lemma 7 yields

$$\widehat{\mathfrak{R}}_S\big(\mathcal{L}_G(\lambda)\big)\le\beta\sqrt{\frac{\lambda}{n}\operatorname{Tr}(L^+\Sigma_S)}.$$

This completes the proof. □

Theorem 5 (Task-Conditioned SCG Generalization Bound)**.** *Assume $0\le\ell_a(x)\le M$ for all x and all $a\in\mathbf{1}^\perp$ satisfying $a^\top La\le\lambda$. Then, for any $\delta>0$, with probability at least $1-\delta$ over the draw of $S=\{x_1,\dots,x_n\}$, the following holds uniformly for all $a\in\mathbf{1}^\perp$ with $a^\top La\le\lambda$:*

$$\mathbb{E}_{x\sim\mathcal{D}}[\ell_a(x)]\le\frac{1}{n}\sum_{i=1}^{n}\ell_a(x_i)+2\beta\sqrt{\frac{\lambda}{n}\operatorname{Tr}(L^+\Sigma_S)}+3M\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. We use the standard Rademacher uniform-convergence bound. For any function class $\mathcal{F}$ with functions mapping into $[0,M]$, with probability at least $1-\delta$, uniformly for all $f\in\mathcal{F}$,

$$\mathbb{E}[f(x)]\le\frac{1}{n}\sum_{i=1}^{n}f(x_i)+2\widehat{\mathfrak{R}}_S(\mathcal{F})+3M\sqrt{\frac{\log(2/\delta)}{2n}}.$$

Apply this result to

$$\mathcal{F}=\mathcal{L}_G(\lambda).$$

Then, with probability at least $1-\delta$, for all $\ell_a\in\mathcal{L}_G(\lambda)$,

$$\mathbb{E}[\ell_a(x)]\le\frac{1}{n}\sum_{i=1}^{n}\ell_a(x_i)+2\widehat{\mathfrak{R}}_S\big(\mathcal{L}_G(\lambda)\big)+3M\sqrt{\frac{\log(2/\delta)}{2n}}.$$

By Lemma 2,

$$\widehat{\mathfrak{R}}_S\big(\mathcal{L}_G(\lambda)\big)\le\beta\sqrt{\frac{\lambda}{n}\operatorname{Tr}(L^+\Sigma_S)}.$$

Substituting this into the uniform-convergence bound gives

$$\mathbb{E}[\ell_a(x)]\le\frac{1}{n}\sum_{i=1}^{n}\ell_a(x_i)+2\beta\sqrt{\frac{\lambda}{n}\operatorname{Tr}(L^+\Sigma_S)}+3M\sqrt{\frac{\log(2/\delta)}{2n}}.$$

This proves the theorem. □

Spectral interpretation. Let

$$L = \sum_{r=2}^{K} \mu_r u_r u_r^\top, \qquad L^+ = \sum_{r=2}^{K} \frac{1}{\mu_r} u_r u_r^\top.$$

Then

$$\begin{aligned}\mathrm{Tr}(L^+ \Sigma_S) &= \mathrm{Tr}\left(\sum_{r=2}^{K} \frac{1}{\mu_r} u_r u_r^\top \Sigma_S \right) \\ &= \sum_{r=2}^{K} \frac{1}{\mu_r} \mathrm{Tr}(u_r u_r^\top \Sigma_S) \\ &= \sum_{r=2}^{K} \frac{u_r^\top \Sigma_S u_r}{\mu_r}.\end{aligned}$$

Thus, the complexity term is larger when task-induced skill demand places substantial energy on weakly connected, high-resistance modes of the SCG. Conversely, the bound is tighter when task-induced skill demand concentrates on well-integrated, low-resistance modes.

Corollary 2 (Topology-Only Bound as a Coarse Upper Bound)**.** *Since* $\mathrm{Tr}(\Sigma_S) \le 1$*, the task-conditioned complexity satisfies*

$$\mathrm{Tr}(L^+ \Sigma_S) \le \mathrm{Tr}(L^+).$$

Consequently,

$$\sqrt{\frac{\lambda}{n} \mathrm{Tr}(L^+ \Sigma_S)} \le \sqrt{\frac{\lambda}{n} \mathrm{Tr}(L^+)}.$$

Proof. Both L^+ and Σ_S are positive semidefinite. For positive semidefinite matrices A and B,

$$\mathrm{Tr}(AB) \le \mathrm{Tr}(A)\,\mathrm{Tr}(B).$$

Taking $A = L^+$ and $B = \Sigma_S$ gives

$$\mathrm{Tr}(L^+ \Sigma_S) \le \mathrm{Tr}(L^+)\,\mathrm{Tr}(\Sigma_S).$$

Since $\mathrm{Tr}(\Sigma_S) \le 1$,

$$\mathrm{Tr}(L^+ \Sigma_S) \le \mathrm{Tr}(L^+).$$

The stated inequality follows by multiplying both sides by λ/n and taking square roots. □

A.4 Stability-Aware Learned-SCG Generalization Bound

The preceding theorem assumes that the SCG Laplacian L is fixed and known. In practice, however, the SCG is constructed from gradient-induced skill partitions and empirical co-activation statistics. We now extend the analysis to this estimated-graph setting.

A.4.1 Candidate cluster numbers and stability score

Let $\mathcal{K}$ be a finite set of candidate cluster numbers. For each $K \in \mathcal{K}$, we perform B repeated subsampling runs. Let

$$\mathcal{P}_K^{(b)}$$

denote the partition obtained from the b-th subsampled token set, where $b = 1, \dots, B$. These subsampled partitions are used only to estimate the stability of the candidate cluster number K. For two runs $b \neq c$, let

$$\mathrm{ARI}_{b,c}(K)$$

be the Adjusted Rand Index between $\mathcal{P}_K^{(b)}$ and $\mathcal{P}_K^{(c)}$, computed on the intersection of the corresponding subsampled token sets. Define the mean ARI as

$$\overline{\mathrm{ARI}}(K) = \frac{2}{B(B-1)} \sum_{1 \le b < c \le B} \mathrm{ARI}_{b,c}(K),$$

and define the ARI variance as

$$\mathrm{Var}_{\mathrm{ARI}}(K) = \frac{2}{B(B-1)} \sum_{1 \le b < c \le B} \left(\mathrm{ARI}_{b,c}(K) - \overline{\mathrm{ARI}}(K)\right)^2.$$

We define the stability error score

$$s(K) = 1 - \overline{\mathrm{ARI}}(K) + \alpha \sqrt{\mathrm{Var}_{\mathrm{ARI}}(K)} + \eta \Omega(K),$$

where $\alpha, \eta > 0$ are constants and $\Omega(K)$ is a complexity penalty, for example $\Omega(K) = K/N_{\mathrm{tok}}$, with N_{tok} denoting the number of tokens used for skill induction.

