SIGMA: LABEL-EFFICIENT TRAINING OF LANGUAGE MODELS TO FOLLOW CHANGING STRUCTURED GUIDANCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models offer a flexible interface for task-specific decisions, yet accuracy under fixed instructions does not establish reliable behavior when those instructions change. We investigate how limited task annotations can teach models to follow updates that change the required decision while remaining stable under changes that preserve meaning. We introduce SIGMA, which pairs each labeled input with two versions of structured guidance and supervises both correct answers and the expected relationship between their predictions. By aligning the relative scores of all candidate outputs across guidance versions, SIGMA reuses existing annotations to learn how decisions should respond to guidance edits. On BANKING77, CLINC150, and HWU64 with Qwen2.5-7B-Instruct, SIGMA trained on four examples per intent exceeds the mean accuracy of supervised low-rank fine-tuning using sixteen examples per intent under unseen guidance combinations on every dataset, using 75% fewer training annotations with the same development set. The results support supervision across guidance versions as a way to improve annotation efficiency and handle unseen combinations of updates without further training.

1 INTRODUCTION

Language models in task-oriented applications must make accurate decisions from few labeled examples while following domain-specific requirements. Intent descriptions and routing instructions offer a practical interface for specifying these requirements, with refined descriptions improving few-shot intent classification (Park et al., 2025). In applications such as customer support, this interface must remain reliable as requirements change.

The central difficulty is that *correctness under a fixed task specification does not establish the ability to use that specification*. Under fixed guidance, following current requirements and relying on familiar input–output associations can produce identical answers. As illustrated in Figure 1, the same customer request may require a different destination after a routing update, whereas an equivalent restatement should preserve the decision. Reliable guidance use therefore requires both responsiveness to relevant changes and stability under irrelevant ones.

Existing methods improve adaptation through contrastive learning (Zhang et al., 2021), feedback-driven prompt optimization (Agrawal et al., 2026), and joint prompt–parameter optimization (Bo et al., 2026). These advances improve task accuracy, but that objective alone does not ensure reliable responses to later revisions. More directly, KAFT trains contextual controllability alongside robustness to irrelevant information (Li et al., 2023), and Context-DPO improves adherence to context over conflicting parametric knowledge (Bi et al., 2025). Symbol tuning promotes learning arbitrary label mappings (Wei et al., 2023), while recent permutation-aware training reduces dependence on answer presentation (Guda et al., 2025; Zheng et al., 2026). These studies establish that context following and presentation robustness can be learned. How limited task-specific supervision supports *selective responses to evolving guidance*, especially when several changes occur together, remains an open question.

We ask: *Can a small set of task annotations teach a language model to follow changing guidance and generalize this behavior to unseen combinations of updates?* We introduce **SIGMA**, which uses the expected relationship between decisions before and after a guidance change as a training signal. For the same labeled input, it learns when the prescribed output should change and when predictions should be preserved, jointly constraining the relative evidence for all candidate outputs.

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.

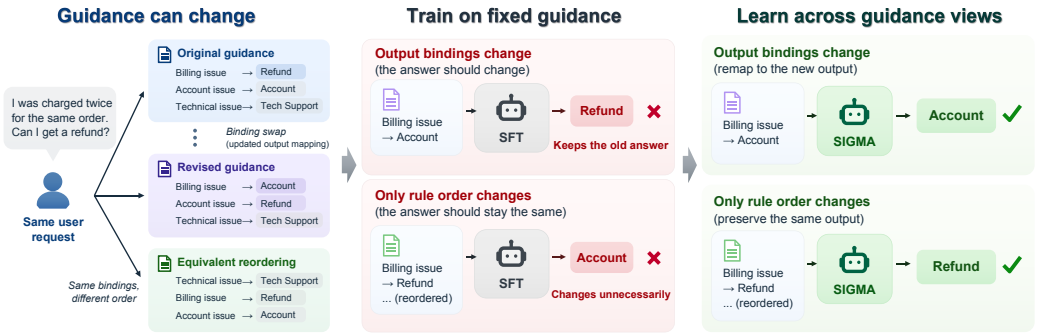


Figure 1: **Correct answers under fixed guidance can mask unreliable responses to edits.** For the same request, changing its output binding should change the answer, whereas reordering equivalent rules should preserve it. Single-view SFT can fail in either direction: retaining an outdated answer or changing an answer unnecessarily. The figure contrasts these schematic failures with the desired responses.

Each annotation thus supervises both task correctness and guidance use; the adapted model then scores outputs under current guidance without further training.

Our contributions are threefold:

- **Insight into guidance use.** We show that correct answers under fixed guidance can coexist with unstable responses to revisions (Section 3). A fixed-feature proxy with noisy responses characterizes when consistency across guidance versions reduces estimation variance and when the bias from imperfect constraints outweighs this reduction (Section 4.4).
- **Transferable adaptation.** SIGMA improves generalization to unseen guidance compositions across three backbones. Controls using identical paired views and token-matched fine-tuning assess gains beyond additional views or computation (Section 5.2; Appendix D.6).
- **Training-label efficiency.** On BANKING77, CLINC150, and HWU64 with Qwen2.5-7B-Instruct, SIGMA improves mean composition accuracy by **6.63 percentage points** over 10-shot LoRA-SFT. Four-shot SIGMA also exceeds the mean composition accuracy of 16-shot LoRA-SFT on all three datasets: **75% fewer training annotations**, with the same development set (Section 5.3).

2 Related Work

Prompt and parameter adaptation. Prompt optimization has progressed from joint instruction–demonstration search in MIPRO (Opsahl-Ong et al., 2024) to reflective evolution in GEPA (Agrawal et al., 2026); MetaTuner (Bo et al., 2026) further co-generates prompts and parameters. StructTuning (Liu et al., 2025) uses knowledge hierarchies to organize domain training. These approaches improve task adaptation, but their objectives do not specify how predictions should transform under prescribed guidance edits. SIGMA trains these relations explicitly and evaluates guidance updates with frozen adapted weights.

Few-shot intent classification. Contrastive representation learning and utterance–label matching (Zhang et al., 2021; Yehudai et al., 2023) address annotation scarcity, while ICDA (Lin et al., 2023) augments training data. Retrieval-based ICL (Milios et al., 2023) shifts adaptation into context; DLR (Park et al., 2025) refines ambiguous label names, and FCSLM (Chen & Wang, 2025) combines small-model prediction with selective LLM reasoning. Their semantic classification objectives do not directly constrain responses to changed output bindings. SIGMA reuses each annotation across guidance views, testing label efficiency under a fixed, complete candidate inventory. Qwen3-Reranker and other semantic classifiers (Zhang et al., 2025; Aarab, 2026) provide strong references when coupled with exact binding lookup.

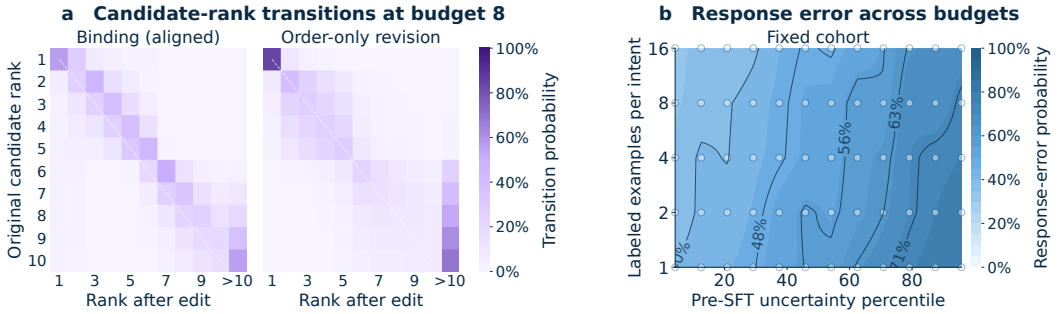


Figure 2: **Guidance sensitivity after single-view SFT. (a)** Candidate-rank transitions at budget 8, with binding edits aligned to original coordinates; diagonal entries indicate unchanged ranks. **(b)** Response error versus reference uncertainty and label budget on a fixed cohort. Markers locate observed bin rates; colors and contours interpolate their rates in percentile and log-budget coordinates.

Context following and permutation consistency. KAFT (Li et al., 2023) and Context-DPO (Bi et al., 2025) train context-faithful responses, while symbol tuning (Wei et al., 2023) teaches arbitrary label mappings. Beyond augmentation, R-Drop (Liang et al., 2021) matches dropout-induced distributions; recent PBM-based LoRA (Guda et al., 2025) and PA-GRPO (Zheng et al., 2026) explicitly train consistency across option permutations. These permutation-consistency objectives address option debiasing; structured guidance updates also require responses to changes in output bindings. SIGMA couples supervised local binding swaps and equivalent rule reorderings through centered-score alignment over all candidates, then tests unseen edit compositions. Its fixed-feature analysis builds on consistency and equivariant-learning theory (Yang et al., 2023; Elesedy & Zaidi, 2021), identifying when variance reduction outweighs mismatch bias without treating edited views as independent labels.

3 Preliminary Study

This preliminary study examines whether accurate predictions under a single guidance view are sufficient evidence of reliable guidance use. We probe guidance use through two edits: swapping output bindings should remap predictions, while reordering equivalent rules should preserve them. We test whether standard fine-tuning learns these responses by examining candidate rankings and response errors after guidance edits.

Diagnostic. We train a single-view supervised fine-tuning (SFT) baseline only under original guidance, with $1, 2, 4, 8$, or 16 labels per intent. We then freeze the model and construct both revisions separately from the original guidance. Rank comparisons undo the known binding permutation. This alignment tests whether candidate preferences are preserved after accounting for the required output remapping. Across budgets, response errors use a fixed cohort that answers correctly under original guidance at every budget. Keeping cohort membership fixed controls for changes in which inputs are answered correctly as budgets increase. Appendix C.7 specifies the cohorts, metrics, and aggregation.

Findings. At budget 8, exact top-10 rank retention is 27.3% after binding alignment and 26.9% after rule reordering. The latter retains the top-ranked answer in 84.2% of cases, illustrating how Top-1 can conceal changes among alternatives. A response error means either failing to follow the binding swap or changing the answer under reordering. Grouping inputs by pre-SFT predictive uncertainty, its rate rises from about 36% in the lowest group to 69% in the highest. Larger budgets generally reduce error, with some local reversals. The uncertainty gradient is descriptive.

Implication

These findings show that correct answers under one guidance view can coexist with unstable responses across views. This motivates training on how predictions should respond to guidance edits.

4 METHOD

SIGMA (**S**tructured **I**nterventions for **G**uidance-aware **M**odel **A**daptation) learns how predictions should respond to guidance edits. Each labeled input is paired with an edited guidance view whose effect on the correct output is known: local binding swaps require predictions to be remapped, while equivalent rule reorderings require them to remain unchanged. Training combines supervision on both views with a penalty that aligns their centered candidate scores. At inference, the adapted model scores outputs under the current guidance with frozen parameters (Figure 3).

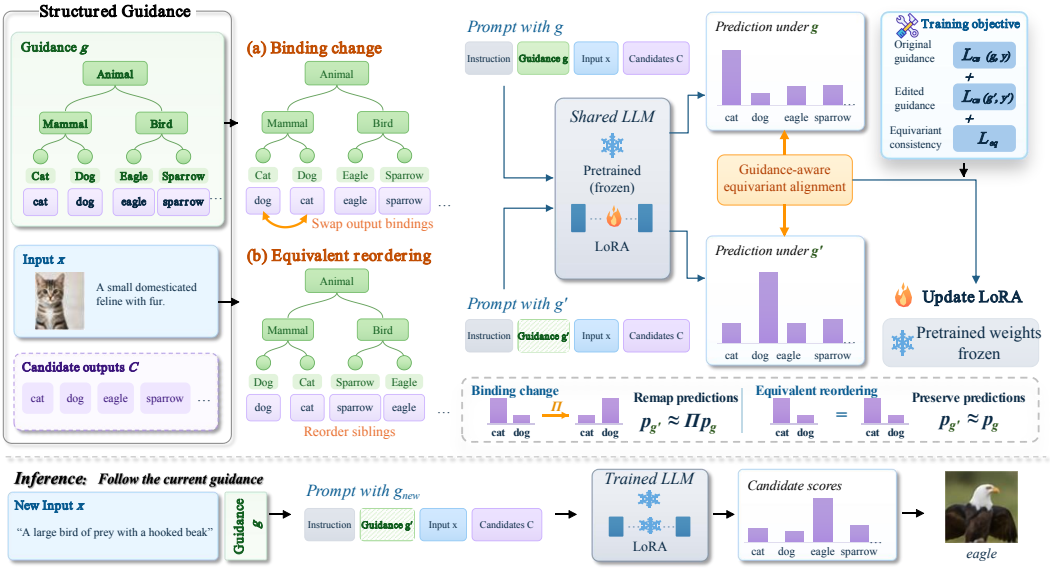


Figure 3: **SIGMA training and inference. Training (top):** input x and candidates $\mathcal{C}$ are shared across guidance g and its edited view g'. **(a)** Binding swaps require remapped predictions; **(b)** equivalent reordering requires unchanged predictions. Paired cross-entropy and centered-score alignment update only LoRA parameters (Eq. (5)). **Inference (bottom):** the frozen adapted model scores all candidates under g_{new} (top-1 output shown). Animal images are illustrative; model inputs and outputs are text. The diagram's $\mathbf{p}_g$ and Π correspond to $\mathbf{q}_g$ and P_e in the text.

4.1 STRUCTURED GUIDANCE INTERVENTIONS

Task and interface. Let $\mathcal{D} = \{(x_i, z_i)\}_{i=1}^n$ contain task inputs and semantic labels in $\mathcal{Z}$. A rooted tree $\mathcal{T}$ places labels at its leaves and semantic groups at its internal nodes. Guidance g specifies this tree, node descriptions, and a bijection $a_g : \mathcal{Z} \to \mathcal{C}$ assigning each label an output string. The ordered set $\mathcal{C}$ contains all task outputs and is fixed across inputs and guidance versions, with $m = |\mathcal{C}| \geq 2$.

All training instances share base guidance g_0 with a fixed child order. Under guidance g, the required answer is $y_i = a_g(z_i)$: changing a binding can change this answer while preserving the input's semantic class. Given the input, guidance, and full candidate inventory, the model directly scores candidates without an intermediate semantic prediction. SIGMA-F (Factorized) predicts the semantic class and applies exact binding lookup (Appendix A.3).

Local binding swaps. Let $\mathcal{V}$ contain non-root nodes with at least two leaf children. A binding edit uniformly samples a node v from $\mathcal{V}$ and then an unordered pair of its leaf children. It exchanges their bound outputs c_a, c_b and leaves all other bindings unchanged. Denoting this transposition of $\mathcal{C}$ by π_e, we regenerate the guidance as

$$g' = T_e(g), \qquad a_{g'} = \pi_e \circ a_g, \qquad y_i' = \pi_e(y_i). \tag{1}$$

Here T_e transforms the guidance document according to edit e. Swaps are restricted to siblings under a non-root parent, preserving the tree, semantic descriptions, and candidate order. Both sampling stages are independent of the instance label and model predictions.

Equivalent views and sampling. A representation edit reverses child order at every internal node in the rendered tree. Sibling order does not indicate rule priority, so all relations and bindings are preserved, with $\pi_e = \mathrm{id}_{\mathcal{C}}$. Odd training steps assign each instance a sampled binding edit; even steps apply the representation edit. If $\mathcal{V}$ is empty, binding steps use the identity edit $g' = g$. Invalid edited documents also revert to the identity, keeping both targets well defined. Appendix A.1 details the sampling law, rendering, validation, and target-change rate.