After computing $s(K)$, we rerun spectral clustering on the full token set under the same candidate K. Let

$$\widehat{\mathcal{P}}_K$$

be the resulting full-data partition, and let

$$\widehat{G}_K$$

be the empirical SCG constructed from this partition and the corresponding empirical co-activation statistics. We denote its Laplacian by

$$\widehat{L}_K.$$

Thus, the subsampled partitions $\{\mathcal{P}_K^{(b)}\}_{b=1}^B$ estimate the stability of the induced skill partition, whereas $\widehat{L}_K$ is the final empirical SCG Laplacian constructed from the full graph-construction data under candidate K.

Assumption 2 (Stability-to-graph perturbation)**.** *For each $K \in \mathcal{K}$, let $L_K^\star$ denote the population co-activation Laplacian associated with the ideal K-component skill structure, and let $\widehat{L}_K$ be the full-data empirical SCG Laplacian. Because cluster indices are arbitrary up to permutation, we assume that the empirical and population skill indices have first been aligned by a permutation, and write $\widehat{L}_K$ for the aligned empirical Laplacian. There exists a constant $C_L > 0$ such that, with probability at least $1 - \delta_G$,*

$$\|\widehat{L}_K - L_K^\star\|_{\mathrm{op}} \le C_L s(K), \qquad \forall K \in \mathcal{K}.$$

This assumption formalizes the stability score as an observable proxy for graph-estimation error. It does not equate ARI stability with Laplacian perturbation. Rather, under a fixed SCG construction pipeline, it links instability of the induced skill partition to instability of the resulting empirical co-activation graph, up to a problem-dependent constant C_L.

For simplicity, define

$$\varepsilon_K = C_L s(K).$$

Let

$$m_K = \mu_2(L_K^\star)$$

be the algebraic connectivity of the population SCG. Throughout this section, we assume

$$0 < \varepsilon_K < m_K.$$

Lemma 9 (Lower bound on the learned graph connectivity)**.** *Under Assumption 2, for every $K \in \mathcal{K}$,*

$$\mu_2(\widehat{L}_K) \ge m_K - \varepsilon_K.$$

Proof. Fix $K \in \mathcal{K}$. For any unit vector $u \in \mathbf{1}^\perp$,

$$u^\top L_K^\star u \ge m_K.$$

Moreover,

$$u^\top \widehat{L}_K u = u^\top L_K^\star u + u^\top (\widehat{L}_K - L_K^\star) u.$$

By the definition of the operator norm,

$$\left| u^\top (\widehat{L}_K - L_K^\star) u \right| \le \|\widehat{L}_K - L_K^\star\|_{\mathrm{op}} \|u\|_2^2 \le \varepsilon_K.$$

Since $\|u\|_2 = 1$,

$$u^\top(\widehat{L}_K - L_K^\star)u \geq -\varepsilon_K.$$

Therefore,

$$u^\top \widehat{L}_K u \geq m_K - \varepsilon_K.$$

Taking the minimum over all unit vectors $u \in \mathbf{1}^\perp$ yields

$$\mu_2(\widehat{L}_K) \geq m_K - \varepsilon_K.$$

□

Lemma 10 (Pseudoinverse perturbation)**.** *Suppose $L_K^\star$ and $\widehat{L}_K$ are connected graph Laplacians with the same nullspace* $\mathrm{span}\{\mathbf{1}\}$*. If*

$$\|\widehat{L}_K - L_K^\star\|_{\mathrm{op}} \leq \varepsilon_K < m_K,$$

then

$$\|\widehat{L}_K^+ - L_K^{\star+}\|_{\mathrm{op}} \leq \frac{\varepsilon_K}{m_K(m_K - \varepsilon_K)}.$$

Proof. Since both $L_K^\star$ and $\widehat{L}_K$ are connected graph Laplacians, their nullspace is $\mathrm{span}\{\mathbf{1}\}$. On the centered subspace $\mathbf{1}^\perp$, both matrices are positive definite, and their pseudoinverses coincide with ordinary inverses. By Lemma 9,

$$\mu_2(\widehat{L}_K) \geq m_K - \varepsilon_K.$$

Therefore,

$$\|\widehat{L}_K^+\|_{\mathrm{op}} = \frac{1}{\mu_2(\widehat{L}_K)} \leq \frac{1}{m_K - \varepsilon_K}.$$

Similarly,

$$\|L_K^{\star+}\|_{\mathrm{op}} = \frac{1}{\mu_2(L_K^\star)} = \frac{1}{m_K}.$$

On $\mathbf{1}^\perp$, the inverse perturbation identity gives

$$\widehat{L}_K^+ - L_K^{\star+} = \widehat{L}_K^+ \left(L_K^\star - \widehat{L}_K\right) L_K^{\star+}.$$

Taking the operator norm and applying submultiplicativity,

$$\begin{aligned}\|\widehat{L}_K^+ - L_K^{\star+}\|_{\mathrm{op}} &\leq \|\widehat{L}_K^+\|_{\mathrm{op}} \, \|L_K^\star - \widehat{L}_K\|_{\mathrm{op}} \, \|L_K^{\star+}\|_{\mathrm{op}} \\ &\leq \frac{1}{m_K - \varepsilon_K} \cdot \varepsilon_K \cdot \frac{1}{m_K}.\end{aligned}$$

Thus

$$\|\widehat{L}_K^+ - L_K^{\star+}\|_{\mathrm{op}} \leq \frac{\varepsilon_K}{m_K(m_K - \varepsilon_K)}.$$

This completes the proof. □

Lemma 11 (Transfer of empirical smoothness to population smoothness)**.** *Suppose $a \in \mathbf{1}^\perp$ satisfies the empirical graph-smoothness constraint*

$$a^\top \widehat{L}_K a \leq \lambda.$$

Then

$$a^\top L_K^\star a \leq \lambda_K, \qquad \lambda_K := \lambda \frac{m_K}{m_K - \varepsilon_K}.$$

Proof. We decompose

$$a^\top L_K^\star a = a^\top \widehat{L}_K a + a^\top (L_K^\star - \widehat{L}_K) a.$$

By assumption,

$$a^\top \widehat{L}_K a \leq \lambda.$$

Moreover,

$$a^\top (L_K^\star - \widehat{L}_K) a \leq \|L_K^\star - \widehat{L}_K\|_{\mathrm{op}} \|a\|_2^2 \leq \varepsilon_K \|a\|_2^2.$$

It remains to control $\|a\|_2^2$. Since $a \in \mathbf{1}^\perp$, by Lemma 9,

$$a^\top \widehat{L}_K a \geq \mu_2(\widehat{L}_K)\|a\|_2^2 \geq (m_K - \varepsilon_K)\|a\|_2^2.$$

Together with $a^\top \widehat{L}_K a \leq \lambda$, this gives

$$\|a\|_2^2 \leq \frac{\lambda}{m_K - \varepsilon_K}.$$

Substituting back,

$$a^\top L_K^\star a \leq \lambda + \varepsilon_K \frac{\lambda}{m_K - \varepsilon_K} = \lambda \frac{m_K}{m_K - \varepsilon_K}.$$