4.2 Equivariant Fine-Tuning

Preserving semantic uncertainty. A binding edit changes the output assigned to a semantic class while preserving the evidence for that class. We express this prior through a semantic posterior $r(z \mid x)$, assumed invariant to both binding and representation edits. The resulting output distribution satisfies

$$q^\star_g(c \mid x) = r(a_g^{-1}(c) \mid x), \qquad \mathbf{q}^\star_{g'} = P_e \mathbf{q}^\star_g. \tag{2}$$

The permutation matrix $P_e \in \mathbb{R}^{m \times m}$ acts as $[P_e \mathbf{v}]_c = \mathbf{v}_{\pi_e^{-1}(c)}$, with all vectors in the fixed candidate order. This relation specifies how probabilities over all candidates should change, extending supervision beyond the paired target labels.

Candidate scoring. Let $\widetilde{x}_g = \mathrm{Render}(x, \mathcal{C}, g)$ contain the task instruction, guidance, input, and candidates. The language model p_θ scores each candidate sequence $c = (c_1, \ldots, c_{|c|})$, including its end-of-answer token, by its average token log probability:

$$s_\theta(c \mid x, g) = \frac{1}{|c|} \sum_{t=1}^{|c|} \log p_\theta(c_t \mid \widetilde{x}_g, c_{<t}), \qquad \mathbf{q}_g = \mathrm{softmax}(\mathbf{s}_g). \tag{3}$$

Here $\mathbf{s}_g \in \mathbb{R}^m$ stacks candidate scores and $q_\theta(c \mid x, g) = [\mathbf{q}_g]_c$. Token averaging and unit softmax temperature are fixed. We omit the fixed $\mathcal{C}$ from conditionals and suppress the dependence of $\mathbf{s}_g, \mathbf{q}_g$ and the centered scores $\mathbf{u}_g$ below on x and θ; terms indexed by g_i use x_i.

Aligning relative evidence. Adding the same constant to every score leaves softmax unchanged. We therefore center each score vector before comparing views:

$$\mathbf{u}_g = \mathbf{s}_g - \frac{\mathbf{1}^\top \mathbf{s}_g}{m} \mathbf{1}, \qquad \mathbf{q}_{g'} = P_e \mathbf{q}_g \iff \mathbf{u}_{g'} = P_e \mathbf{u}_g. \tag{4}$$

Here $\mathbf{1} \in \mathbb{R}^m$ is the all-ones vector. Centering removes the common shift, making aligned distributions equal exactly when their centered scores match. We penalize $\|\mathbf{u}_{g'} - P_e \mathbf{u}_g\|_2^2$, the squared discrepancy in relative candidate evidence after alignment, including low-probability alternatives. Appendix B.1 gives the pairwise log-odds characterization and orthogonal-projection interpretation.

Paired objective. For instance (x_i, z_i) with base guidance $g_i = g_0$, let $g_i' = T_e(g_i)$ and $y_i = a_{g_i}(z_i)$. Its paired loss is

$$\begin{aligned} \ell_\theta(i, e) = & -\tfrac{1}{2} \log q_\theta(y_i \mid x_i, g_i) - \tfrac{1}{2} \log q_\theta(\pi_e(y_i) \mid x_i, g_i') \\ & + \frac{\lambda}{2m} \left\| \mathbf{u}_{g_i'} - P_e \mathbf{u}_{g_i} \right\|_2^2, \end{aligned} \tag{5}$$

where $\lambda \geq 0$ is the only added loss coefficient. The supervised terms favor each view's correct candidate, while the residual aligns relative scores across all candidates. Both branches share parameters and receive gradients. Identity edits reuse the original score vector, giving zero residual and one original-view cross-entropy from the two half-weighted terms.

Let $\rho_{\mathcal{T}}$ equally mix the binding-edit and representation-edit laws, including identity replacements. The training objective is

$$\mathcal{L}_{\mathrm{SIGMA}}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}_{e \sim \rho_{\mathcal{T}}} [\ell_\theta(i, e)]. \tag{6}$$

The alternating schedule assigns one step to each edit type per two-step block. Setting $\lambda = 0$ gives supervised fine-tuning on the same paired views. Each pair reuses one semantic annotation: the extra view adds computation but no independently labeled input.

4.3 Adaptation and Guidance Updates

We optimize only low-rank adaptation (LoRA) parameters (Hu et al., 2022), freezing the pretrained weights. Training instances are sampled uniformly, and minibatches average Eq. (5), retaining the full supervised weight of identity pairs. We select λ and the checkpoint on the development split, keeping each instance's views in its original split. Transfer evaluation uses composed guidance documents outside the single-edit training support and excludes them from model selection. Centering and permutation add $O(m)$ arithmetic per pair. Appendix A.2 gives the full procedure and scoring conventions.

At inference, we freeze the adapted parameters $\widehat{\theta}$. Given x and current guidance g_{new}, the model scores each candidate once and returns the top K_{top}, with $1 \le K_{\text{top}} \le m$. We rank scores in descending order, break ties by the fixed candidate order, and apply no external output permutation. Only the current guidance is needed.

Composed edits. For a fixed input and candidate inventory, consider successive edits $g_j = T_{e_j}(g_{j-1})$, $j = 1, \dots, k$. Let $\mathbf{u}_j = \mathbf{u}_{g_j}$ and define each step's alignment error as $\varepsilon_j = \|\mathbf{u}_j - P_{e_j}\mathbf{u}_{j-1}\|_2$. Since permutation matrices are orthogonal,

$$\|\mathbf{u}_k - P_{e_k} \cdots P_{e_1}\mathbf{u}_0\|_2 \le \sum_{j=1}^{k} \varepsilon_j. \tag{7}$$

Repeated substitution writes the final discrepancy as a sum of permuted local residuals, whose norms are unchanged. The bound controls accumulated error along the specified sequence; its value depends on the local discrepancies encountered there. Appendix B.2 proves this bound and a sufficient condition for preserving the remapped top-ranked answer.

4.4 A Fixed-Feature Variance-Reduction Perspective

To isolate the effect of alignment, we compare SIGMA with paired supervision on the *same* inputs, edits, and labels. We use a fixed-feature squared-loss proxy, following analyses of consistency regularization and equivariant estimation (Yang et al., 2023; Elesedy & Zaidi, 2021). Inputs and the label-independent edit law are fixed. Noise blocks for original labels are independent and mean-zero, with finite, possibly unequal covariances; all edited views reuse the original label.

Let $\widehat{w}_\lambda$ minimize the proxy and $w^\dagger$ minimize its unregularized population loss; $\lambda = 0$ gives paired supervision. The paired Gram matrix $S \succ 0$ defines the prediction-error metric, while the aligned contrast Gram matrix $C \succeq 0$ measures disagreement between views. Thus $Cw^\dagger = 0$ means that the best paired predictor is equivariant (Appendix B.3).

Theorem 4.1 (Risk reduction from equivariance). *If $Cw^\dagger = 0$, then $\|\widehat{w}_\lambda - w^\dagger\|_S^2 \le \|\widehat{w}_0 - w^\dagger\|_S^2$ for every realized noise vector, simultaneously for all $\lambda \ge 0$. For any fixed $\lambda > 0$, expected improvement is strict exactly when $\operatorname{tr}[C \operatorname{Cov}(\widehat{w}_0)] > 0$: the penalty must act on a direction carrying estimation noise. With equivariance mismatch, expected risk improves exactly when the variance removed exceeds the squared bias introduced.*

Appendix B.4 gives the exact risk identity and proof; Appendix B.5 gives sharp mismatch thresholds.

What sibling edits constrain. In the model $f_{W,a}(x, g) = P_g W \phi(x) + a$, sibling swaps penalize within-group differences in the guidance-independent offset a, leaving the semantic signal and group-constant offsets unconstrained. The benefit depends on which noisy directions the edits constrain. Appendix B.6 derives the spectrum and exact risk for this model under SIGMA's edit law, including a nonzero categorical example.

Statistical perspective. This proxy isolates how alignment reduces estimation variance from noisy responses at fixed inputs. Appendix B.7 quantifies the corresponding budget of independent repeated response draws at the same design points; Appendices B.8–B.9 characterize probability and minimax risk in this setting. The few-shot experiments measure accuracy on unseen inputs as the number of distinct training utterances increases, with an essentially deterministic label for each utterance. Section 5.3 measures this training-label efficiency under nonlinear cross-entropy adaptation.

Table 1: **Generalization to unseen guidance compositions.** Qwen2.5-7B-Instruct, 10 shots per intent; five-seed mean accuracy (%). Base uses g_0; Comp. averages six held-out composition conditions. SIGMA-F uses composition-matched instance weights for both columns. Labeled subsets and the full inventory are shared. **Bold**/underline: best/second-best across all rows. SIGMA-F uses exact binding lookup (Appendix A.3); MetaTuner-J uses an auxiliary generator.

	BANKING77		CLINC150		HWU64	
Method	**Base**	**Comp.**	**Base**	**Comp.**	**Base**	**Comp.**
Frozen prompting	81.73	72.46	91.18	84.37	80.94	71.62
LoRA-SFT (Hu et al., 2022)	88.41	80.96	95.27	90.61	88.73	81.54
Paired SFT ($\lambda = 0$)	89.18	85.37	95.84	93.42	89.71	86.11
Transported KL (Liang et al., 2021)	89.62	86.83	96.03	94.17	90.06	86.94
ICL-Full (Milios et al., 2023)	86.47	78.51	94.71	89.64	86.82	78.29
DLR-Full (Park et al., 2025)	88.09	82.36	95.63	92.18	88.97	82.84
MIPROv2 (Opsahl-Ong et al., 2024)	87.54	80.11	95.19	90.38	87.63	80.27
GEPA (Agrawal et al., 2026)	88.82	84.16	95.91	93.09	89.53	84.77
MetaTuner-J (Bo et al., 2026)	90.17	86.02	96.31	94.26	90.92	87.15
Our method and variants						
SIGMA	**90.71**	88.56	**96.68**	95.37	**91.34**	89.08
SIGMA-F	89.91	**89.91**	96.25	**96.25**	90.49	**90.49**

5 Experiments

5.1 Experimental Setup

Benchmarks and metrics. Primary comparisons use BANKING77 (Casanueva et al., 2020), CLINC150 (Larson et al., 2019), and HWU64 (Liu et al., 2021). We report Base/Comp. accuracy and label-efficiency curves; Table 2 gives guidance-use diagnostics. MASSIVE-EN (FitzGerald et al., 2023), HINT3 (Arora et al., 2020), and complementary metrics are reported in Tables 6–7.

Comparison protocol. The primary 10-shot comparison shares labeled subsets, guidance, and complete candidate inventories across the methods in Table 1. Paired SFT and transported KL isolate paired supervision and the consistency objective; SIGMA-F provides an exact-mapping reference. Appendix C.4 specifies the adaptations. Appendix D.6 reports additional controls and paired statistical comparisons.

Backbones and implementation. Qwen2.5-7B-Instruct (Yang et al., 2024) is the primary backbone. LoRA-SFT, paired SFT, and SIGMA additionally use Llama-3-8B-Instruct (Grattafiori et al., 2024) and Qwen3-8B (Yang et al., 2025) at 5/10 shots. LoRA rank is 16; experiments use five paired seeds. Development accuracy selects the regularization coefficient and checkpoint. Appendix C details the budgets and protocol.

5.2 Main Results

Composition generalization. Across the three datasets, SIGMA reduces the Base-to-Comp. accuracy drop from 6.43 percentage points with LoRA-SFT to 1.91 (Table 1). Its average gain is 2.11 points on Base and 6.63 on Comp., with the larger gain on unseen compositions. SIGMA-F has exactly zero Base-to-Composition drop under matched evaluation support. This follows from its factorized construction: the semantic prediction is held fixed across binding changes and the current output is obtained through exact binding lookup. Accordingly, Base and composition correctness coincide instance-wise for every valid guidance composition. Table 2 covers guidance-use diagnostics and ablations.

Backbone comparison. Across all three backbones and both shot budgets, SIGMA outperforms LoRA-SFT and paired SFT on every dataset under both Base and Comp. (Figure 4). Its composition-accuracy gains over paired SFT range from 1.85 to 3.87 percentage points.

Table 2: **Guidance-use diagnostics and ablations.** Qwen2.5-7B-Instruct, 10 shots; equal-weight three-dataset means. Chg. Acc./PairAcc: target-changing/all eligible pairs. Binding uses the original test population; compositions use accepted-pair weights and provide consistency metrics. Accuracy is in percent; $\mathrm{TV}_{\mathrm{eq}} \in [0, 1]$ (lower is better). Gray subscripts: variant-minus-SIGMA (accuracy: percentage points; TV: absolute units). Best/second-best values across all rows are bold/underlined; SIGMA-F uses exact binding lookup.

	Binding swaps		Compositions		Consistency	
Variant	**Chg. Acc.**	**PairAcc**	**Chg. Acc.**	**PairAcc**	**EquivAcc**	$\mathrm{TV}_{\mathrm{eq}}\downarrow$
Training controls						
SFT (token-matched)	$61.37_{-28.30}$	$83.16_{-8.87}$	$52.44_{-33.17}$	$74.21_{-14.93}$	$72.84_{-22.74}$	$0.1487_{+0.1298}$
Paired SFT ($\lambda = 0$)	$81.26_{-8.41}$	$88.91_{-3.12}$	$73.18_{-12.43}$	$83.72_{-5.42}$	$87.61_{-7.97}$	$0.0719_{+0.0530}$
Transported KL	$84.11_{-5.56}$	$89.83_{-2.20}$	$77.41_{-8.20}$	$85.73_{-3.41}$	$91.28_{-4.30}$	$0.0424_{+0.0235}$
Our method and variants						
SIGMA: binding only	$88.76_{-0.91}$	$91.22_{-0.81}$	$82.18_{-3.43}$	$86.73_{-2.41}$	$92.88_{-2.70}$	$0.0321_{+0.0132}$
SIGMA: reversal only	$62.41_{-27.26}$	$84.69_{-7.34}$	$55.87_{-29.74}$	$78.83_{-10.31}$	$84.31_{-11.27}$	$0.0875_{+0.0686}$
SIGMA	89.67	92.03	85.61	89.14	95.58	0.0189
SIGMA-F	**92.18**	**92.21**	**92.09**	**92.22**	**100.00**	**0.0000**

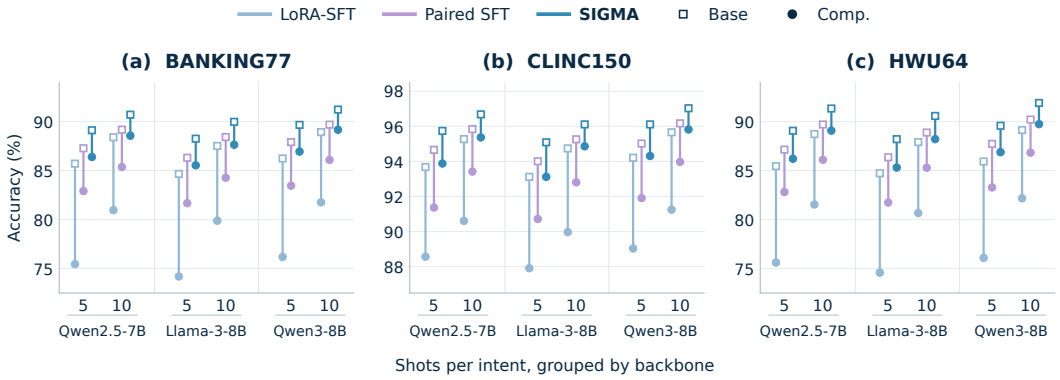


Figure 4: **Backbone comparison at 5 and 10 shots per intent.** Five-seed mean accuracy across three datasets (horizontal panels). Colors identify methods; segments connect Base (open squares) and Comp. (filled circles) accuracies for each backbone and shot budget. Base uses the original guidance; Comp. averages six held-out composition conditions. Within each backbone, methods share labeled subsets, initialization, and candidate scoring. Qwen2.5-7B and Llama-3-8B use Instruct checkpoints; Qwen3-8B uses non-thinking mode. Panels use dataset-specific vertical ranges.