This proves the claim. □

Lemma 12 (Perturbation of task-conditioned resistance)**.** *Let*

$$\Sigma_{S,K}^\star = \frac{1}{n}\sum_{i=1}^n z_{i,K}^\star z_{i,K}^{\star\top}$$

be the ideal task-conditioned second-moment matrix under the population skill representation, and let

$$\widehat{\Sigma}_{S,K} = \frac{1}{n}\sum_{i=1}^n \widehat{z}_{i,K}\widehat{z}_{i,K}^\top$$

be the empirical task-conditioned second-moment matrix under the learned skill representation, using the same aligned skill indexing as $\widehat{L}_K$. Assume

$$\mathrm{Tr}(\Sigma_{S,K}^\star) \leq 1.$$

Define

$$\varepsilon_{\Sigma,K}^{(1)} = \|\Sigma_{S,K}^\star - \widehat{\Sigma}_{S,K}\|_*,$$

where $\|\cdot\|_*$ denotes the nuclear norm. Then

$$\mathrm{Tr}(L_K^{\star+}\Sigma_{S,K}^\star) \leq \mathrm{Tr}(\widehat{L}_K^+\widehat{\Sigma}_{S,K}) + \Delta_K,$$

where

$$\Delta_K = \frac{\varepsilon_K}{m_K(m_K - \varepsilon_K)} + \frac{\varepsilon_{\Sigma,K}^{(1)}}{m_K - \varepsilon_K}.$$

Proof. We start from

$$\mathrm{Tr}(L_K^{\star+}\Sigma_{S,K}^\star) - \mathrm{Tr}(\widehat{L}_K^+\widehat{\Sigma}_{S,K}).$$

Add and subtract $\mathrm{Tr}(\widehat{L}_K^+\Sigma_{S,K}^\star)$:

$$\begin{aligned}&\mathrm{Tr}(L_K^{\star+}\Sigma_{S,K}^\star) - \mathrm{Tr}(\widehat{L}_K^+\widehat{\Sigma}_{S,K})\\ &= \mathrm{Tr}\left((L_K^{\star+} - \widehat{L}_K^+)\Sigma_{S,K}^\star\right) + \mathrm{Tr}\left(\widehat{L}_K^+(\Sigma_{S,K}^\star - \widehat{\Sigma}_{S,K})\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathrm{Tr}(L_K^{\star+}\Sigma_{S,K}^\star) = \mathrm{Tr}(&\widehat{L}_K^+\widehat{\Sigma}_{S,K})\\ &+ \mathrm{Tr}\left((L_K^{\star+} - \widehat{L}_K^+)\Sigma_{S,K}^\star\right)\\ &+ \mathrm{Tr}\left(\widehat{L}_K^+(\Sigma_{S,K}^\star - \widehat{\Sigma}_{S,K})\right).\end{aligned}$$

We bound the two perturbation terms separately. First, by trace duality,

$$\left|\mathrm{Tr}\left((L_K^{\star+} - \widehat{L}_K^+)\Sigma_{S,K}^\star\right)\right| \leq \|L_K^{\star+} - \widehat{L}_K^+\|_{\mathrm{op}}\,\|\Sigma_{S,K}^\star\|_*.$$

Since $\Sigma_{S,K}^\star$ is positive semidefinite,

$$\|\Sigma_{S,K}^\star\|_* = \mathrm{Tr}(\Sigma_{S,K}^\star) \leq 1.$$

By Lemma 10,

$$\|L_K^{\star+} - \widehat{L}_K^{+}\|_{\mathrm{op}} \le \frac{\varepsilon_K}{m_K(m_K - \varepsilon_K)}.$$

Hence

$$\left|\mathrm{Tr}\left((L_K^{\star+} - \widehat{L}_K^{+})\Sigma_{S,K}^{\star}\right)\right| \le \frac{\varepsilon_K}{m_K(m_K - \varepsilon_K)}.$$

Second, again by trace duality,

$$\left|\mathrm{Tr}\left(\widehat{L}_K^{+}(\Sigma_{S,K}^{\star} - \widehat{\Sigma}_{S,K})\right)\right| \le \|\widehat{L}_K^{+}\|_{\mathrm{op}}\, \|\Sigma_{S,K}^{\star} - \widehat{\Sigma}_{S,K}\|_{*}.$$

By Lemma 9,

$$\|\widehat{L}_K^{+}\|_{\mathrm{op}} = \frac{1}{\mu_2(\widehat{L}_K)} \le \frac{1}{m_K - \varepsilon_K}.$$

By definition,

$$\|\Sigma_{S,K}^{\star} - \widehat{\Sigma}_{S,K}\|_{*} = \varepsilon_{\Sigma,K}^{(1)}.$$

Therefore,

$$\left|\mathrm{Tr}\left(\widehat{L}_K^{+}(\Sigma_{S,K}^{\star} - \widehat{\Sigma}_{S,K})\right)\right| \le \frac{\varepsilon_{\Sigma,K}^{(1)}}{m_K - \varepsilon_K}.$$

Combining the two bounds gives

$$\mathrm{Tr}(L_K^{\star+}\Sigma_{S,K}^{\star}) \le \mathrm{Tr}(\widehat{L}_K^{+}\widehat{\Sigma}_{S,K}) + \frac{\varepsilon_K}{m_K(m_K - \varepsilon_K)} + \frac{\varepsilon_{\Sigma,K}^{(1)}}{m_K - \varepsilon_K}.$$

This proves the lemma. □

Theorem 6 (Stability-Aware Learned-SCG Generalization Bound)**.** *Let $\mathcal{K}$ be a finite set of candidate cluster numbers. Suppose Assumption 2 holds with probability at least $1 - \delta_G$. For each $K \in \mathcal{K}$, let $\widehat{L}_K$ be the full-data empirical SCG Laplacian and let $\widehat{\Sigma}_{S,K}$ be the empirical task-conditioned skill second-moment matrix induced by the learned skill representation. Assume $0 \le \ell_a(x) \le M$. Then, with probability at least $1 - \delta - \delta_G$, the following bound holds simultaneously for all $K \in \mathcal{K}$ and all $a \in \mathbf{1}^{\perp}$ satisfying $a^{\top}\widehat{L}_K a \le \lambda$:*

$$\mathbb{E}[\ell_a(x)] \le \frac{1}{n}\sum_{i=1}^{n} \ell_a(x_i) + 2\beta\sqrt{\frac{\lambda_K}{n}\left[\mathrm{Tr}(\widehat{L}_K^{+}\widehat{\Sigma}_{S,K}) + \Delta_K\right]} + 3M\sqrt{\frac{\log(2|\mathcal{K}|/\delta)}{2n}},$$

where

$$\lambda_K = \lambda\frac{m_K}{m_K - \varepsilon_K},$$

and

$$\Delta_K = \frac{\varepsilon_K}{m_K(m_K - \varepsilon_K)} + \frac{\varepsilon_{\Sigma,K}^{(1)}}{m_K - \varepsilon_K}.$$

Equivalently, since $\varepsilon_K = C_L s(K)$,

$$\Delta_K = \frac{C_L s(K)}{m_K(m_K - C_L s(K))} + \frac{\varepsilon_{\Sigma,K}^{(1)}}{m_K - C_L s(K)}.$$