5.3 Label Efficiency

Figure 5 shows the empirical trade-off between training annotations and generalization to unseen guidance compositions. SIGMA forms the Pareto frontier among the four methods and tested budgets. At four examples per intent, it exceeds the mean Comp. accuracy of 16-shot LoRA-SFT on all three datasets, using **75% fewer training labels** with the same development set.

5.4 Margin–Residual Diagnostics

Figure 6 analyzes 4,926 initially correct trajectories per method from 12-candidate subsets of BANKING77, CLINC150, and HWU64 with $k \in \{2, 4, 8\}$. Accuracies are conditional and distinct from full-inventory results. From saved scores and swaps, we reconstruct $E_{\mathrm{local}} = \sum_j \|\mathbf{u}_j - P_j \mathbf{u}_{j-1}\|_2$, $\mathcal{R}_{\mathrm{mr}} = \sqrt{2} E_{\mathrm{local}} / \gamma_0$, and final correctness. Corollary B.4 guarantees a unique correct top-1 for $\mathcal{R}_{\mathrm{mr}} < 1$ (above the dashed line).

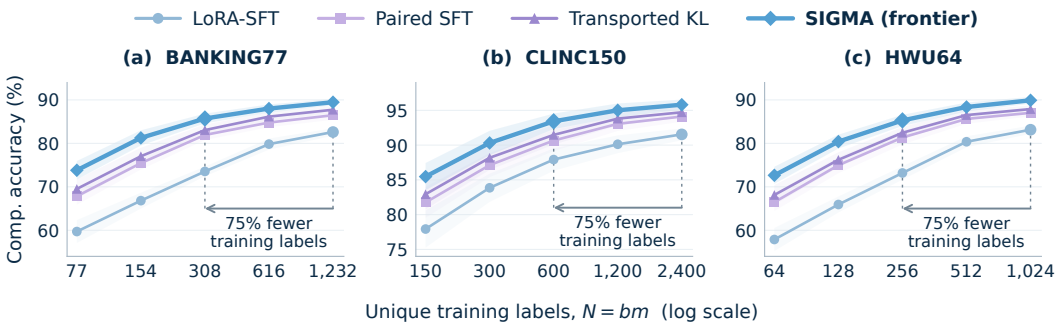


Figure 5: **Annotation–generalization Pareto frontier.** Qwen2.5-7B-Instruct; five-seed means over six held-out composition conditions, with ± one standard deviation shaded. Points use $N = bm$ unique labels (m intents; $b \in \{1, 2, 4, 8, 16\}$) from shared, nested subsets. Lines connect observations; the thick line marks the empirical Pareto frontier across the four methods and tested budgets. Arrows show the smallest tested SIGMA budget exceeding 16-shot LoRA-SFT's mean accuracy (4-shot throughout), excluding fixed development annotations from savings. Table 5 provides full statistics and AULC.

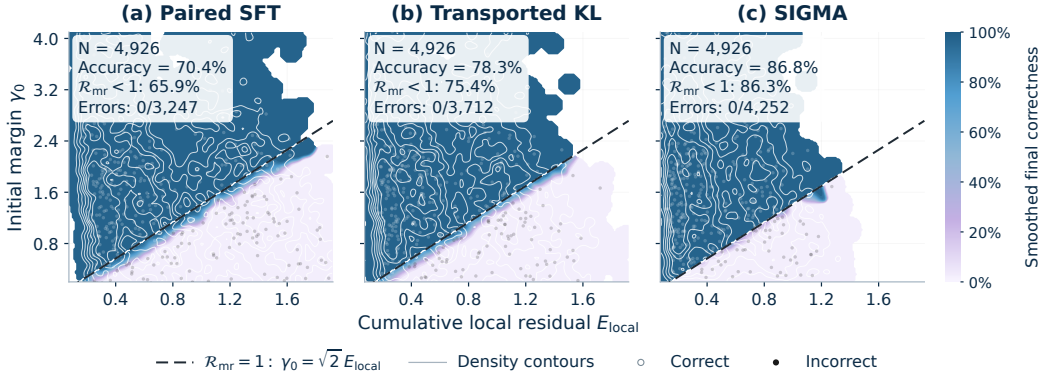


Figure 6: **Margin–residual score diagnostic.** Each panel pools 4,926 initially correct trajectories from three 12-candidate diagnostic subsets and $k \in \{2, 4, 8\}$, so accuracies are conditional on this restricted evaluation. Color shows final correctness smoothed separately across $\mathcal{R}_{\mathrm{mr}} = 1$; white contours show density and dots a deterministic subset without jitter. Above the dashed line ($\mathcal{R}_{\mathrm{mr}} < 1$), Corollary B.4 guarantees a unique correct top-1; insets summarize all records, with "Errors" counting failures in this certified region.

For Paired SFT, transported KL, and SIGMA, respectively, certified coverage is **65.9%, 75.4%, and 86.3%**, with **0/3,247, 0/3,712, and 0/4,252** errors; pooled final accuracy is **70.4%, 78.3%, and 86.8%**. Certified trajectories satisfy the bound and preserve correctness throughout. Within $\gamma_0 \in [1.4, 1.8]$, SIGMA's median cumulative residual is **42.8%** lower than Paired SFT's and **24.7%** lower than transported KL's (descriptive comparison). Appendix D.7 details reconstruction and scope.

6 Conclusion

We presented SIGMA to address the gap between task accuracy under fixed guidance and reliable behavior under guidance updates. The method makes each annotation supervise both the task decision and how candidate predictions should transform across guidance versions. Across the evaluated datasets and backbones, gains on unseen edit combinations indicate that local training interventions can support transferable guidance use, with improved training-label efficiency on the three primary intent benchmarks. A natural next question is how to select guidance edits that best promote transfer to broader families of updates under fixed annotation and computation budgets.

AI USE STATEMENT

We used generative AI tools to assist with drafting and editing the manuscript, organizing the literature, preparing figures and tables, and editing LaTeX source files. The authors take full responsibility for the final manuscript, including the accuracy of its content, references, and any AI-assisted text or artifacts.

REFERENCES

Ilias Aarab. BTZSC: A benchmark for zero-shot text classification across cross-encoders, embedding models, and rerankers. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/417e1c15b3d49852fceded8aa104107d-Abstract-Conference.html.

Lakshya A. Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J. Ryan, Meng Jiang, Christopher Potts, Koushik Sen, Alex Dimakis, Ion Stoica, Dan Klein, Matei Zaharia, and Omar Khattab. GEPA: Reflective prompt evolution can outperform reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/0e9e708b6f48e14fd0ac29e167413f76-Abstract-Conference.html.

Gaurav Arora, Chirag Jain, Manas Chaturvedi, and Krupal Modi. HINT3: Raising the bar for intent detection in the wild. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pp. 100–105, 2020. doi: 10.18653/v1/2020.insights-1.16. URL https://aclanthology.org/2020.insights-1.16/.

Baolong Bi, Shaohan Huang, Yiwei Wang, Tianchi Yang, Zihan Zhang, Haizhen Huang, Lingrui Mei, Junfeng Fang, Zehao Li, Furu Wei, Weiwei Deng, Feng Sun, Qi Zhang, and Shenghua Liu. Context-DPO: Aligning language models for context-faithfulness. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10280–10300, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.536. URL https://aclanthology.org/2025.findings-acl.536/.

Xiaohe Bo, Rui Li, Zexu Sun, Quanyu Dai, Zeyu Zhang, Zihang Tian, Xu Chen, and Zhenhua Dong. Prompt and parameter co-optimization for large language models. In *Proceedings of the International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=WPJnLvQrP3. ICLR 2026 poster.

Iñigo Casanueva, Tadas Temčinas, Daniela Gerz, Matthew Henderson, and Ivan Vulić. Efficient intent detection with dual sentence encoders. In *Proceedings of the 2nd Workshop on Natural Language Processing for Conversational AI*, pp. 38–45, 2020. doi: 10.18653/v1/2020.nlp4convai-1.5.

Peng Chen and Bang Wang. On collaborating small and large models for few-shot intent detection. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 13937–13953, 2025. doi: 10.18653/v1/2025.findings-emnlp.749. URL https://aclanthology.org/2025.findings-emnlp.749/.

Bryn Elesedy and Sheheryar Zaidi. Provably strict generalisation benefit for equivariant models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 2959–2969, 2021. URL https://proceedings.mlr.press/v139/elesedy21a.html.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, et al. MASSIVE: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4277–4302, 2023. doi: 10.18653/v1/2023.acl-long.235. URL https://aclanthology.org/2023.acl-long.235/.

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.

Blessed Guda, Lawrence Francis, Gabrial Zencha Ashungafac, Carlee Joe-Wong, and Moise Busogi. Quantifying and mitigating selection bias in LLMs: A transferable LoRA fine-tuning and efficient majority voting approach. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2019–2038, Mumbai, India, December 2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-303-6. doi: 10.18653/v1/2025.findings-ijcnlp.127. URL https://aclanthology.org/2025.findings-ijcnlp.127/.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. An evaluation dataset for intent classification and out-of-scope prediction. In *Proceedings of EMNLP-IJCNLP*, pp. 1311–1316, 2019. doi: 10.18653/v1/D19-1131.

Daliang Li, Ankit Singh Rawat, Manzil Zaheer, Xin Wang, Michal Lukasik, Andreas Veit, Felix Yu, and Sanjiv Kumar. Large language models with controllable working memory. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 1774–1793, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.112. URL https://aclanthology.org/2023.findings-acl.112/.

Xiaobo Liang, Lijun Wu, Juntao Li, Yue Wang, Qi Meng, Tao Qin, Wei Chen, Min Zhang, and Tie-Yan Liu. R-Drop: Regularized dropout for neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 10890–10905, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5a66b9200f29ac3fa0ae244cc2a51b39-Abstract.html.

Yen-Ting Lin, Alexandros Papangelis, Seokhwan Kim, Sungjin Lee, Devamanyu Hazarika, Mahdi Namazifar, Di Jin, Yang Liu, and Dilek Hakkani-Tur. Selective in-context data augmentation for intent detection using pointwise V-information. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 1463–1476, 2023. doi: 10.18653/v1/2023.eacl-main.107. URL https://aclanthology.org/2023.eacl-main.107/.

Kai Liu, Ze Chen, Zhihang Fu, Wei Zhang, Rongxin Jiang, Fan Zhou, Yaowu Chen, Yue Wu, and Jieping Ye. Structure-aware domain knowledge injection for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29443–29464, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.1425. URL https://aclanthology.org/2025.acl-long.1425/.

Xingkun Liu, Arash Eshghi, Pawel Swietojanski, and Verena Rieser. Benchmarking natural language understanding services for building conversational agents. In *Increasing Naturalness and Flexibility in Spoken Dialogue Interaction*, pp. 165–183. Springer, 2021. doi: 10.1007/978-981-15-9323-9_15.

Aristides Milios, Siva Reddy, and Dzmitry Bahdanau. In-context learning for text classification with many labels. In *Proceedings of the 1st GenBench Workshop on (Benchmarking) Generalisation in NLP*, pp. 173–184, 2023. doi: 10.18653/v1/2023.genbench-1.14. URL https://aclanthology.org/2023.genbench-1.14/.

Krista Opsahl-Ong, Michael J. Ryan, Josh Purtell, David Broman, Christopher Potts, Matei Zaharia, and Omar Khattab. Optimizing instructions and demonstrations for multi-stage language model programs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*

Processing, pp. 9340–9366, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.525. URL https://aclanthology.org/2024.emnlp-main.525/.

Gyutae Park, Ingeol Baek, Byeongjeong Kim, Joongbo Shin, and Hwanhee Lee. Dynamic label name refinement for few-shot dialogue intent classification. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 41–52, 2025. doi: 10.18653/v1/2025.acl-short.3. URL https://aclanthology.org/2025.acl-short.3/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of EMNLP-IJCNLP*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc Le. Symbol tuning improves in-context learning in language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 968–979, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.61. URL https://aclanthology.org/2023.emnlp-main.61/.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025. URL https://arxiv.org/abs/2505.09388.

Shuo Yang, Yijun Dong, Rachel Ward, Inderjit S. Dhillon, Sujay Sanghavi, and Qi Lei. Sample efficiency of data augmentation consistency regularization. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 3825–3853, 2023. URL https://proceedings.mlr.press/v206/yang23c.html.

Asaf Yehudai, Matan Vetzler, Yosi Mass, Koren Lazar, Doron Cohen, and Boaz Carmeli. QAID: Question answering inspired few-shot intent detection. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=3ZPESALKXO.

Jianguo Zhang, Trung Bui, Seunghyun Yoon, Xiang Chen, Zhiwei Liu, Congying Xia, Quan Hung Tran, Walter Chang, and Philip Yu. Few-shot intent detection via contrastive pre-training and fine-tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 1906–1912, 2021. doi: 10.18653/v1/2021.emnlp-main.144. URL https://aclanthology.org/2021.emnlp-main.144/.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv preprint arXiv:2506.05176, 2025. URL https://arxiv.org/abs/2506.05176.

Jinquan Zheng, Jia Yuan, Jiacheng Yao, Chenyang Gu, Pujun Zheng, and Guoxiu He. Mitigating selection bias in large language models via permutation-aware GRPO. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 35125–35143, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.1621. URL https://aclanthology.org/2026.acl-long.1621/.

APPENDIX CONTENTS

A Technical Details **14**

A.1 Guidance Construction and Intervention Sampling 14
A.2 Training and Candidate Scoring 15
A.3 SIGMA-F: Factorized Semantic Classification 15

B Theoretical Analysis and Proofs **17**

B.1 Geometry of the Equivariance Penalty 17
B.2 Ordered Compositions and Ranking Stability 17
B.3 Edit-Averaged Linear Proxy 18
B.4 Shared Noise, Exact Risk, and Strict Improvement 19
B.5 Sharp Mismatch Bound 21
B.6 Hierarchy-Dependent Model with Nonzero Semantic Signal 22
B.7 Effective Dimension and Annotation Budgets 26
B.8 Gaussian Probability Guarantee for the Risk Gain 27
B.9 Gaussian Minimax Floor on the Equivariant Subspace 29

C Experimental Details **32**

C.1 Benchmarks, Splits, and Annotation Budgets 32
C.2 Guidance and Held-Out Interventions 32
C.3 Metrics and Statistical Reporting 33
C.4 Baseline Adaptations and Fair Comparisons 34
C.5 Backbones, Optimization, and Resource Accounting 34
C.6 Result Aggregation 35
C.7 Preliminary Study Protocol and Scoring 35

D Additional Experimental Results **38**

D.1 Average Accuracy and Composition Gap 38
D.2 Label Budgets 38
D.3 Additional Benchmarks and Metrics 40
D.4 Computational Cost 41
D.5 Preliminary Study Results 42
D.6 Controls for Guidance Use and Label Efficiency 43
D.7 Margin–Residual Diagnostic 47
D.8 Which Spectral Directions Does SIGMA Contract? 48
D.9 Where Does Guidance-Aware Behavior Reside? 50

A TECHNICAL DETAILS

A.1 GUIDANCE CONSTRUCTION AND INTERVENTION SAMPLING

Task record and rendering. A task record stores a rooted tree $\mathcal{T}$, one immutable identifier and semantic description per node, and a bijection $a : \mathcal{Z} \to \mathcal{C}$. Leaves are the semantic labels. The candidate inventory contains distinct, nonempty output strings with distinct token sequences. Node identifiers and candidate strings determine fixed lexicographic orders. Guidance is serialized as a nested list: each internal-node record contains its identifier, description, and child records; each leaf record contains its identifier, semantic description, and designated output string. Records are unambiguously delimited. The base guidance g_0 uses ascending child-identifier order. Within a task, every training instance uses this base guidance, so $g_i = g_0$ in Eq. (5).