Proof. Fix a candidate $K \in \mathcal{K}$. Consider the population SCG Laplacian $L_K^{\star}$ and the corresponding task-conditioned second-moment matrix $\Sigma_{S,K}^{\star}$. For fixed K, Theorem 5 applied to the population graph gives that, with probability at least $1 - \delta_K$, for all $a \in \mathbf{1}^{\perp}$ satisfying $a^{\top}L_K^{\star}a \le \lambda_K$,

$$\mathbb{E}[\ell_a(x)] \le \frac{1}{n}\sum_{i=1}^{n} \ell_a(x_i) + 2\beta\sqrt{\frac{\lambda_K}{n}\,\mathrm{Tr}(L_K^{\star+}\Sigma_{S,K}^{\star})} + 3M\sqrt{\frac{\log(2/\delta_K)}{2n}}.$$

In practice, the smoothness constraint is imposed with respect to the empirical graph:

$$a^{\top}\widehat{L}_K a \le \lambda.$$

By Lemma 11, this implies

$$a^\top L_K^\star a \le \lambda_K, \qquad \lambda_K = \lambda \frac{m_K}{m_K - \varepsilon_K}.$$

Thus, the population-graph bound applies to every function satisfying the empirical smoothness constraint.

Next, by Lemma 12,

$$\mathrm{Tr}(L_K^{\star+} \Sigma_{S,K}^\star) \le \mathrm{Tr}(\widehat{L}_K^+ \widehat{\Sigma}_{S,K}) + \Delta_K.$$

Since the square-root function is monotone increasing on the nonnegative real line,

$$\sqrt{\mathrm{Tr}(L_K^{\star+} \Sigma_{S,K}^\star)} \le \sqrt{\mathrm{Tr}(\widehat{L}_K^+ \widehat{\Sigma}_{S,K}) + \Delta_K}.$$

Substituting this into the population-graph generalization bound yields, for fixed K,

$$\mathbb{E}[\ell_a(x)] \le \frac{1}{n}\sum_{i=1}^n \ell_a(x_i) + 2\beta\sqrt{\frac{\lambda_K}{n}\left[\mathrm{Tr}(\widehat{L}_K^+ \widehat{\Sigma}_{S,K}) + \Delta_K\right]} + 3M\sqrt{\frac{\log(2/\delta_K)}{2n}}.$$

We now make the bound uniform over all candidate cluster numbers. Set

$$\delta_K = \frac{\delta}{|\mathcal{K}|}.$$

Then

$$\log(2/\delta_K) = \log\left(\frac{2|\mathcal{K}|}{\delta}\right).$$

For each $K \in \mathcal{K}$, let $\mathcal{E}_K$ denote the event that the above generalization bound holds for that K. Since

$$P(\mathcal{E}_K^c) \le \delta_K,$$

the union bound gives

$$P\left(\bigcup_{K\in\mathcal{K}} \mathcal{E}_K^c\right) \le \sum_{K\in\mathcal{K}} P(\mathcal{E}_K^c) \le \delta.$$

Equivalently,

$$P\left(\bigcap_{K\in\mathcal{K}} \mathcal{E}_K\right) \ge 1 - \delta.$$

Let $\mathcal{E}_{\mathrm{graph}}$ denote the event in Assumption 2, namely

$$\|\widehat{L}_K - L_K^\star\|_{\mathrm{op}} \le C_L s(K), \qquad \forall K \in \mathcal{K}.$$

By assumption,

$$P(\mathcal{E}_{\mathrm{graph}}) \ge 1 - \delta_G.$$

The final bound requires both the uniform generalization event and the graph perturbation event. Therefore, by another union bound,

$$P\left(\left(\bigcap_{K\in\mathcal{K}} \mathcal{E}_K\right) \cap \mathcal{E}_{\mathrm{graph}}\right) \ge 1 - \delta - \delta_G.$$

On this joint event, the stated bound holds simultaneously for all $K \in \mathcal{K}$ and all $a \in \mathbf{1}^\perp$ satisfying $a^\top \widehat{L}_K a \le \lambda$. This completes the proof. □

Implication for cluster-number selection. Theorem 6 shows that graph estimation affects the complexity term through both the resistance perturbation Δ_K and the inflated smoothness radius λ_K. Ignoring terms independent of K, an ideal bound-aware criterion would therefore minimize

$$K_{\mathrm{BA}}^\star = \arg\min_{K\in\mathcal{K}} \left\{\lambda_K \left[\mathrm{Tr}(\widehat{L}_K^+ \widehat{\Sigma}_{S,K}) + \Delta_K\right]\right\}.$$

Both λ_K and Δ_K, however, depend on population quantities such as m_K and the unknown perturbation constants. The stability score $s(K)$ therefore serves as an observable proxy for the reliability of the estimated SCG. In our empirical pipeline, $K^\star$ is selected using this stability criterion before held-out evaluation, avoiding the use of held-out task performance or resistance for graph construction.

Table 2: Definitions of empirical SCG diagnostics.

Symbol	Definition	Meaning
$K^\star$	$K^\star = \arg\min_{K \in \mathcal{K}} \mathcal{J}_{\mathrm{stab}}(K)$.	Stability-selected cluster number before semantic filtering.
$\widetilde{K}$	Number of retained non-empty clusters after semantic filtering.	Number of skills retained as SCG nodes.
$\mathcal{J}_{\mathrm{stab}}(K)$	$(1 - \overline{\mathrm{ARI}}(K)) + \alpha_{\mathrm{stab}} \mathrm{Std}_{\mathrm{ARI}}(K) + \eta_{\mathrm{stab}} K/\bar{n}_K$.	Empirical stability-selection objective. A smaller value indicates a more stable skill partition under subsampling.
$\widehat{G}_{\mathrm{tr}}$	Binary co-activation graph constructed from the training split using the selected and filtered skill vocabulary.	Train-induced empirical SCG.
$\widehat{L}$	$\widehat{L} = \widehat{D} - \widehat{W}$.	Unnormalized Laplacian of the train-induced SCG.
$\widehat{L}^+$	Moore–Penrose pseudoinverse of $\widehat{L}$.	Resistance operator of the train-induced SCG.
U_{test}	$U_{\mathrm{test}} \in \{0,1\}^{n_{\mathrm{test}} \times \widetilde{K}}$.	Test sample–skill usage matrix obtained by matching test tokens to the frozen training-split skill vocabulary.
C	$\frac{1}{n_{\mathrm{test}}} \sum_{i=1}^{n_{\mathrm{test}}} \mathbb{I}\{\|U_{\mathrm{test}}[i]\|_1 > 0\}$.	Held-out sample coverage, i.e., the fraction of test examples activating at least one train-induced skill.
ϕ_i	$U_{\mathrm{test}}[i]/\|U_{\mathrm{test}}[i]\|_1$, for nonzero rows.	Normalized skill-usage distribution of a covered test example.
z_i	$\phi_i - \frac{1}{\widetilde{K}}\mathbf{1}$.	Centered held-out skill-usage vector.
$\widehat{\Sigma}_{\mathrm{test}}$	$\frac{1}{n_{\mathrm{valid}}} \sum_{i:\|U_{\mathrm{test}}[i]\|_1 > 0} z_i z_i^\top$.	Held-out task-conditioned skill second-moment matrix, where n_{valid} is the number of covered test examples.
$\bar{T}_L$	$\mathrm{Tr}(\widehat{L}^+)/\widetilde{K}$.	Per-node topology-only resistance diagnostic of the train-induced SCG.
$R_{\mathrm{task}}^{\mathrm{eval}}$	$\mathrm{Tr}(\widehat{L}^+ \widehat{\Sigma}_{\mathrm{test}})$.	Held-out task-conditioned resistance diagnostic.
R_{norm}	$R_{\mathrm{task}}^{\mathrm{eval}}/(\mathrm{Tr}(\widehat{\Sigma}_{\mathrm{test}}) + \epsilon)$.	Normalized held-out task-conditioned resistance, with $\epsilon > 0$ for numerical stability.
$\lambda_2^{\mathrm{norm}}$	$\lambda_2(\widehat{D}^{-1/2} \widehat{L} \widehat{D}^{-1/2})$.	Auxiliary normalized-connectivity measure of the train-induced SCG.