The prompt concatenates the task instruction, the complete guidance, the input x, and the full candidate list in fixed order. The instruction states that the model must select the output bound to the input's semantic class and that sibling order does not denote priority. The instance label is never included. All splits use the same prompt fields, tokenizer, and candidate inventory. The guidance is constructed from the task schema; evaluation inputs and instance-level test annotations do not determine its wording.

Binding-edit law. For an internal node v, let $L(v)$ be its set of leaf children and $d_v = |L(v)|$. Define $\mathcal{V} = \{v : v \neq \mathrm{root}(\mathcal{T}),\ d_v \geq 2\}$. When $\mathcal{V} \neq \varnothing$, the probability of selecting parent v and pair $\{z_a, z_b\} \subseteq L(v)$ is

$$\rho_{\mathrm{bind}}(v, \{z_a, z_b\}) = \frac{1}{|\mathcal{V}|\binom{d_v}{2}}. \tag{A.1}$$

The selected output strings $a_g(z_a), a_g(z_b)$ are exchanged in the binding map, and the complete document is regenerated. Descriptions remain attached to semantic nodes. The renderer obtains all output references from the updated binding map. The operation preserves the semantic leaves, descriptions, candidates, and candidate order. For $\mathcal{V} = \varnothing$, ρ_{bind} is a point mass on the identity edit. A flat tree therefore uses the identity edit on binding steps.

Representation-edit law. The representation transformation reverses the child list at every internal node, retaining all record contents and parent-child edges. Applying it twice restores the original serialization. Its output map is $\mathrm{id}_{\mathcal{C}}$, and ρ_{repr} is a point mass on this transformation. The objective uses $\rho_{\mathcal{T}} = \frac{1}{2}\rho_{\mathrm{bind}} + \frac{1}{2}\rho_{\mathrm{repr}}$, with odd and even optimization steps assigned to the two components, respectively. For binding steps, the node and pair are sampled independently for each minibatch instance.

Validation and identity replacement. Before training, each task record is checked for a valid rooted tree, complete semantic descriptions, a bijective binding map, and unique candidates. Malformed task records are rejected. Each constructed guidance document is parsed back into its tree, descriptions, and binding map. Validation requires the same semantic tree and descriptions, the same candidate inventory, and $a_{g'}(z) = \pi_e(a_g(z))$ for every $z \in \mathcal{Z}$. A binding edit must fix all unselected bindings; a representation edit must fix every binding. A failed construction is replaced by $(g', \pi_e) = (g, \mathrm{id}_{\mathcal{C}})$ and retained in the minibatch. Its single forward score vector is reused in both loss positions, yielding exactly $-\log q_\theta(y \mid x, g)$ with zero equivariance penalty. There is no rejection resampling and no change to the minibatch denominator.

Intervention coverage. Edits are independent of the gold label, so many valid pairs preserve their target. For a label z whose parent v_z belongs to $\mathcal{V}$, a sampled binding swap changes its output with probability

$$\Pr[\pi_e(a_g(z)) \neq a_g(z) \mid e \sim \rho_{\mathrm{bind}}] = \frac{2}{|\mathcal{V}| d_{v_z}}. \tag{A.2}$$

For all other leaves, this probability is zero. The formula follows because $d_{v_z} - 1$ of the $\binom{d_{v_z}}{2}$ pairs contain z. Under the equal mixture, the probability is halved. These expressions describe valid symbolic constructions. Construction failures can only reduce the realized target-change rate. The empirical rate is the number of pairs with $y' \neq y$ divided by the number of attempted pairs; it is

also recorded separately for binding steps. Identity-replacement counts are recorded with the same denominator. These quantities distinguish target-changing supervision from invariance supervision.

A.2 Training and Candidate Scoring

Scoring conventions. Each candidate is tokenized at the answer boundary and terminated by the model's fixed end-of-answer token. Equation (3) averages log probabilities over the resulting nonempty sequence, including that token. Only answer tokens contribute to the score. All candidates are evaluated with teacher forcing under the same prompt; the candidate softmax is computed with a log-sum-exp normalization. No temperature parameter is learned. Dropout is disabled in the paired scorer, so discrepancies measure changes induced by guidance. Gradients pass through both branches and through score centering. An identity pair reuses the same computed scores and their computation graph.

Optimization procedure. The procedure takes the task record, labeled training split, pretrained language model, a LoRA configuration, an optimizer configuration, a training-step budget, and the loss coefficient λ as inputs.

1. Validate the task record and render g_0. Split original instances before constructing guidance views. Freeze the task schema, candidate order, and prompt fields.
2. Initialize the LoRA parameters and keep the pretrained parameters fixed. Number optimizer steps from $b = 1$.
3. At step b, sample a minibatch of original instances uniformly with replacement. For odd b, independently draw a binding edit per instance using Eq. (A.1). For even b, apply the representation edit to each instance's base guidance.
4. Validate each edit and apply the identity replacement defined in Appendix A.1. Set $y_i = a_{g_0}(z_i)$ and $y_i' = \pi_e(y_i)$.
5. Score the complete candidate inventory under both guidance views. Reuse the original score vector for an identity pair. Center the scores and align them by candidate-index permutation.
6. Average Eq. (5) over all minibatch instances and take one optimizer step. Run an even number of optimizer steps, completing equally many binding and representation steps.
7. Evaluate checkpoints after even-numbered steps. Select the checkpoint and λ by top-1 accuracy on a fixed development pool containing one base-guidance view and one edit drawn from $\rho_{\mathcal{T}}$ per development instance. Keep this pool identical across compared settings. Break ties by the smaller λ, then the earlier checkpoint.
8. Freeze the selected model. For each test input, score the complete candidate inventory under the test guidance and sort by descending score, breaking ties by the fixed candidate order.

The task's LoRA rank, target layers, optimizer, and training budget are shared with paired supervised fine-tuning. SIGMA adds only λ to that configuration. Let $\mathcal{G}_{\text{train}}$ contain g_0, its reversed serialization, and every single binding-swap document reachable from g_0. Training and development views lie in this set. Unseen-guidance evaluation uses composed edits whose final document is outside $\mathcal{G}_{\text{train}}$; compositions returning a member of this set are excluded from that evaluation. The held-out composed documents do not participate in checkpoint selection.

Cost accounting. A nonidentity pair requires candidate scoring under two guidance views. Mean subtraction, candidate-index permutation, and the residual norm require $O(m)$ arithmetic and $O(m)$ score storage; dense $m \times m$ matrices are unnecessary. Candidate-score computation still depends on prompt length and candidate token lengths. Label budgets count original instances, while processed guidance views and training tokens are counted separately. Inference uses one current guidance view and the same candidate inventory as training.

A.3 SIGMA-F: Factorized Semantic Classification

SIGMA-F separates semantic recognition from output binding under the same bijective task specification. Let $\widehat{r}_\phi(z \mid x)$ be a semantic classifier trained on the same original labeled instances, using the

same semantic descriptions and a canonical guidance representation. Its candidate probabilities under g are obtained by exact map lookup:

$$q_\phi^{\mathrm{fact}}(c \mid x, g) = \widehat{r}_\phi(a_g^{-1}(c) \mid x). \tag{A.3}$$

The classifier's semantic distribution is held fixed when bindings change. Since $a_{g'}^{-1} = a_g^{-1} \circ \pi_e^{-1}$, this predictor satisfies $\mathbf{q}_{g'}^{\mathrm{fact}} = P_e \mathbf{q}_g^{\mathrm{fact}}$ exactly, irrespective of its semantic accuracy. Its top-1 decision is $a_g(\arg\max_z \widehat{r}_\phi(z \mid x))$, with semantic ties resolved in a fixed semantic-label order.

SIGMA-F requires explicit access to the binding map and a semantic prediction interface. SIGMA directly scores outputs from guidance text and uses centered-score alignment to encourage equivariance. Prediction accuracy and label efficiency are evaluated separately from this structural property.

B Theoretical Analysis and Proofs

B.1 Geometry of the Equivariance Penalty

Fix an input, model parameters, and candidate inventory. Let $P_\perp = I_m - \mathbf{1}\mathbf{1}^\top/m$, so $\mathbf{u}_g = P_\perp \mathbf{s}_g$. The identities $P_\perp^\top = P_\perp$, $P_\perp^2 = P_\perp$, and $P_\perp \mathbf{1} = 0$ make $P_\perp$ the orthogonal projector onto the zero-sum subspace. A permutation matrix satisfies $P_e^\top P_e = I_m$, $P_e \mathbf{1} = \mathbf{1}$, and $P_\perp P_e = P_e P_\perp$.

Proposition B.1 (Centered-score characterization)**.** *For finite score vectors $\mathbf{s}_g, \mathbf{s}_{g'}$, the candidate distributions satisfy $\mathbf{q}_{g'} = P_e \mathbf{q}_g$ if and only if $P_\perp \mathbf{s}_{g'} = P_e P_\perp \mathbf{s}_g$.*

Proof. For any finite score vector $\mathbf{s}$, $\log \mathrm{softmax}(\mathbf{s}) = \mathbf{s} - \log(\sum_c \exp s_c)\mathbf{1}$. Taking logarithms coordinatewise in $\mathbf{q}_{g'} = P_e \mathbf{q}_g$ therefore gives $\mathbf{s}_{g'} - P_e \mathbf{s}_g = b\mathbf{1}$ for a scalar b. Applying $P_\perp$ proves the forward direction. Conversely, $P_\perp(\mathbf{s}_{g'} - P_e \mathbf{s}_g) = 0$ implies the difference lies in $\mathrm{span}\{\mathbf{1}\}$. Softmax removes this constant difference and commutes with candidate permutations, proving the reverse direction. □

Relative evidence. Let $\boldsymbol{\delta} = \mathbf{u}_{g'} - P_e \mathbf{u}_g$. Its entries sum to zero. For distinct candidates c, d, define the aligned log-odds discrepancy

$$\Delta_{cd} = \log \frac{q_\theta(c \mid x, g')}{q_\theta(d \mid x, g')} - \log \frac{q_\theta(\pi_e^{-1}(c) \mid x, g)}{q_\theta(\pi_e^{-1}(d) \mid x, g)} = \delta_c - \delta_d. \tag{B.1}$$

Here $c < d$ below refers to the fixed candidate order. Expanding the pairwise differences and using $\sum_c \delta_c = 0$ yields

$$\sum_{c<d} \Delta_{cd}^2 = m \sum_c \delta_c^2 - \left(\sum_c \delta_c \right)^2 = m \|\boldsymbol{\delta}\|_2^2. \tag{B.2}$$

Consequently, the regularizer in Eq. (5) equals $\lambda (2m^2)^{-1} \sum_{c<d} \Delta_{cd}^2$. This identity explains its sensitivity to relative evidence among all candidates, including alternatives with low probability. Preserving these odds under binding transport adds a constraint beyond paired correctness.

Proposition B.2 (Distance to the aligned subspace)**.** *Let $\mathbf{r} = [\mathbf{u}_g^\top, \mathbf{u}_{g'}^\top]^\top \in \mathbb{R}^{2m}$ and $\mathcal{E}_e = \{(\mathbf{v}, P_e \mathbf{v}) : \mathbf{v} \in \mathbb{R}^m\}$. The orthogonal projector onto $\mathcal{E}_e$ and its squared residual are*

$$\Pi_e = \frac{1}{2} \begin{bmatrix} I_m & P_e^\top \\ P_e & I_m \end{bmatrix}, \qquad \|(I_{2m} - \Pi_e)\mathbf{r}\|_2^2 = \frac{1}{2} \|\mathbf{u}_{g'} - P_e \mathbf{u}_g\|_2^2. \tag{B.3}$$

Proof. The closest aligned pair minimizes $\|\mathbf{u}_g - \mathbf{v}\|_2^2 + \|\mathbf{u}_{g'} - P_e \mathbf{v}\|_2^2$. Differentiation gives $\overline{\mathbf{u}} = (\mathbf{u}_g + P_e^\top \mathbf{u}_{g'})/2$. Thus $\Pi_e \mathbf{r} = (\overline{\mathbf{u}}, P_e \overline{\mathbf{u}})$. Direct multiplication verifies $\Pi_e^\top = \Pi_e$ and $\Pi_e^2 = \Pi_e$. The residual blocks are $-P_e^\top \boldsymbol{\delta}/2$ and $\boldsymbol{\delta}/2$. Orthogonality of P_e makes their squared norms sum to $\|\boldsymbol{\delta}\|_2^2/2$. □

For centered inputs, $\overline{\mathbf{u}}$ is also centered, so $(\overline{\mathbf{u}}, P_e \overline{\mathbf{u}})$ belongs to the centered part of $\mathcal{E}_e$. The loss therefore measures distance to compatible paired predictions after removing the softmax offset freedom. Training penalizes the residual while retaining both original branch predictions; $\Pi_e \mathbf{r}$ gives their geometric projection.

B.2 Ordered Compositions and Ranking Stability

Local binding swaps generate permutations within the leaf-sibling groups specified by $\mathcal{V}$. Groups have disjoint leaf sets; their generated permutation group is the product of the symmetric groups on those sets. Leaves without an eligible parent remain fixed. The representation transformation preserves all output bindings. The following result applies to ordered compositions in this intervention family, without requiring individual edits to commute.

Proposition B.3 (Accumulation of local discrepancies)**.** *For $g_j = T_{e_j}(g_{j-1})$ on a fixed input and candidate inventory, let $P_j = P_{e_j}$, $\mathbf{d}_j = \mathbf{u}_j - P_j \mathbf{u}_{j-1}$, and $\varepsilon_j = \|\mathbf{d}_j\|_2$. Then Eq. (7) holds.*

Proof. Repeatedly substituting $\mathbf{u}_j = P_j\mathbf{u}_{j-1} + \mathbf{d}_j$ gives

$$\mathbf{u}_k - P_k \cdots P_1\mathbf{u}_0 = \sum_{j=1}^{k} P_k \cdots P_{j+1}\mathbf{d}_j, \tag{B.4}$$

where the product for $j = k$ is the identity. Every product is orthogonal, so $\|P_k \cdots P_{j+1}\mathbf{d}_j\|_2 = \varepsilon_j$. The triangle inequality proves the bound. □

Corollary B.4 (Transported top-1 margin)**.** *Suppose the correct initial output y_0 has centered-score margin $\gamma_0 = \mathbf{u}_0[y_0] - \max_{c\neq y_0} \mathbf{u}_0[c] > 0$. Set $y_k = (\pi_{e_k} \circ \cdots \circ \pi_{e_1})(y_0)$. If $\gamma_0 > \sqrt{2}\sum_{j=1}^{k} \varepsilon_j$, then y_k is the unique top-ranked candidate under g_k.*

Proof. Let $\mathbf{v} = P_k \cdots P_1\mathbf{u}_0$ and $\mathbf{h} = \mathbf{u}_k - \mathbf{v}$. The permutation preserves the initial margin around the transported output y_k. For every $c \neq y_k$, Cauchy–Schwarz gives

$$\mathbf{u}_k[y_k] - \mathbf{u}_k[c] \geq \gamma_0 - |\mathbf{h}[y_k] - \mathbf{h}[c]| \geq \gamma_0 - \sqrt{2}\|\mathbf{h}\|_2 \geq \gamma_0 - \sqrt{2}\sum_{j=1}^{k} \varepsilon_j > 0. \tag{B.5}$$

The factor $\sqrt{2}$ is the Euclidean norm of the difference of two coordinate basis vectors. Centering preserves candidate rankings, so the conclusion also holds for the original scores. □

Both statements are deterministic bounds along a specified sequence. Statistical generalization to unseen guidance requires control of the local discrepancies on those views.

B.3 An edit-averaged linear proxy with one response per original input

The analysis conditions on original inputs, features, and a label-independent edit law. Prediction-risk identities require only second moments; the concentration and minimax results additionally assume Gaussian noise. Fixed-penalty expectations use a deterministic λ. A validation-selected λ can depend on training noise through the fitted models and requires separate analysis. Under exact equivariance, samplewise contraction holds simultaneously for every nonnegative coefficient.

Let $Q \in \mathbb{R}^{m\times h}$, $h = m - 1$, have orthonormal columns spanning $\mathbf{1}^{\perp}$. A candidate permutation induces $R_e = Q^{\top}P_eQ$. Because P_e preserves $\mathbf{1}^{\perp}$, $R_e^{\top}R_e = I_h$. For each original example i, fix $A_i, A'_{ie} \in \mathbb{R}^{h\times d}$ and a finite edit law ρ_i. The method's law $\rho_{\mathcal{T}}$ is a special case. A point mass represents one fixed edit. Set $B_{ie} = R_e^{\top}A'_{ie}$ and observe only $t_i = \mu_i + \varepsilon_i$. The blocks ε_i are independent, with mean zero and finite covariance Σ_i. All edited targets are R_et_i, with no new noise draws.

Replace each cross-entropy in Eq. (5) by the candidate-normalized squared loss $\|Aw - t\|_2^2/(2m)$, retaining the method's discrepancy coefficient $\lambda/(2m)$. Multiplying the entire proxy objective by m leaves its minimizer unchanged and gives

$$\begin{aligned}\widehat{\mathcal{L}}_\lambda(w) = \frac{1}{4n}\sum_{i=1}^{n} \mathbb{E}_{e\sim\rho_i}\left[\|A_iw - t_i\|_2^2 + \|B_{ie}w - t_i\|_2^2\right] \\ + \frac{\lambda}{2n}\sum_{i=1}^{n} \mathbb{E}_{e\sim\rho_i}\|(B_{ie} - A_i)w\|_2^2, \qquad \widehat{w}_\lambda = \arg\min_w \widehat{\mathcal{L}}_\lambda(w).\end{aligned} \tag{B.6}$$

The baseline $\lambda = 0$ retains exactly the same paired views and reused labels; only the discrepancy penalty changes. The factor $\tau = 4\lambda$ below follows from this normalization. Alignment is exact: $\|A'_{ie}w - R_et_i\|_2 = \|B_{ie}w - t_i\|_2$, and the corresponding discrepancy norms agree by orthogonality.

Define

$$\begin{aligned}M_{ie} = \frac{A_i + B_{ie}}{2}, \qquad D_{ie} = \frac{B_{ie} - A_i}{2}, \qquad \overline{M}_i = \mathbb{E}_eM_{ie}, \\ C = \frac{1}{n}\sum_i \mathbb{E}_eD_{ie}^{\top}D_{ie}, \qquad S = \frac{1}{n}\sum_i \mathbb{E}_e(M_{ie}^{\top}M_{ie} + D_{ie}^{\top}D_{ie}) \succ 0, \\ K = S^{-1/2}CS^{-1/2}, \qquad F_\tau = (I_d + \tau K)^{-1}.\end{aligned} \tag{B.7}$$

Stack $\overline{M}_i$ into $\overline{M}$, and stack $t_i, \mu_i, \varepsilon_i$ into t, μ, ε. For a positive semidefinite matrix H, $H^\dagger$ denotes its Moore–Penrose inverse and $H^{\dagger/2} = (H^\dagger)^{1/2}$. Positive definiteness identifies the finite-dimensional proxy; the assumption is confined to this model.

Proof roadmap. The parallelogram identity separates paired supervision into a mean-view regression and a contrast penalty. Whitening then expresses the regularized estimator as a contraction of the unregularized estimate. Computing noise covariance after averaging edits preserves the correlations induced by label reuse. Its bias and trace terms yield the exact risk without Gaussianity or independent spectral coordinates; an equivariant oracle lies in the penalty's null space and is therefore unchanged.

Lemma B.5 (Normal equations, paired oracle, and spectral range)**.** *Let* $\mathcal{L}(w) = 2\mathbb{E}_\varepsilon \widehat{\mathcal{L}}_0(w)$. *Then*

$$\begin{aligned} \widehat{w}_\lambda = (S+\tau C)^{-1}\frac{\overline{M}^\top t}{n}, \quad w^\dagger = S^{-1}\frac{\overline{M}^\top \mu}{n}, \\ \mathcal{L}(w) - \mathcal{L}(w^\dagger) = \|w - w^\dagger\|_S^2, \qquad 0 \preceq K \preceq I_d. \end{aligned} \tag{B.8}$$

Proof. For every pair, the mixed terms cancel in the parallelogram identity:

$$\begin{aligned} \|A_i w - t_i\|_2^2 + \|B_{ie} w - t_i\|_2^2 &= \|(M_{ie} w - t_i) - D_{ie} w\|_2^2 + \|(M_{ie} w - t_i) + D_{ie} w\|_2^2 \\ &= 2\|M_{ie} w - t_i\|_2^2 + 2\|D_{ie} w\|_2^2. \end{aligned} \tag{B.9}$$

Consequently,

$$\begin{aligned} \widehat{\mathcal{L}}_\lambda(w) &= \frac{1}{2n}\sum_i \mathbb{E}_e \|M_{ie} w - t_i\|_2^2 + \frac{1+\tau}{2} w^\top C w \\ &= \frac{1}{2} w^\top (S+\tau C) w - \frac{1}{n} w^\top \overline{M}^\top t + \frac{1}{2n}\|t\|_2^2. \end{aligned} \tag{B.10}$$

For nonzero z, $z^\top(S+\tau C)z \geq z^\top S z > 0$. Thus the first-order condition gives the unique minimizer. Taking expectations and completing the square gives

$$\begin{aligned} \mathcal{L}(w) - \mathcal{L}(w^\dagger) &= w^\top S w - 2 w^\top S w^\dagger + (w^\dagger)^\top S w^\dagger \\ &= (w - w^\dagger)^\top S (w - w^\dagger) \geq 0. \end{aligned} \tag{B.11}$$

Finally, for every z,

$$\begin{aligned} 0 \leq z^\top K z &= \frac{1}{n}\sum_i \mathbb{E}_e \|D_{ie} S^{-1/2} z\|_2^2 \\ &\leq \frac{1}{n}\sum_i \mathbb{E}_e \left[\|D_{ie} S^{-1/2} z\|_2^2 + \|M_{ie} S^{-1/2} z\|_2^2\right] \\ &= z^\top S^{-1/2} S S^{-1/2} z = \|z\|_2^2. \end{aligned} \tag{B.12}$$

No commutation of the unwhitened Gram matrices is needed. □

Categorical responses. For Y_i with class-probability vector p_i, one original annotation gives

$$t_i = Q^\top e_{Y_i}, \qquad \mu_i = Q^\top p_i, \qquad \Sigma_i = Q^\top \left[\operatorname{diag}(p_i) - p_i p_i^\top\right] Q. \tag{B.13}$$

These moments meet the model assumptions. Indeed, $\mathbb{E} e_{Y_i} = p_i$ and $\mathbb{E} e_{Y_i} e_{Y_i}^\top = \operatorname{diag}(p_i)$; subtracting the outer product of the mean proves the covariance formula. Moreover, $Q^\top e_{\pi_e(Y_i)} = R_e Q^\top e_{Y_i}$ because $QQ^\top = I_m - \mathbf{1}\mathbf{1}^\top/m$ and permutations preserve $\mathbf{1}$. Edited targets are therefore deterministic transports of one categorical label. The squared-loss target is $Q^\top p_i$; softmax-optimal logits are represented by $Q^\top \log p_i$ when probabilities are positive.

B.4 Shared noise, exact risk, and strict improvement

Let $v = S^{1/2} w^\dagger$, and let $K = U \operatorname{diag}(\kappa_j) U^\top$ with $\beta = U^\top v$. Define

$$\begin{aligned} \eta &= S^{-1/2}\frac{\overline{M}^\top \varepsilon}{n}, \\ \Omega = \mathbb{E}\eta\eta^\top &= \frac{1}{n^2} S^{-1/2}\left(\sum_i \overline{M}_i^\top \Sigma_i \overline{M}_i\right) S^{-1/2}, \qquad \omega_j = u_j^\top \Omega u_j. \end{aligned} \tag{B.14}$$

The term $\overline{M}_i^\top \Sigma_i \overline{M}_i$ preserves correlations between edits sharing one response. Averaging single-edit covariances as $\mathbb{E}_e[M_{ie}^\top \Sigma_i M_{ie}]$ would discard those correlations.

Theorem B.6 (Exact risk with reused labels)**.** *Let $w^\dagger$ minimize the expected paired squared loss, and let $\mathcal{R}_\lambda = \mathbb{E}\|\widehat{w}_\lambda - w^\dagger\|_S^2$ denote its excess risk. In an orthonormal eigenbasis u_j of K, let $\kappa_j \in [0,1]$, $\beta_j = u_j^\top S^{1/2} w^\dagger$, and $\omega_j = u_j^\top \Omega u_j$, where Ω is the whitened covariance of the original-label noise after edit averaging. For every fixed $\lambda \geq 0$, the exact identity is $\mathcal{R}_\lambda = \sum_j [(\tau\kappa_j\beta_j)^2 + \omega_j]/(1+\tau\kappa_j)^2$.*

Regularization strictly improves risk precisely when its squared bias is smaller than the variance removed. If $Cw^\dagger = 0$, it cannot increase the realized S-error for any noise vector, simultaneously for all $\lambda \geq 0$. For $\lambda > 0$, expected improvement is strict exactly when some direction has $\kappa_j > 0$ and $\omega_j > 0$.

Complete proof of Theorem B.6. Factoring $S + \tau C = S^{1/2}(I_d + \tau K)S^{1/2}$ in the normal equation gives

$$\begin{aligned} S^{1/2}\widehat{w}_\lambda &= F_\tau(v + \eta), \\ S^{1/2}(\widehat{w}_\lambda - w^\dagger) &= b_\tau + F_\tau \eta, \qquad b_\tau = -\tau F_\tau K v. \end{aligned} \tag{B.15}$$

Since $\mathbb{E}\eta = 0$, the bias-noise cross term vanishes. Cyclicity of trace gives

$$\begin{aligned} \mathcal{R}_\lambda &= \mathbb{E}\|b_\tau + F_\tau \eta\|_2^2 \\ &= \|b_\tau\|_2^2 + \operatorname{tr}(F_\tau \Omega F_\tau) \\ &= \|\tau F_\tau K v\|_2^2 + \operatorname{tr}(F_\tau^2 \Omega) \\ &= \sum_j \frac{\tau^2\kappa_j^2\beta_j^2 + \omega_j}{(1+\tau\kappa_j)^2}. \end{aligned} \tag{B.16}$$

Put $G_\tau = \operatorname{tr}[(I_d - F_\tau^2)\Omega]$. The baseline has $\mathcal{R}_0 = \operatorname{tr}\Omega$, so

$$\mathcal{R}_\lambda - \mathcal{R}_0 = \|\tau F_\tau K v\|_2^2 - G_\tau, \qquad G_\tau = \sum_j \omega_j \left[1 - (1+\tau\kappa_j)^{-2}\right]. \tag{B.17}$$

This proves the necessary and sufficient pointwise condition.

If $Cw^\dagger = 0$, then $Kv = 0$. For every realized noise vector and every $\tau \geq 0$,

$$\begin{aligned} 0 \leq \|\widehat{w}_\lambda - w^\dagger\|_S^2 = \eta^\top F_\tau^2 \eta &= \sum_j \frac{(u_j^\top \eta)^2}{(1+\tau\kappa_j)^2} \\ &\leq \sum_j (u_j^\top \eta)^2 = \|\widehat{w}_0 - w^\dagger\|_S^2. \end{aligned} \tag{B.18}$$

For $\tau > 0$, G_τ is positive precisely when some $\kappa_j > 0$ and $\omega_j > 0$. Almost-sure strictness additionally requires $\Pr\{\eta \in \ker K\} = 0$; discrete noise can place positive mass on $\ker K$. □

Connection to Theorem 4.1. Theorem 4.1 follows from Eqs. (B.17)–(B.18). In particular, $\operatorname{Cov}(\widehat{w}_0) = S^{-1/2}\Omega S^{-1/2}$, so $\operatorname{tr}[C \operatorname{Cov}(\widehat{w}_0)] = \operatorname{tr}(K\Omega) = \sum_j \kappa_j\omega_j$. Since these summands are nonnegative, positivity is exactly the strict-improvement criterion above.

Lemma B.7 (Edit averaging and attainable oracle directions)**.** *Define the whitened averaged-design Gram*

$$H_0 = S^{-1/2}\frac{\overline{M}^\top \overline{M}}{n} S^{-1/2}, \qquad \mathcal{E} = \operatorname{range}(H_0). \tag{B.19}$$

As μ varies over real response means, the attainable v are exactly $\mathcal{E}$. In addition,

$$0 \preceq H_0 \preceq I_d - K, \qquad \ker K \subseteq \mathcal{E}, \qquad H_0 z = z \text{ for every } z \in \ker K. \tag{B.20}$$

If $\Sigma_i = \sigma^2 I_h$, then $\Omega = (\sigma^2/n)H_0$.