B METHOD

In this section, we describe how empirical SCGs are constructed from a pretrained language model. Following Section 3, we use the term *skills* to denote model-internal skill-indexed components induced from token-level predictive geometry. The construction pipeline consists of two stages. First, we induce skills by clustering token-level loss-gradient directions. Second, we construct a binary SCG from their empirical co-activation relations and compute task-conditioned SCG diagnostics. Since the induced skills are defined geometrically rather than by human annotations, we further apply a post-hoc semantic validation procedure to assess their interpretability. This procedure is used only for interpretation and does not affect skill induction, SCG construction, or cluster-number selection.

B.1 SKILL-INDEXED COMPONENT INDUCTION

Let M_θ be an autoregressive language model and $(x_{<t}, x_t)$ a token prediction instance. We compute the token-level loss $\ell_t(\theta) = -\log P_\theta(x_t \mid x_{<t})$ and stratify tokens into five loss bands: Very Low, Low, Medium, High, and Very High. In the main experiments, skills are induced from High-loss tokens, which provide a balance between informative prediction signals and noisy high-loss cases. For each selected token, we compute the loss gradient with respect to a fixed parameter subset, by default the last Transformer block, $g_t = \nabla_{\theta_{\mathcal{B}}} \ell_t(\theta)$, and normalize it as $\widetilde{g}_t = g_t/(\|g_t\|_2 + \epsilon)$. We quantify similarity between gradient directions by $S_{ij} = 1 - \arccos(\mathrm{clip}(\widetilde{g}_i^\top \widetilde{g}_j, -1, 1))/\pi$, and apply spectral clustering to induce skill-indexed components. The cluster number $K^\star$ is selected

using the stability criterion. For each candidate K, we repeatedly cluster random subsamples and measure partition consistency using pairwise adjusted Rand index scores over overlapping token instances, together with a mild fragmentation penalty. We then rerun spectral clustering on all selected tokens using $K^\star$. Finally, we apply IDF and POS filtering, retaining informative nouns, verbs, and adjectives, and remove clusters with no remaining tokens. This yields the final skill set $\widetilde{\mathcal{S}} = \{\widetilde{s}_1, \ldots, \widetilde{s}_{\widetilde{K}}\}$, where $\widetilde{K} \leq K^\star$, which is used for SCG construction.

B.2 SCG Construction and Task-Conditioned Diagnostics

Given the filtered skills $\widetilde{\mathcal{S}} = \{\widetilde{s}_1, \ldots, \widetilde{s}_{\widetilde{K}}\}$, we construct a binary skill co-activation graph over the graph-construction corpus $\mathcal{C}$. For each example $X \in \mathcal{C}$, we record which skills are activated by its tokens. Two skills are connected if they are co-activated in at least one example. This binary construction captures the existence of empirical co-activation relations rather than their frequencies. Let $\widehat{G}_M$ be the empirical SCG of model M, and let $\widehat{L}_M$ be its graph Laplacian. To relate the SCG to a downstream task T, we map each task example onto the frozen skill vocabulary induced from the graph-construction split, obtain its skill-usage profile, and compute the corresponding task-conditioned skill second-moment matrix $\widehat{\Sigma}_{T,M}$. We then use $\mathrm{Tr}(\widehat{L}_M^{+}\widehat{\Sigma}_{T,M})$ as the empirical task-conditioned resistance diagnostic. A smaller value indicates that task-induced skill demands concentrate more strongly on low-resistance modes of the empirical SCG. We also report a normalized version to control for the overall magnitude of task-induced skill variation. Table 2 provides detailed definitions of the empirical SCG diagnostics reported in the main text.

B.3 Post-hoc Semantic Interpretation

The induced skills are defined by token-level loss-gradient geometry and do not carry predefined semantic labels. To obtain human-readable descriptions, we apply a post-hoc interpretation procedure using a fixed language-model prompt. The input is a `skill_to_token` dictionary, where each skill is associated with the tokens assigned to it by gradient-based clustering. For each skill, the annotator produces a short label, a functional description, primary and secondary interpretation types, an open-ended fine-grained type, representative evidence tokens, artifact types and severity, an interpretation-confidence score, and a confidence rationale.

The prompt explicitly asks the annotator to characterize the shared prediction or functional pattern reflected by the tokens rather than merely summarize their lexical content. It also distinguishes coherent semantic patterns from template artifacts, high-frequency function words, generic relation words, subword fragments, and semantically mixed clusters. The interpretation step is strictly post hoc: it does not modify the induced skill partition, determine the cluster number, or affect SCG construction. The confidence score reflects the reliability of the natural-language interpretation based on token coherence, lexical specificity, artifact severity, and description specificity, rather than the existence or functional relevance of the underlying skill. The complete prompt used in our experiments is provided below.

B.4 Full Prompt for Post-hoc Skill Interpretation

For reproducibility, we provide the complete prompt used for post-hoc semantic interpretation of the gradient-induced skills. The prompt takes the `skill_to_token` dictionary produced by the skill-induction pipeline as input and returns structured semantic interpretations together with interpretation-confidence annotations. The following is an English translation of the prompt used in our experiments, preserving its instructions, output schema, interpretation categories, artifact criteria, confidence-scoring rules, and quality-control requirements.

Prompt:

> **Task Name**
>
> Gradient-Induced Skill Post-hoc Interpretation with Confidence Scoring
>
> **Your Role**
>
> You are an interpretability research assistant for large language models. Given a `skill_to_token` dictionary obtained from token-level loss-gradient clustering, your

task is to provide a post-hoc semantic interpretation for each skill node and assign an interpretation-confidence score.