Proof. With $X = S^{-1/2}\overline{M}^\top/\sqrt{n}$, $XX^\top = H_0$ and $v = X\mu/\sqrt{n}$. The identity $\ker(XX^\top) = \ker X^\top$, obtained from $z^\top XX^\top z = \|X^\top z\|_2^2$, implies $\operatorname{range} X = \operatorname{range} H_0$. For any $v \in \mathcal{E}$, an explicit attaining mean is

$$\mu_v = \overline{M} S^{-1/2} H_0^\dagger v, \qquad S^{-1/2}\frac{\overline{M}^\top \mu_v}{n} = H_0 H_0^\dagger v = v. \tag{B.21}$$

Jensen's inequality in quadratic form yields

$$\begin{aligned} 0 \le z^\top H_0 z &= \frac{1}{n}\sum_i \|\mathbb{E}_e M_{ie} S^{-1/2} z\|_2^2 \\ &\le \frac{1}{n}\sum_i \mathbb{E}_e \|M_{ie} S^{-1/2} z\|_2^2 \\ &= z^\top (I_d - K) z \le \|z\|_2^2. \end{aligned} \tag{B.22}$$

More precisely, because $M_{ie} = A_i + D_{ie}$, their centered edit fluctuations agree. Hence

$$\begin{aligned} &H_0 = I_d - K - J_0, \\ &J_0 = \frac{1}{n} S^{-1/2} \sum_i \mathbb{E}_e \left[(D_{ie} - \mathbb{E}_e D_{ie})^\top (D_{ie} - \mathbb{E}_e D_{ie}) \right] S^{-1/2}, \\ &0 \preceq J_0 \preceq K. \end{aligned} \tag{B.23}$$

The final order follows by subtracting the positive-semidefinite outer product of the mean from the second moment. If $Kz = 0$, then $0 \le z^\top J_0 z \le z^\top K z = 0$, so $J_0 z = 0$ and $H_0 z = z$. Thus z belongs to $\operatorname{range} H_0$. The spherical covariance formula follows by substituting $\Sigma_i = \sigma^2 I_h$ into Eq. (B.14). □

B.5 A SHARP MISMATCH BOUND ON THE ATTAINABLE PARAMETER SPACE

The discrepancy $\Delta = (w^\dagger)^\top C w^\dagger$ measures equivariance mismatch of the best paired linear predictor. The next theorem varies unrestricted real means under a fixed additive-noise law. Categorical covariances depend on their means, so categorical comparisons use the pointwise identity in Theorem B.6. Let V have orthonormal columns spanning $\mathcal{E}$, and use the zero-dimensional convention when this space is trivial. For $\tau > 0$, define

$$\begin{aligned} &K_{\mathcal{E}} = V^\top K V, \qquad B_{\mathcal{E}} = \tau^2 V^\top K F_\tau^2 K V, \\ &L_\tau^\star = \lambda_{\max}\left(K_{\mathcal{E}}^{\dagger/2} B_{\mathcal{E}} K_{\mathcal{E}}^{\dagger/2} \right), \qquad \Delta(\mu) = v^\top K v. \end{aligned} \tag{B.24}$$

Set $L_\tau^\star = 0$ when $K_{\mathcal{E}} = 0$.

Theorem B.8 (Attained worst-case risk difference)**.** *For every $\delta \ge 0$,*

$$\sup_{\mu:\, \Delta(\mu) \le \delta} (\mathcal{R}_\lambda - \mathcal{R}_0) = \delta L_\tau^\star - G_\tau. \tag{B.25}$$

Thus uniform strict improvement is equivalent to $\delta L_\tau^\star < G_\tau$. This formulation includes the degenerate case without dividing by a zero constant.

Proof. Write an attainable v as Vx. Positive semidefiniteness implies $x^\top K_{\mathcal{E}} x = 0$ exactly when $KVx = 0$. Since F_τ is invertible, this is also exactly when $x^\top B_{\mathcal{E}} x = 0$. Thus both matrices have the same kernel, and $B_{\mathcal{E}}$ is supported on $\operatorname{range} K_{\mathcal{E}}$. Let $Z = K_{\mathcal{E}}^{\dagger/2} B_{\mathcal{E}} K_{\mathcal{E}}^{\dagger/2}$. The complete upper-bound chain is

$$\begin{aligned} \mathcal{R}_\lambda - \mathcal{R}_0 &= x^\top B_{\mathcal{E}} x - G_\tau \\ &= (K_{\mathcal{E}}^{1/2} x)^\top Z (K_{\mathcal{E}}^{1/2} x) - G_\tau \\ &\le \lambda_{\max}(Z) \|K_{\mathcal{E}}^{1/2} x\|_2^2 - G_\tau \\ &= L_\tau^\star v^\top K v - G_\tau \\ &\le L_\tau^\star \delta - G_\tau. \end{aligned} \tag{B.26}$$

If $K_{\mathcal{E}} \neq 0$, then $L_\tau^\star > 0$. Choose a unit top eigenvector $z_\star$ of Z in $\operatorname{range} K_{\mathcal{E}}$ and set

$$x_\star = \sqrt{\delta}\, K_{\mathcal{E}}^{\dagger/2} z_\star, \qquad v_\star = V x_\star, \qquad \mu_\star = \overline{M} S^{-1/2} H_0^\dagger v_\star. \tag{B.27}$$

Lemma B.7 proves feasibility, while

$$\begin{aligned} v_\star^\top K v_\star &= \delta z_\star^\top K_{\mathcal{E}}^{\dagger/2} K_{\mathcal{E}} K_{\mathcal{E}}^{\dagger/2} z_\star = \delta, \\ x_\star^\top B_{\mathcal{E}} x_\star &= \delta z_\star^\top Z z_\star = \delta L_\tau^\star. \end{aligned} \tag{B.28}$$

Every inequality is therefore attained. If $K_{\mathcal{E}} = 0$, all attainable oracles have zero bias, and the risk difference is constantly $-G_\tau$, proving the remaining case. □

A simpler envelope and its limitation. All spectral maxima in this paragraph are defined as zero when their index sets are empty. The unrestricted spectral maximum supplies

$$\begin{aligned} \|\tau F_\tau K v\|_2^2 &= \sum_{\kappa_j > 0} (\kappa_j \beta_j^2) \frac{\tau^2 \kappa_j}{(1+\tau\kappa_j)^2} \\ &\leq \left[\max_{\kappa_j > 0} \frac{\tau^2 \kappa_j}{(1+\tau\kappa_j)^2} \right] \Delta(\mu). \end{aligned} \tag{B.29}$$

This can lose sharpness when the maximizing direction is not attainable after edit averaging. The restricted constant in Eq. (B.24) avoids that relaxation. For a single edit per original example, $H_0 = I_d - K$ and $\mathcal{E} = \operatorname{range}(I_d - K)$, so $L_\tau^\star$ reduces to the maximum over $0 < \kappa_j < 1$. Multiple reused edits require the restricted constant above.

Corollary B.9 (Small penalties)**.** *If $\operatorname{tr}(K\Omega) > 0$, then for every fixed finite μ, some sufficiently small positive penalty strictly improves the expected paired risk.*

Proof. Differentiating the exact finite sum at $\tau = 0$ gives

$$\left. \frac{d\mathcal{R}_{\tau/4}}{d\tau} \right|_{\tau=0} = -2 \sum_j \kappa_j \omega_j = -2 \operatorname{tr}(K\Omega) < 0. \tag{B.30}$$

The squared bias has derivative zero there, and continuity proves the result. The admissible positive coefficient depends on the instance. □

B.6 An explicit hierarchy-dependent model with nonzero semantic signal

The following model isolates a guidance-independent output preference from an equivariant semantic predictor. It makes the effect of the prescribed sibling-edit law explicit, while retaining the reuse of each original response across all edited views. In particular, the edit expectation is taken inside the empirical objective, before analyzing its label noise.

Let $\mathcal{H} = \mathbf{1}^\perp \subset \mathbb{R}^m$ and $h = m - 1$. For $p \geq 1$, consider the parameter space $\Theta = \mathcal{L}(\mathbb{R}^p, \mathcal{H}) \oplus \mathcal{H}$, with $\|(W,a)\|_\Theta^2 = \|W\|_F^2 + \|a\|_2^2$ and dimension $d = h(p+1)$. We identify this space with $\mathbb{R}^d$ using orthonormal bases; adjoints and identity operators below are taken in these bases. The centered-score predictor is

$$f_{W,a}(x,g) = P_g W \phi(x) + a, \qquad (W,a) \in \Theta. \tag{B.31}$$

Here P_g is the output-binding permutation relative to the base guidance. The term a represents an output preference that does not follow the guidance. The semantic features $\phi_i \in \mathbb{R}^p$ are fixed independently of the response noise and satisfy

$$\frac{1}{n} \sum_{i=1}^n \phi_i = 0, \qquad \frac{1}{n} \sum_{i=1}^n \phi_i \phi_i^\top = I_p, \qquad n \geq p+1. \tag{B.32}$$

Let $G_1, \ldots, G_J$, with $J \geq 1$, be the disjoint groups of leaf-child outputs belonging to the valid non-root parents in the method. Write $b_v = |G_v| \geq 2$. Assume the prescribed binding edits pass validation.

The edit distribution assigns probability $1/2$ to the representation edit, which has permutation I, and probability $[2J\binom{b_v}{2}]^{-1}$ to each transposition within G_v. Define the within-group contrast projectors

$$\begin{aligned} \Pi_v &= \mathrm{diag}(\mathbf{1}_{G_v}) - \frac{1}{b_v}\mathbf{1}_{G_v}\mathbf{1}_{G_v}^\top, \quad c_v = \frac{1}{2J(b_v-1)}, \\ C_{\mathrm{out}} &= \sum_{v=1}^{J} c_v \Pi_v, \qquad\qquad r = \sum_{v=1}^{J}(b_v-1). \end{aligned} \tag{B.33}$$

The operators Π_v and C_{out} are restricted to $\mathcal{H}$.

Proposition B.10 (Exact geometry and risk under the sibling-edit law)**.** *For each original response $t_i \in \mathcal{H}$, reuse $P_e t_i$ as the response of edit e. Consider the fully edit-averaged paired objective*

$$\begin{aligned} \widehat{\mathcal{L}}_\lambda(W,a) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_e\Big[\frac{1}{4}\big(\|W\phi_i + a - t_i\|_2^2 + \|W\phi_i + P_e^\top a - t_i\|_2^2\big) \\ + \frac{\lambda}{2}\|(P_e^\top - I)a\|_2^2\Big], \qquad \lambda \ge 0. \end{aligned} \tag{B.34}$$

Its paired Gram matrix is $S = I_\Theta$. Its whitened contrast matrix is $K = 0 \oplus C_{\mathrm{out}}$, with eigenvalues c_v, each of multiplicity $b_v - 1$, and zero of multiplicity $d - r$. In particular, $0 < c_v \le 1/2$.

Suppose the original responses satisfy

$$t_i = W_\star \phi_i + a_\star + \varepsilon_i, \qquad C_{\mathrm{out}} a_\star = 0, \qquad \mathbb{E}\varepsilon_i = 0, \qquad \mathrm{Cov}(\varepsilon_i) = \Sigma_i, \tag{B.35}$$

where the noises are independent across original examples and have finite second moments. Let $\tau = 4\lambda$, $L = I_{\mathcal{H}} - C_{\mathrm{out}}$, $F_{\mathrm{out}} = (I_{\mathcal{H}} + \tau C_{\mathrm{out}})^{-1}$, and $\bar{t} = n^{-1}\sum_i t_i$. The unique minimizer is

$$\widehat{W}_\lambda = \frac{1}{n}\sum_{i=1}^{n} t_i \phi_i^\top, \qquad \widehat{a}_\lambda = F_{\mathrm{out}} L \bar{t}. \tag{B.36}$$

It is unbiased for $(W_\star, a_\star)$. Define

$$\begin{aligned} \eta &= \left(\frac{1}{n}\sum_i \varepsilon_i \phi_i^\top, \frac{L}{n}\sum_i \varepsilon_i\right), \qquad \Omega = \mathrm{Cov}(\eta), \\ F &= (I_\Theta + \tau K)^{-1}, \qquad\qquad \mathcal{R}_\lambda = \mathbb{E}\big[\|\widehat{W}_\lambda - W_\star\|_F^2 + \|\widehat{a}_\lambda - a_\star\|_2^2\big]. \end{aligned} \tag{B.37}$$

Then $\mathcal{R}_\lambda = \mathrm{tr}(F\Omega F)$. This identity does not require Gaussian noise or isotropic covariances.

If additionally $\Sigma_i = \sigma^2 I_{\mathcal{H}}$ for every i, with $\sigma > 0$, then

$$\begin{aligned} \Omega &= \frac{\sigma^2}{n}(I_\Theta - K)^2, \\ \mathcal{R}_\lambda &= \frac{\sigma^2}{n}\left[d - r + \sum_{v=1}^{J}\frac{(b_v-1)(1-c_v)^2}{(1+4\lambda c_v)^2}\right]. \end{aligned} \tag{B.38}$$

Every $\lambda > 0$ strictly improves this risk relative to paired supervision with $\lambda = 0$. The baseline and limiting risks are

$$\begin{aligned} \mathcal{R}_0 &= \frac{\sigma^2}{n}\left[d - 1 + \sum_{v=1}^{J}\frac{1}{4J^2(b_v-1)}\right], \\ \lim_{\lambda\to\infty}\mathcal{R}_\lambda &= \frac{\sigma^2}{n}(d-r). \end{aligned} \tag{B.39}$$

Proof roadmap. Averaging transpositions within each sibling group gives a complete-graph Laplacian with explicit eigenvalues and multiplicities. Feature centering separates the semantic and offset Gram blocks. The edit-averaged noise operator on offsets is $I_{\mathcal{H}} - C_{\mathrm{out}}$, so its covariance contains $(I_{\mathcal{H}} - C_{\mathrm{out}})^2$. Label reuse therefore squares the noise factor $1 - c_v$, giving $(1-c_v)^2$ in Eq. (B.38).

Proof. For a transposition of outputs a, b, write $q_{ab} = e_a - e_b$. Its matrix satisfies $I - P_{ab} = q_{ab}q_{ab}^\top$ and $P_{ab}^\top = P_{ab}$. The complete-graph Laplacian on G_v obeys

$$\begin{aligned} \sum_{\{a,b\}\subset G_v} q_{ab}q_{ab}^\top &= b_v \operatorname{diag}(\mathbf{1}_{G_v}) - \mathbf{1}_{G_v}\mathbf{1}_{G_v}^\top = b_v\Pi_v, \\ \mathbb{E}[I - P_e \mid v, \text{binding edit}] &= \frac{b_v}{\binom{b_v}{2}}\Pi_v = \frac{2}{b_v - 1}\Pi_v. \end{aligned} \tag{B.40}$$

For every supported edit, $P_e^2 = I$. Therefore its mean and contrast operators on the offset are orthogonal projectors up to the sign of the contrast:

$$\begin{aligned} J_e = \frac{I + P_e}{2}, \qquad N_e = \frac{P_e - I}{2}, \\ J_e^\top J_e = J_e, \qquad N_e^\top N_e = \frac{I - P_e}{2}, \\ J_e^\top J_e + N_e^\top N_e = I. \end{aligned} \tag{B.41}$$

Combining Eq. (B.40) with the half binding mixture gives the exact constants

$$\begin{aligned} \mathbb{E}_e[N_e^\top N_e] &= \frac{1}{4}\mathbb{E}[I - P_e \mid \text{binding edit}] \\ &= \frac{1}{4J}\sum_{v=1}^{J}\frac{2}{b_v - 1}\Pi_v = C_{\text{out}}, \\ \mathbb{E}_e[J_e] &= I - C_{\text{out}} = L. \end{aligned} \tag{B.42}$$

For $\vartheta = (W, a)$, the paired mean and contrast maps are $M_{i,e}\vartheta = W\phi_i + J_e a$ and $D_{i,e}\vartheta = N_e a$. The feature design in Eq. (B.32) makes the cross term vanish. More explicitly,

$$\begin{aligned} \langle \vartheta, S\vartheta\rangle &= \frac{1}{n}\sum_i \mathbb{E}_e\left[\|W\phi_i + J_e a\|_2^2 + \|N_e a\|_2^2\right] \\ &= \frac{1}{n}\sum_i \|W\phi_i\|_2^2 + 2\left\langle W\frac{1}{n}\sum_i \phi_i, La\right\rangle + \|a\|_2^2 \\ &= \operatorname{tr}\left(W\left[\frac{1}{n}\sum_i \phi_i\phi_i^\top\right]W^\top\right) + \|a\|_2^2 \\ &= \|W\|_F^2 + \|a\|_2^2 = \|\vartheta\|_\Theta^2. \end{aligned} \tag{B.43}$$

Since S is self-adjoint, equality of these quadratic forms gives $S = I_\Theta$. Similarly,

$$\langle \vartheta, C\vartheta\rangle = \frac{1}{n}\sum_i \mathbb{E}_e\|D_{i,e}\vartheta\|_2^2 = a^\top C_{\text{out}} a, \qquad K = C = 0 \oplus C_{\text{out}}. \tag{B.44}$$

Each Π_v is symmetric and idempotent, has range $\{z : \operatorname{supp}(z) \subseteq G_v,\ \mathbf{1}_{G_v}^\top z = 0\}$, and has rank $b_v - 1$. Disjointness of the G_v gives $\Pi_v\Pi_u = 0$ for $u \neq v$. Consequently, Eq. (B.33) is an orthogonal spectral decomposition. It has r positive eigenvalues, with the stated multiplicities, and $h - r$ zero eigenvalues on $\mathcal{H}$. The hp semantic parameter directions supply the remaining zeros. The inequalities $J \geq 1$ and $b_v - 1 \geq 1$ give $0 < c_v \leq 1/2$.