Important Premise

These skills have already been defined from the model's internal predictive behavior: tokens assigned to the same skill are grouped together because they have similar token-level loss-gradient directions. Your task is not to reconstruct the skills or validate the SCG, but to provide human-readable post-hoc interpretations of these components that have already been induced from gradient geometry.

In other words:

- the functional basis of each skill comes from gradient-based clustering;
- your interpretation is used only for post-hoc interpretation;
- the confidence score measures whether the natural-language interpretation is clear, stable, and credible, rather than whether the skill itself exists.

Input

You will be given a JSON dictionary in the following format:

```
{
  "skill_id_1": ["token1", "token2", "token3", ...],
  "skill_id_2": ["token1", "token2", "token3", ...]
}
```

Each key is a skill ID, and each value is the list of tokens associated with that skill. The token list may contain repeated tokens, case variants, BPE subword fragments, high-frequency template words, or semantically mixed items.

Overall Goal

For each skill, output:

1. a short English label;
2. an English functional description;
3. a primary interpretation category;
4. optional secondary interpretation categories;
5. an open-ended fine-grained type description;
6. representative tokens supporting the interpretation;
7. possible artifact types;
8. artifact severity;
9. an interpretation-confidence score;
10. a rationale for the confidence score.

Core Requirement

Do not simply restate the token list. Instead, interpret the shared *prediction pattern* or *functional pattern* suggested by these tokens. For example:

```
water, fish, freshwater, aquatic, swim
-> water and aquatic-environment prediction pattern

heat, temperature, thermal, hot
-> heat- and temperature-related physical-state prediction pattern

greater, highest, smaller, increase, decrease
-> quantitative comparison or magnitude-change prediction pattern

answer, answer, answer
-> answer-format or output-template artifact

have, has, are, be
-> grammatical/function-word prediction pattern

ing, tion, subword fragments
-> tokenization artifact
```

Output Language

Except for the JSON field names, all contents of `short_label`, `functional_description`, `open_ended_type`, `confidence_rationale`, and other natural-language fields must be written in English.

Output Format

Return a JSON array. Each element corresponds to one skill. Strictly use the following fields:

```
[
  {
    "skill_id": "...",
    "short_label": "...",
    "functional_description": "...",
    "primary_interpretation_type": "...",
    "secondary_interpretation_types": ["..."],
    "open_ended_type": "...",
    "evidence_tokens": ["...", "...", "..."],
    "artifact_types": ["..."],
    "artifact_severity": "...",
    "interpretation_confidence": 0.0,
    "confidence_rationale": "..."
  }
]
```

Field Definitions

1. `skill_id`

Preserve the skill ID exactly as provided in the input. Do not rewrite or rename it.

2. `short_label`

Provide a short English label, typically two to six words. The label should summarize the primary prediction pattern represented by the skill.

Examples:

- Water and aquatic contexts
- Heat and temperature
- Biological cells
- Comparative magnitude
- Answer-format pattern
- Generic function words

3. `functional_description`

Use one sentence to describe the predictive function that the skill may support. The description should characterize a *prediction pattern* or *functional pattern*, rather than merely list a semantic topic.

Recommended formulations include:

> "This skill appears to support predictions involving ..."

or

> "This skill appears to capture a ... prediction pattern."

4. `primary_interpretation_type`

Choose the single most appropriate primary category from the list below. If none of the predefined categories is suitable, select `other_or_open_category`. Note that this category is used only for post-hoc statistical analysis and should not directly determine the confidence score.

Available categories:

```
semantic_domain
scientific_concept
physical_process
biological_process
environmental_or_ecological_context
quantitative_or_comparative_relation
```

```
causal_or_change_relation
action_or_process_pattern
object_or_material_property
spatial_or_temporal_relation
grammatical_or_function_word_pattern
template_or_answer_format
subword_or_tokenization_artifact
mixed_or_low_coherence
other_or_open_category
```

5. `secondary_interpretation_types`

If the skill involves multiple functional patterns, select zero to three secondary categories from the same category list. If no secondary category is needed, return an empty list `[]`.

Do not add secondary categories merely to fill the field.

6. `open_ended_type`

Use a short English phrase to freely describe a more fine-grained functional type for the skill. This field is important because the predefined categories may not cover every possible pattern.

Examples:

- heat-related physical state and temperature-driven change
- aquatic ecosystem and water-environment entities
- answer-slot and response-format lexical pattern
- generic auxiliary and possession-related function words

7. `evidence_tokens`

List five to twelve representative tokens that most strongly support the interpretation.

Prioritize:

- frequent tokens;
- semantically clear tokens;
- tokens that cover the major pattern of the skill;
- nonredundant tokens rather than repeated instances.

If the total number of tokens in the skill is small, fewer than five evidence tokens are allowed.

8. `artifact_types`

Select all applicable artifact types. If no clear artifact is present, return `["none"]`. The field may contain multiple artifact types; `["none"]` should be used only when no obvious artifact is present.

Available artifact types:

```
none
template_or_answer_format
high_frequency_function_word
generic_verb_or_relation_word
subword_fragment
mixed_semantics
other_artifact
```

Use the following rules:

- If a skill is dominated by words such as *answer*, *option*, *choice*, *correct*, or *question*, it should generally be marked as `template_or_answer_format`.
- If a skill is dominated by words such as *have*, *has*, *be*, *are*, *is*, *do*, or *does*, it should generally be marked as `high_frequency_function_word`.
- If a skill is dominated by generic verbs or relation words such as *use*, *make*, *take*, *put*, *get*, *go*, *come*, or *provide*, it should generally be marked as `generic_verb_or_relation_word`.
- If a skill contains many BPE subword fragments, incomplete word pieces, or tokens that are difficult to interpret independently, mark it as `subword_fragment`.
- If the tokens in a skill clearly come from multiple unrelated semantic domains, mark it as `mixed_semantics`.

9. `artifact_severity`

Choose one of the following four levels:

```
none
low
medium
high
```

Use the following criteria:

- **none**: There are essentially no obvious artifacts. The tokens are semantically concentrated and interpretable.
- **low**: A small amount of noisy content is present, but it does not substantially affect the main interpretation.
- **medium**: Template words, high-frequency words, subword fragments, or semantic mixing are clearly present. An interpretation can still be provided, but it should be treated cautiously.
- **high**: The skill is dominated by template words, high-frequency function words, subword fragments, or strongly mixed tokens, making it unsuitable for a clear semantic-skill interpretation.

10. `interpretation_confidence`

Assign a score between 0 and 1, rounded to two decimal places. The score represents the reliability of the natural-language interpretation you have provided.

Suggested scoring guidelines:

- **0.85–1.00**: The tokens are highly coherent, the interpretation is clear, and artifacts are minimal.
- **0.70–0.84**: Most tokens support the same interpretation, with only minor noise or generic terms.
- **0.50–0.69**: A meaningful semantic pattern exists, but substantial mixing is present and the interpretation is only moderately reliable.
- **0.30–0.49**: The skill is dominated by template words, high-frequency words, generic verbs, or mixed semantics, and the interpretation is relatively weak.
- **0.00–0.29**: The skill is dominated by subword fragments, severe semantic mixing, or is otherwise nearly uninterpretable.