The average of the mean-view map is $\bar{M}_i(W, a) = W\phi_i + La$. Its adjoint acts on $z \in \mathcal{H}$ as $\bar{M}_i^* z = (z\phi_i^\top, Lz)$. The fully edit-averaged normal equation is therefore

$$\begin{aligned} (I_\Theta + \tau K)(\widehat{W}_\lambda, \widehat{a}_\lambda) &= \frac{1}{n}\sum_i \bar{M}_i^* t_i \\ &= \left(\frac{1}{n}\sum_i t_i\phi_i^\top, L\bar{t}\right). \end{aligned} \tag{B.45}$$

The operator on the left is positive definite for every $\tau \geq 0$. Solving its two blocks proves Eq. (B.36). Under Eq. (B.35), centering and whitening give $n^{-1}\sum_i \mathbb{E}[t_i]\phi_i^\top = W_\star$ and $\mathbb{E}\bar{t} = a_\star$. Since

$C_{\text{out}}a_\star = 0$, both L and F_{out} fix $a_\star$. Thus the estimator is unbiased, and its entire error vector is exactly

$$(\widehat{W}_\lambda - W_\star, \widehat{a}_\lambda - a_\star) = F\eta, \qquad \mathcal{R}_\lambda = \mathbb{E}\|F\eta\|_\Theta^2 = \operatorname{tr}(F\Omega F). \tag{B.46}$$

For general Σ_i, the covariance retains dependence between output coordinates. Let $\eta_{W,k} \in \mathcal{H}$ be column k of the matrix block of η, and let η_a be its offset block. Independence across original examples yields

$$\begin{aligned}
\operatorname{Cov}(\eta_{W,k}, \eta_{W,\ell}) &= \frac{1}{n^2}\sum_i \phi_{ik}\phi_{i\ell}\Sigma_i, \\
\operatorname{Cov}(\eta_{W,k}, \eta_a) &= \frac{1}{n^2}\sum_i \phi_{ik}\Sigma_i L, \\
\operatorname{Cov}(\eta_a) &= \frac{1}{n^2} L\left(\sum_i \Sigma_i\right) L.
\end{aligned} \tag{B.47}$$

If every $\Sigma_i = \sigma^2 I_{\mathcal{H}}$, the feature moment identities reduce these three blocks respectively to $(\sigma^2/n)\delta_{k\ell} I_{\mathcal{H}}$, zero, and $(\sigma^2/n)L^2$. Hence

$$\Omega = \frac{\sigma^2}{n}\left(I_{\mathcal{L}(\mathbb{R}^p,\mathcal{H})} \oplus L^2\right) = \frac{\sigma^2}{n}(I_\Theta - K)^2. \tag{B.48}$$

The square on $I_\Theta - K$ arises because every edit of example i reuses ε_i: the covariance contains the averaged operator $\bar{M}_i$ on both sides.

Substituting the spectral decomposition into Eq. (B.46) proves Eq. (B.38). For $\lambda > 0$, the gain is

$$\begin{aligned}
\mathcal{R}_0 - \mathcal{R}_\lambda &= \frac{\sigma^2}{n}\sum_{v=1}^{J}(b_v - 1)(1 - c_v)^2\left[1 - \frac{1}{(1 + 4\lambda c_v)^2}\right] \\
&> 0,
\end{aligned} \tag{B.49}$$

because each weight is positive and each bracket is positive. Expanding the baseline risk gives:

$$\begin{aligned}
d - r + \sum_v (b_v - 1)(1 - c_v)^2 &= d - 2\sum_v (b_v - 1)c_v + \sum_v (b_v - 1)c_v^2 \\
&= d - 2\sum_v \frac{1}{2J} + \sum_v \frac{1}{4J^2(b_v - 1)} \\
&= d - 1 + \sum_v \frac{1}{4J^2(b_v - 1)}.
\end{aligned} \tag{B.50}$$

Every positive-eigenvalue contribution converges to zero as $\lambda \to \infty$, proving the limiting expression. □

A nonzero categorical instance of the semantic model. A categorical response model also supports a nonzero equivariant signal. Fix features satisfying Eq. (B.32), and choose a nonzero $W_0 \in \mathcal{L}(\mathbb{R}^p, \mathcal{H})$. Define $B = \max_i \|W_0\phi_i\|_\infty$. This number is strictly positive, because

$$\frac{1}{n}\sum_i \|W_0\phi_i\|_2^2 = \operatorname{tr}\left(W_0\left[\frac{1}{n}\sum_i \phi_i\phi_i^\top\right]W_0^\top\right) = \|W_0\|_F^2 > 0. \tag{B.51}$$

Choose $0 < \nu \le (2mB)^{-1}$ and set

$$W_\star = \nu W_0, \qquad a_\star = 0, \qquad \mathbf{p}_i = \frac{1}{m}\mathbf{1} + W_\star\phi_i. \tag{B.52}$$

These are strictly positive probability vectors. Indeed, for every coordinate c,

$$[\mathbf{p}_i]_c \ge \frac{1}{m} - \nu\|W_0\phi_i\|_\infty \ge \frac{1}{m} - \nu B \ge \frac{1}{2m} > 0, \qquad \mathbf{1}^\top\mathbf{p}_i = 1. \tag{B.53}$$

Draw independent $Y_i \sim \mathrm{Categorical}(\mathbf{p}_i)$ and take $t_i = e_{Y_i} - \mathbf{1}/m$. Then $\mathbb{E}[t_i] = W_\star \phi_i$ is nonzero for at least one input, and the same realized response has edited version $P_e t_i$. The centered categorical noise has covariance

$$\Sigma_i = \mathrm{diag}(\mathbf{p}_i) - \mathbf{p}_i \mathbf{p}_i^\top \quad \text{on } \mathcal{H}. \tag{B.54}$$

This categorical example uses the general covariance identity in Eq. (B.46); the isotropic formula in Eq. (B.38) requires additional covariance assumptions.

The categorical example also has a strict expected squared-risk benefit for every $\lambda > 0$. To see this directly, for any nonzero $z \in \mathcal{H}$ let $\bar{z}_i = \sum_c [\mathbf{p}_i]_c z_c$. Strict positivity of the probabilities gives

$$z^\top \Sigma_i z = \sum_c [\mathbf{p}_i]_c (z_c - \bar{z}_i)^2 > 0. \tag{B.55}$$

Equality would require all z_c to coincide, which is incompatible with nonzero $z \in \mathbf{1}^\perp$. For a unit $z \in \mathrm{range}(\Pi_v)$, the corresponding offset noise direction therefore has variance

$$\mathrm{Var}(z^\top \eta_a) = \frac{(1 - c_v)^2}{n^2} \sum_i z^\top \Sigma_i z > 0. \tag{B.56}$$

Since F strictly contracts this direction, the trace identity gives $\mathcal{R}_0 - \mathcal{R}_\lambda = \mathrm{tr}[\Omega(I_\Theta - F^2)] > 0$. This gives a finite-design categorical instance with nonzero signal and a strict squared-risk benefit.

Consequences of the sibling-edit law. The penalty leaves hp semantic directions available for label-based learning and removes only the r within-group offset directions in the hard-constraint limit. The remaining $h - r$ offset directions are constant within each edited group; they include contrasts between distinct groups and coordinates associated with uncovered labels, subject to global centering. Sibling edits cannot identify these directions through equivariance alone. The factors $1 + 4\lambda c_v = 1 + 2\lambda/[J(b_v - 1)]$ also expose the finite-penalty effect of the group sizes and the specified sampling probabilities. Comparing arbitrary trees by prediction risk requires further analysis. The semantic term in Eq. (B.31) is equivariant by construction, and representation edits leave it unchanged. The proposition thus isolates correction of a guidance-independent output preference. Learning semantic features and representation invariance in a nonlinear language model lies outside this proxy.

B.7 Effective dimension and an exact annotation-budget comparison

Assume $Cw^\dagger = 0$ and $\Sigma_i = \sigma^2 I_h$, with $\sigma > 0$. Second moments suffice in this subsection. We use H_0 from Lemma B.7, which accounts for label reuse across edits and can differ from $I_d - K$. Define

$$d_{\mathrm{eff}}(\tau) = \mathrm{tr}(F_\tau^2 H_0), \qquad p_0 = \dim \ker K, \qquad h_j = u_j^\top H_0 u_j. \tag{B.57}$$

Proposition B.11 (Effective dimension and replicated-design budgets)**.** *The exact risk and its hard-constraint limit are*

$$\mathcal{R}_\lambda = \frac{\sigma^2}{n} d_{\mathrm{eff}}(\tau), \qquad \lim_{\tau \to \infty} d_{\mathrm{eff}}(\tau) = p_0. \tag{B.58}$$

If $\mathcal{A} = \{j : \kappa_j > 0,\ h_j > 0\}$ is nonempty, let $\gamma = \min_{j \in \mathcal{A}} \kappa_j$. Then

$$p_0 \le d_{\mathrm{eff}}(\tau) \le p_0 + \frac{d_{\mathrm{eff}}(0) - p_0}{(1 + \tau\gamma)^2} \le d_{\mathrm{eff}}(0). \tag{B.59}$$

For $\tau > 0$, the middle inequality is an equality exactly when every active eigenvalue equals γ. If $\mathcal{A}$ is empty, $d_{\mathrm{eff}}(\tau) = p_0$ for every τ.

For a fixed base design of size n_0, repeat each original design point k times using independent response draws with the same means and covariances. Use the same edit law for each repeat, so $n = kn_0$. For a target risk $\epsilon > 0$, the smallest admissible annotation budget is exactly

$$n_\lambda(\epsilon) = n_0 \max\left\{1, \left\lceil \frac{\sigma^2 d_{\mathrm{eff}}(\tau)}{n_0 \epsilon} \right\rceil \right\}. \tag{B.60}$$

Proof. By Lemma B.7, $\Omega = (\sigma^2/n)H_0$. The zero-bias case of Theorem B.6 proves the risk identity. For a zero-eigenvalue unit vector, $H_0 u_j = u_j$, so $h_j = 1$. For a positive eigenvalue, $0 \le h_j \le 1 - \kappa_j$. The complete spectral comparison is

$$\begin{aligned} p_0 &\le p_0 + \sum_{j \in \mathcal{A}} \frac{h_j}{(1+\tau\kappa_j)^2} = d_{\text{eff}}(\tau) \\ &\le p_0 + \frac{\sum_{j \in \mathcal{A}} h_j}{(1+\tau\gamma)^2} \\ &= p_0 + \frac{d_{\text{eff}}(0) - p_0}{(1+\tau\gamma)^2} \le d_{\text{eff}}(0). \end{aligned} \tag{B.61}$$

Every weight in $\mathcal{A}$ is positive. For $\tau > 0$, strict monotonicity of $(1+\tau x)^{-2}$ in x proves the equality criterion. Taking the limit in the finite sum proves the floor.

Replicating the design leaves $S, C, \overline{M}^\top \overline{M}/n$, H_0, and $w^\dagger$ unchanged. Each replicate uses a fresh independent original response, whereas all its edited views share that response. Thus the noise covariance scales as $1/k$. In particular,

$$\begin{aligned} \mathcal{R}_\lambda(kn_0) \le \epsilon \quad &\Longleftrightarrow \quad k \ge \frac{\sigma^2 d_{\text{eff}}(\tau)}{n_0 \epsilon}, \\ \min\{k \in \mathbb{N} : k \ge 1,\ \mathcal{R}_\lambda(kn_0) \le \epsilon\} &= \max\left\{1, \left\lceil \frac{\sigma^2 d_{\text{eff}}(\tau)}{n_0 \epsilon} \right\rceil \right\}. \end{aligned} \tag{B.62}$$

This proves the exact budget, including a zero effective dimension. □

The exact comparison between two penalties uses Eq. (B.60) for both. If only integer rounding is relaxed, their budget ratio is

$$\frac{\max\{n_0,\ \sigma^2 d_{\text{eff}}(\tau)/\epsilon\}}{\max\{n_0,\ \sigma^2 d_{\text{eff}}(0)/\epsilon\}}. \tag{B.63}$$

It reduces to $d_{\text{eff}}(\tau)/d_{\text{eff}}(0)$ when both unconstrained budgets are at least n_0 and the denominator is positive. This budget counts independent response draws at fixed inputs. With deterministic labels, querying the same input adds no response variation; transforming its annotation also reuses the same draw. Sample complexity for new inputs requires a separate analysis.

Prediction geometry matters. The loss analyzed above is the paired S-risk. For a correctly specified equivariant signal $w_\star$, suppose an unseen-input linear predictor has test Gram $T \succeq 0$. Set $A_T = S^{-1/2} T S^{-1/2}$. The same zero-bias error representation gives

$$\begin{aligned} \mathbb{E}\|\widehat{w}_\lambda - w_\star\|_T^2 &= \operatorname{tr}(A_T F_\tau \Omega F_\tau), \\ \mathbb{E}\|\widehat{w}_0 - w_\star\|_T^2 - \mathbb{E}\|\widehat{w}_\lambda - w_\star\|_T^2 &= \operatorname{tr}[A_T(\Omega - F_\tau \Omega F_\tau)]\,. \end{aligned} \tag{B.64}$$

These equalities follow by expanding the quadratic form and taking its trace. Nonnegativity for every test Gram is equivalent to $\Omega - F_\tau \Omega F_\tau \succeq 0$: sufficiency follows from the nonnegative trace of a product of positive semidefinite matrices; necessity follows by taking $A_T = zz^\top$ for each vector z. Commutation of Ω and K is sufficient, since a common orthogonal eigenbasis then gives entries $\omega_j[1 - (1+\tau\kappa_j)^{-2}] \ge 0$. Improvement on unseen inputs therefore depends on the test geometry. The fixed-design identities concern squared prediction risk; guarantees for classification error require further assumptions.