When assigning the confidence score, jointly consider:

- **token coherence**: whether the tokens concentrate around the same semantic or functional pattern;
- **lexical specificity**: whether the tokens are specific content words rather than overly generic terms;
- **artifact severity**: whether template words, high-frequency words, subword fragments, or semantic mixing are present;
- **description specificity**: whether the functional description is specific rather than vague.

Do not automatically assign a high or low confidence score based on the `primary_interpretation_type`. The category is used only for post-hoc statistics. Confidence should instead depend on token coherence, clarity, and artifact severity.

11. `confidence_rationale`

Use one or two sentences to explain why the confidence score was assigned.

Explicitly indicate:

- which tokens support the interpretation;
- whether noise or artifacts are present;
- why the confidence is high, medium, or low.

Special-Case Handling

1. Clearly Semantic Skills

If the tokens are clearly concentrated within one semantic domain or scientific concept, provide a clear functional description and assign a relatively high confidence score.

Example input tokens:

```
water, fish, freshwater, aquatic, swim
```

The output should be similar to:

```
{
  "short_label": "Water and aquatic contexts",
  "functional_description":
    "This skill appears to support predictions involving water,
     aquatic organisms, freshwater environments, and swimming-related
     contexts.",
  "primary_interpretation_type":
    "environmental_or_ecological_context",
  "secondary_interpretation_types": ["semantic_domain"],
  "open_ended_type":
    "aquatic ecosystem and water-environment prediction pattern",
  "artifact_types": ["none"],
  "artifact_severity": "none",
  "interpretation_confidence": 0.91
}
```

2. Template- or Answer-Format Skills

Do not force such skills into interpretations as abstract reasoning abilities. Instead, label them as `template_or_answer_format` and reduce the interpretation-confidence score.

Example input tokens:

```
answer, answer, answer, correct, option
```

The output should be similar to:

```
{
  "short_label": "Answer-format pattern",
  "functional_description":
    "This skill appears to capture answer-slot or response-format
     tokens rather than a specific semantic concept.",
  "primary_interpretation_type": "template_or_answer_format",
  "secondary_interpretation_types": [],
  "open_ended_type":
    "answer-slot and output-format lexical pattern",
  "artifact_types": ["template_or_answer_format"],
  "artifact_severity": "high",
  "interpretation_confidence": 0.36
}
```

3. High-Frequency Function-Word Skills

If the tokens mainly consist of words such as *have*, *has*, *be*, *are*, *do*, or *does*, do not interpret the skill as a specific knowledge skill. Instead, characterize it as a grammatical or function-word prediction pattern.

4. Generic-Verb or Relation-Word Skills

If the tokens mainly consist of words such as *use*, *make*, *take*, *put*, *get*, *go*, *come*, *provide*, or *require*, generally mark the skill as `generic_verb_or_relation_word`. These tokens may still represent a prediction pattern, but the interpretation-confidence score should generally not be high unless the tokens are highly coherent and their functional relation is clear.

5. Semantically Mixed Skills

If the tokens come from multiple unrelated semantic domains, select `mixed_or_low_coherence` as the primary interpretation type and mark `mixed_semantics` under `artifact_types`. The `functional_description` should explicitly acknowledge that the skill lacks a single coherent semantic interpretation.

6. Subword-Fragment Skills

If most tokens are incomplete word fragments, select `subword_or_tokenization_artifact` as the primary interpretation type and mark `subword_fragment` under `artifact_types`. Do not over-interpret such skills.

Quality-Control Requirements

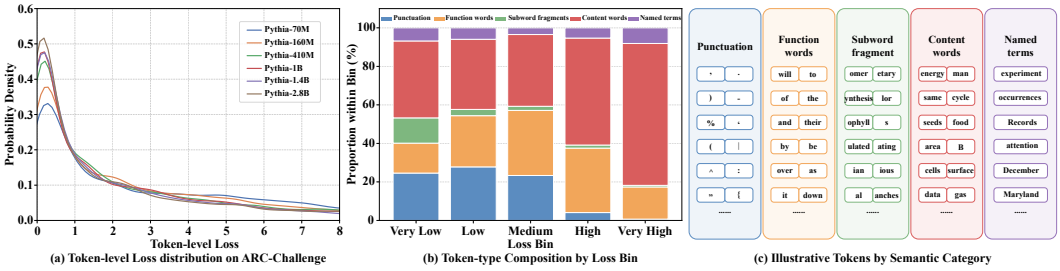

Figure 4: **Token-level loss stratification and semantic composition.** (a) Token-level loss distributions on ARC-Challenge for Pythia models from 70M to 2.8B. (b) Token-type composition of Pythia-2.8B tokens after sorting by loss and splitting them into five equal-sized bins: Very Low, Low, Medium, High, and Very High. Each bin is decomposed into Punctuation, Function words, Subword fragments, Content words, and Named terms. (c) Illustrative tokens from each semantic category.

- Do not interpret every skill as a high-level cognitive capability.
- Do not force template words, function words, or subword fragments into semantic-skill interpretations.
- Do not make strong interpretations based on a single token.
- If a skill is semantically mixed, explicitly acknowledge `mixed_or_low_coherence`.
- If the interpretation is uncertain, lower the `interpretation_confidence` score.
- The `functional_description` should be concise, specific, and readable.
- All `evidence_tokens` must come from the input token list. Do not invent tokens.
- The output must be a valid JSON array. Do not add Markdown, explanatory text, or code fences.

Final Task

Given the `skill_to_token` JSON dictionary provided next, generate the post-hoc interpretation results for every skill using the JSON-array format specified above.

C Experiment

C.1 Token-Loss Stratification and Token-Type Composition

Figure 4 motivates the loss-stratified skill induction used in our main experiments. Figure 4(a) shows that token-level losses on ARC-Challenge are strongly right-skewed across Pythia scales, with larger models placing more probability mass in the lower-loss region. Figure 4(b) further shows that token-type composition varies systematically across loss bands. Lower-loss bands contain substantial proportions of punctuation and function words, whereas the High-loss band contains markedly fewer punctuation tokens and a larger proportion of content words. Figure 4(c) provides representative tokens from these categories. These observations indicate that token loss reflects not only prediction difficulty but also the types of lexical signals contributing to prediction. We therefore construct the main SCGs from High-loss tokens, reducing the influence of frequent or semantically uninformative tokens on gradient-based skill induction while retaining prediction-relevant content.

C.2 SCG Diagnostics across Qwen2.5 Model Scales

To complement the main Pythia scaling analysis, we apply the same SCG construction and evaluation pipeline to Qwen2.5 models at 0.5B, 1.5B, and 3B scales on ARC-Challenge and SciQ. Table 3 summarizes the resulting structural and task-conditioned diagnostics across the three model scales.

Table 3 summarizes the results. The Qwen2.5 models exhibit substantial variation in both the number of induced skills and their task-conditioned resistance across scales and datasets. Because only three Qwen2.5 scales are included here, we treat these results as complementary evidence.

Table 3: **Empirical SCG diagnostics for Qwen2.5 models on ARC-Challenge and SciQ.** $K^\star$ is the stability-selected cluster number before filtering, $\widetilde{K}$ is the number of retained SCG nodes, $\mathcal{J}_{\mathrm{stab}}$ is the stability-selection objective (smaller is more stable), and R_{norm} is the normalized held-out task-conditioned resistance.

Dataset	Model	$K^\star$	$\widetilde{K}$	$\mathcal{J}_{\mathrm{stab}}$	R_{norm}
ARC-Challenge	Qwen2.5-0.5B	182	175	0.7080	0.007944
	Qwen2.5-1.5B	160	152	0.7685	0.008535
	Qwen2.5-3B	185	178	0.7335	0.007810
SciQ	Qwen2.5-0.5B	202	191	0.7291	0.005553
	Qwen2.5-1.5B	291	270	0.7262	0.004270
	Qwen2.5-3B	260	250	0.7638	0.004482

Table 4: **Functional validation of induced skill subspaces across Qwen2.5 model scales on ARC-Challenge and SciQ.** We report the increase in per-token gold-answer NLL (ΔNLL$\uparrow$) after ablating task-relevant skill subspaces (Task) or same-rank random subspaces (RandSub).

Dataset	Intervention	0.5B	1.5B	3B
ARC-Challenge	Task	2.611	3.858	2.653
	RandSub	0.031	0.077	0.078
SciQ	Task	2.155	0.338	1.354
	RandSub	-0.028	-0.185	0.325

C.3 Functional Validation across Qwen2.5 Model Scales

We further perform the same intervention-based functional validation on Qwen2.5 models. We compare ablation of task-relevant skill subspaces with same-rank random representation subspaces on ARC-Challenge and SciQ. As shown in Table 4, task-relevant skill subspace ablation produces a larger increase in per-token gold-answer NLL than the random-subspace control in all six model–task settings. These results provide complementary evidence that the induced skill directions capture functionally relevant representation structure in Qwen2.5 models.

C.4 SCG-Guided Fine-Tuning Data Selection

This appendix provides implementation details for the SCG-guided fine-tuning experiment reported in the main text. The experiment keeps the supervised fine-tuning procedure fixed within each model–task setting and varies only the rule used to select a limited number of fine-tuning examples. The train-induced SCG and skill vocabulary are constructed before data selection and remain frozen throughout the experiment.

C.4.1 Experimental Setting

We separate the data used for graph construction, fine-tuning selection, and final evaluation. Let $\mathcal{D}_{\mathrm{graph}}$ denote the split used to induce skills and construct the empirical SCG, $\mathcal{D}_{\mathrm{pool}}$ the candidate pool from which fine-tuning examples are selected, and $\mathcal{D}_{\mathrm{test}}$ the held-out evaluation split. The test split is never used for example selection or hyperparameter tuning. For each model–task setting, all selection strategies operate on the same frozen SCG, the same frozen skill vocabulary, and the same candidate pool, and select the same number n_{ft} of examples. Each strategy starts fine-tuning from a fresh copy of the same pretrained model and uses the same fine-tuning budget within that setting.

Fine-tuning uses standard supervised causal language modeling with loss computed only on the target answer span. For a prompt p and target answer $y = (y_1, \ldots, y_m)$, the objective is

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = -\sum_{j=1}^{m} \log P_\theta(y_j \mid p, y_{<j}).$$

Prompt tokens are retained as context but masked from the training loss. In our implementation, we fine-tune the last Transformer block, the final layer normalization, and the output head when appli-

Table 5: **Fine-tuning data-selection strategies.** All strategies select the same number of examples from the same valid candidate pool within each model–task setting.

Strategy	Selection rule	Role in the comparison
Random	Uniformly sample examples from the valid candidate pool.	Provides a non-structured data-selection baseline.
High-LM-Loss	Select examples with the largest full-text language-model loss $\ell_{\mathrm{LM}}(x)$.	Controls for conventional hard-example selection based on predictive difficulty.
Skill Coverage	Greedily select examples whose active skill sets introduce the largest number of previously uncovered skills; resistance is used only as a secondary tie-breaking signal when needed.	Tests whether broad coverage of the induced skill vocabulary is sufficient without directly optimizing SCG resistance.
SCG High-R	Select examples with the largest sample-level SCG resistance $r(x)$.	Tests whether examples whose task-induced skill demands lie more strongly on high-resistance SCG modes provide a useful structural selection signal.
SCG Low-R	Select examples with the smallest sample-level SCG resistance $r(x)$.	Provides the reverse control for the resistance-based selection criterion.

cable, while earlier Transformer blocks remain frozen. The remaining optimization configuration is kept fixed across selection strategies within each model–task setting.

C.4.2 Candidate Pool Construction

Fine-tuning examples are selected from a candidate pool that is disjoint from both the graph-construction examples and the held-out test split. The default candidate pool is the validation split. When the validation split is too small to support the required selection budget, we enlarge the pool with auxiliary examples that are not used for SCG construction:

$$\mathcal{D}_{\mathrm{pool}} = \mathcal{D}_{\mathrm{val}} \cup \mathcal{D}_{\mathrm{aux}}.$$

To prevent overlap with the graph-construction data, auxiliary examples whose identifiers occur in the graph-construction set are removed:

$$\mathcal{D}_{\mathrm{aux}} = \{x \in \mathcal{D}_{\mathrm{candidate}} : \mathrm{id}(x) \notin \mathcal{I}_{\mathrm{graph}}\},$$

where $\mathcal{I}_{\mathrm{graph}}$ contains the identifiers of examples contributing tokens to SCG construction. Any overlap with $\mathcal{D}_{\mathrm{test}}$ is also removed. When the validation split is sufficiently large, we simply use $\mathcal{D}_{\mathrm{pool}} = \mathcal{D}_{\mathrm{val}}$.

Each candidate is then projected onto the frozen skill vocabulary. Let $u_x \in \{0,1\}^K$ denote the binary skill-usage vector of example x. We retain candidates with $\|u_x\|_1 > 0$, ensuring that every selectable example activates at least one skill induced from the graph-construction split. All selection strategies operate on this same valid candidate set.

C.4.3 Selection Strategies

For each valid candidate x, we compute the quantities required by the selection strategies. Its binary skill-usage vector u_x is normalized to a skill-attribution vector ϕ_x and centered as z_x. Given the Laplacian pseudoinverse $\widehat{L}^+$ of the frozen train-induced SCG, the sample-level resistance is

$$r(x) = z_x^\top \widehat{L}^+ z_x.$$

Larger $r(x)$ indicates that the skill demands of the example place more weight on high-resistance modes of the empirical SCG.

We additionally compute the mean next-token language-model loss

$$\ell_{\mathrm{LM}}(x) = -\frac{1}{T-1}\sum_{t=1}^{T-1} \log P_\theta(x_{t+1} \mid x_{\le t}),$$

using the full-text view of the candidate example. The skill-usage vector u_x is also used to measure skill coverage. These quantities define the five selection strategies used in the main experiment.