B.8 A Gaussian probability guarantee for the coupled risk gain

Proposition B.12 (A direct high-probability gain under equivariance mismatch)**.** *Condition on the features and edits, and fix $\tau \ge 0$ independently of the training noise. For this concentration result, additionally assume that the whitened original noise is Gaussian, $\eta \sim \mathcal{N}(0, \Omega)$, where $\Omega \succeq 0$ may be singular. Let $F_\tau = (I_d + \tau K)^{-1}$ and define*

$$\begin{aligned} b_\tau &= -\tau F_\tau K v, & H_\tau &= I_d - F_\tau^2, \\ \mathsf{J}_\tau &= \Omega^{1/2} H_\tau \Omega^{1/2}, & \ell_\tau &= -\Omega^{1/2} F_\tau b_\tau, \\ \Gamma_\tau &= \operatorname{tr}(\mathsf{J}_\tau) - \|b_\tau\|_2^2, & V_\tau &= \operatorname{tr}(\mathsf{J}_\tau^2) + 2\|\ell_\tau\|_2^2. \end{aligned} \tag{B.65}$$

For the two estimators trained using the same original responses, let $Q_\lambda = \|\widehat{w}_\lambda - w^\dagger\|_S^2$ *and* $\widehat{G}_\lambda = Q_0 - Q_\lambda$. *Then* $\mathbb{E}\widehat{G}_\lambda = \Gamma_\tau$, *and, for every* $\delta \in (0,1)$,

$$\Pr\Big\{\widehat{G}_\lambda \geq \Gamma_\tau - 2\sqrt{V_\tau \log(1/\delta)}\Big\} \geq 1-\delta. \tag{B.66}$$

Consequently, $\Gamma_\tau > 2\sqrt{V_\tau \log(1/\delta)}$ *is sufficient for strict realized improvement with probability at least* $1-\delta$. *If* $V_\tau = 0$, *then* $\widehat{G}_\lambda = \Gamma_\tau$ *almost surely; this deterministic case is included in* (B.66). *No commutation between* Ω *and* K *is required.*

Proof. The whitened normal equation gives the coupled error representation

$$\begin{aligned} S^{1/2}(\widehat{w}_0 - w^\dagger) &= \eta, \\ S^{1/2}(\widehat{w}_\lambda - w^\dagger) &= F_\tau \eta + b_\tau. \end{aligned} \tag{B.67}$$

Since $K \succeq 0$, its eigenvalues κ_j are nonnegative and F_τ is symmetric. Its eigenvalues are $(1+\tau\kappa_j)^{-1} \in (0,1]$, so

$$H_\tau = U \operatorname{diag}\big(1-(1+\tau\kappa_j)^{-2}\big)_{j=1}^d U^\top \succeq 0, \qquad \mathsf{J}_\tau = \Omega^{1/2} H_\tau \Omega^{1/2} \succeq 0. \tag{B.68}$$

Here $\Omega^{1/2}$ is the symmetric positive semidefinite square root, which is defined for singular covariances as well.

Represent $\eta = \Omega^{1/2} Z$, where $Z \sim \mathcal{N}(0, I_d)$. The same η occurs in both estimator errors. Expanding their squared norms gives the exact identity

$$\begin{aligned} \widehat{G}_\lambda &= \|\eta\|_2^2 - \|F_\tau \eta + b_\tau\|_2^2 \\ &= \eta^\top H_\tau \eta - 2 b_\tau^\top F_\tau \eta - \|b_\tau\|_2^2 \\ &= Z^\top \mathsf{J}_\tau Z + 2\ell_\tau^\top Z - \|b_\tau\|_2^2 \\ &= \Gamma_\tau + Z^\top \mathsf{J}_\tau Z - \operatorname{tr}(\mathsf{J}_\tau) + 2\ell_\tau^\top Z. \end{aligned} \tag{B.69}$$

The identities $\mathbb{E}Z = 0$ and $\mathbb{E}ZZ^\top = I_d$ imply $\mathbb{E}\widehat{G}_\lambda = \Gamma_\tau$.

Take an orthogonal eigendecomposition of the gain matrix: write $\mathsf{J}_\tau = W \operatorname{diag}(\rho_1, \dots, \rho_d) W^\top$ with W orthogonal and $\rho_j \geq 0$. Set $\zeta = W^\top Z$ and $\widetilde{\ell} = W^\top \ell_\tau$. Then the ζ_j are independent standard Gaussian variables, and

$$\begin{aligned} X := \widehat{G}_\lambda - \Gamma_\tau &= \sum_{j=1}^d \big[\rho_j(\zeta_j^2 - 1) + 2\widetilde{\ell}_j \zeta_j\big], \\ \sum_{j=1}^d (\rho_j^2 + 2\widetilde{\ell}_j^2) &= \operatorname{tr}(\mathsf{J}_\tau^2) + 2\|\ell_\tau\|_2^2 = V_\tau. \end{aligned} \tag{B.70}$$

If $V_\tau = 0$, all ρ_j and $\widetilde{\ell}_j$ vanish, and $X = 0$ almost surely. This proves the deterministic conclusion. Henceforth assume $V_\tau > 0$.

Let $c = \max_j \rho_j$, with $(2c)^{-1} = +\infty$ when $c = 0$. For $t < (2c)^{-1}$, completing the square in the Gaussian integral gives

$$\begin{aligned} &\mathbb{E}\exp\Big\{t[\rho_j(\zeta_j^2 - 1) + 2\widetilde{\ell}_j \zeta_j]\Big\} \\ &\quad = \frac{e^{-t\rho_j}}{\sqrt{2\pi}} \int_{\mathbb{R}} \exp\Big\{-\frac{1-2t\rho_j}{2} z^2 + 2t\widetilde{\ell}_j z\Big\}\, dz \\ &\quad = \frac{\exp\Big\{-t\rho_j + \frac{2t^2\widetilde{\ell}_j^2}{1-2t\rho_j}\Big\}}{\sqrt{1-2t\rho_j}}. \end{aligned} \tag{B.71}$$

Independence gives the exact cumulant generating function

$$\psi(t) := \log \mathbb{E}e^{tX} = \sum_{j=1}^d \left[-t\rho_j - \frac{1}{2}\log(1-2t\rho_j) + \frac{2t^2\widetilde{\ell}_j^2}{1-2t\rho_j}\right]. \tag{B.72}$$

Only negative arguments are needed for the lower confidence bound. For every $a \geq 0$,

$$0 \leq a - \log(1+a) = \int_0^a \frac{z}{1+z}\, dz \leq \int_0^a z\, dz = \frac{a^2}{2}. \tag{B.73}$$

Consequently, for every $s > 0$,

$$\begin{aligned} \psi(-s) &= \sum_{j=1}^{d} \left[\frac{2s\rho_j - \log(1+2s\rho_j)}{2} + \frac{2s^2\widetilde{\ell}_j^2}{1+2s\rho_j} \right] \\ &\leq \sum_{j=1}^{d} \left[s^2\rho_j^2 + \frac{2s^2\widetilde{\ell}_j^2}{1+2s\rho_j} \right] \\ &\leq s^2 \sum_{j=1}^{d} (\rho_j^2 + 2\widetilde{\ell}_j^2) = s^2 V_\tau. \end{aligned} \tag{B.74}$$

Let $x = \log(1/\delta) > 0$ and $s_x = \sqrt{x/V_\tau}$. Markov's inequality yields

$$\begin{aligned} \Pr\{X \leq -2\sqrt{V_\tau x}\} &= \Pr\{e^{-s_x X} \geq e^{2s_x\sqrt{V_\tau x}}\} \\ &\leq \exp\Big\{-2s_x\sqrt{V_\tau x} + \psi(-s_x)\Big\} \\ &\leq \exp\Big\{-2s_x\sqrt{V_\tau x} + s_x^2 V_\tau\Big\} \\ &= e^{-x} = \delta. \end{aligned} \tag{B.75}$$

This proves (B.66) and its strict-improvement consequence. The argument bounds the coupled gain directly through the shared noise vector. □

Corollary B.13 (Exact equivariance)**.** *If $Kv = 0$, then $b_\tau = 0$ and $\widehat{G}_\lambda = \eta^\top H_\tau \eta \geq 0$ for every noise realization. Under the Gaussian assumption of Proposition B.12,*

$$\begin{aligned} \widehat{G}_\lambda &\overset{\mathrm{d}}{=} \sum_{j=1}^{d} \rho_j \zeta_j^2, \\ \Pr\Big\{\widehat{G}_\lambda \geq \max\Big\{0,\ \mathrm{tr}(\mathsf{J}_\tau) - 2\sqrt{\mathrm{tr}(\mathsf{J}_\tau^2)\log(1/\delta)}\Big\}\Big\} &\geq 1-\delta. \end{aligned} \tag{B.76}$$

Here the ρ_j are the eigenvalues of J_τ and the ζ_j are independent standard Gaussian variables.

Proof. The condition $Kv = 0$ gives $b_\tau = 0$ and $\ell_\tau = 0$. Equation (B.69) and $H_\tau \succeq 0$ prove nonnegativity without a distributional assumption on the noise. Under Gaussian noise, (B.70) gives the stated law. Substituting $\Gamma_\tau = \mathrm{tr}(\mathsf{J}_\tau)$ and $V_\tau = \mathrm{tr}(\mathsf{J}_\tau^2)$ into (B.66), then using deterministic nonnegativity, proves the confidence bound. If $\mathsf{J}_\tau = 0$, the gain is zero almost surely under the Gaussian model, which includes the degenerate case. □

The closed-form confidence bound is conservative: the comparison in (B.74) discards distributional information. Using the exact cumulant gives, for every $r > 0$,

$$\Pr\{\widehat{G}_\lambda \leq \Gamma_\tau - r\} \leq \inf_{s>0} \exp\{-sr + \psi(-s)\}. \tag{B.77}$$

This one-dimensional optimization gives a Chernoff bound using all eigenvalues and linear terms. Exact conditional quantiles under equivariance follow from the weighted chi-square law. Gaussianity is used for these distributional and concentration statements; the mean identity follows directly from the zero mean and covariance of η.

B.9 An attained Gaussian minimax floor on the equivariant subspace

The following Gaussian experiment has an unrestricted equivariant parameter. Matching upper and lower bounds establish the limiting variance term for this experiment.

Theorem B.14 (Exact minimax risk with one original response per input)**.** *Fix the designs and edit laws from Appendix B.3 and let $\Theta_0 = \ker C$, with $p_0 = \dim \Theta_0$. For $w_\star \in \Theta_0$, observe*

$$t_i = A_i w_\star + \varepsilon_i, \qquad \varepsilon_i \overset{\text{ind}}{\sim} \mathcal{N}(0, \sigma^2 I_h), \qquad \sigma > 0. \tag{B.78}$$

All edited responses are the known deterministic transports $R_e t_i$. For all measurable estimators based on these observations,

$$\inf_{\widetilde{w}} \sup_{w_\star \in \Theta_0} \mathbb{E}_{w_\star} \|\widetilde{w} - w_\star\|_S^2 = \frac{\sigma^2 p_0}{n}. \tag{B.79}$$

The constrained least-squares estimator attains this value. It is also the limit of $\widehat{w}_\lambda$ as $\lambda \to \infty$.

Proof. First identify the information in the experiment. For any $w \in \Theta_0$,

$$0 = w^\top C w = \frac{1}{n} \sum_i \sum_{e:\rho_i(e)>0} \rho_i(e) \|D_{ie} w\|_2^2. \tag{B.80}$$

Every summand is nonnegative. Hence $D_{ie} w = 0$ on the edit support, which gives $A_i w = B_{ie} w = M_{ie} w = \overline{M}_i w$. The original responses determine every edited response, and the collection of paired responses includes the originals. Thus both collections generate the same observation sigma-field. Adding edited targets does not increase the available statistical information.

Let $U_0 \in \mathbb{R}^{d \times p_0}$ have orthonormal columns spanning $\ker K$. Parameterize $w_\star = S^{-1/2} U_0 \theta$, with $\theta \in \mathbb{R}^{p_0}$, and stack the original designs as A. Lemma B.7 and the preceding kernel identity give

$$\begin{aligned} X_0 &= A S^{-1/2} U_0 = \overline{M} S^{-1/2} U_0, \\ \frac{1}{n} X_0^\top X_0 &= U_0^\top H_0 U_0 = I_{p_0}, \\ t &= X_0 \theta + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_{nh}). \end{aligned} \tag{B.81}$$

Thus all p_0 equivariant coordinates are identifiable from the original responses.

For the upper bound, define

$$\widehat{\theta} = \frac{1}{n} X_0^\top t, \qquad \widehat{w}_\infty = S^{-1/2} U_0 \widehat{\theta}. \tag{B.82}$$

Since $X_0^\top X_0 = n I_{p_0}$, direct calculation gives

$$\begin{aligned} \mathbb{E}_\theta \|\widehat{w}_\infty - w_\star\|_S^2 &= \mathbb{E}_\theta \|\widehat{\theta} - \theta\|_2^2 \\ &= \frac{1}{n^2} \mathbb{E} \|X_0^\top \varepsilon\|_2^2 \\ &= \frac{\sigma^2}{n^2} \operatorname{tr}(X_0^\top X_0) = \frac{\sigma^2 p_0}{n}. \end{aligned} \tag{B.83}$$

This risk is constant in θ.

For the lower bound, take an arbitrary estimator $\widetilde{w}(t)$ and let $\widetilde{\theta}(t) = U_0^\top S^{1/2} \widetilde{w}(t)$. Orthogonal decomposition gives the pointwise comparison

$$\begin{aligned} \|\widetilde{w} - w_\star\|_S^2 &= \|\widetilde{\theta} - \theta\|_2^2 + \|(I_d - U_0 U_0^\top) S^{1/2} \widetilde{w}\|_2^2 \\ &\geq \|\widetilde{\theta} - \theta\|_2^2 \geq 0. \end{aligned} \tag{B.84}$$

Use a proper Gaussian prior $\theta \sim \mathcal{N}(0, \xi^2 I_{p_0})$ for any finite $\xi > 0$. The likelihood and prior densities yield a Gaussian posterior by completing the square:

$$\begin{aligned} \pi(\theta \mid t) &\propto \exp\left\{ -\frac{\|t - X_0\theta\|_2^2}{2\sigma^2} - \frac{\|\theta\|_2^2}{2\xi^2} \right\} \\ &\propto \exp\left\{ -\frac{\|\theta - m_\xi(t)\|_2^2}{2 v_\xi} \right\}, \\ v_\xi &= \left(\frac{n}{\sigma^2} + \frac{1}{\xi^2} \right)^{-1}, \qquad m_\xi(t) = \frac{v_\xi}{\sigma^2} X_0^\top t. \end{aligned} \tag{B.85}$$

For every observed t, centering the squared error at the posterior mean cancels its mixed term. Consequently,

$$\begin{aligned}\mathbb{E}[\|\widetilde{\theta}(t)-\theta\|_2^2 \mid t] &= \|\widetilde{\theta}(t)-m_\xi(t)\|_2^2 + \mathbb{E}[\|\theta-m_\xi(t)\|_2^2 \mid t] \\ &\geq \operatorname{tr}(v_\xi I_{p_0}) = p_0 v_\xi.\end{aligned} \tag{B.86}$$

Integrating this nonnegative loss under the prior and using Eq. (B.84) proves the full lower-bound chain:

$$\begin{aligned}\sup_{w_\star\in\Theta_0}\mathbb{E}_{w_\star}\|\widetilde{w}-w_\star\|_S^2 &\geq \sup_{\theta\in\mathbb{R}^{p_0}}\mathbb{E}_\theta\|\widetilde{\theta}-\theta\|_2^2 \\ &\geq \int \mathbb{E}_\theta\|\widetilde{\theta}-\theta\|_2^2\,\pi_\xi(d\theta) \\ &= \mathbb{E}_t\mathbb{E}[\|\widetilde{\theta}(t)-\theta\|_2^2 \mid t] \\ &\geq \frac{p_0}{n/\sigma^2+1/\xi^2}.\end{aligned} \tag{B.87}$$

The bound applies to every estimator and every finite ξ. Taking the infimum over estimators and then $\xi\to\infty$ gives $\sigma^2 p_0/n$. Together with Eq. (B.83), this proves the exact minimax value.

Finally, minimizing paired loss within Θ_0 is ordinary least squares in $X_0\theta$, since the contrast vanishes there. It therefore gives Eq. (B.82). The spectral limit $F_\tau\to U_0U_0^\top$ also gives

$$\begin{aligned}\lim_{\lambda\to\infty}\widehat{w}_\lambda &= S^{-1/2}U_0U_0^\top S^{-1/2}\frac{\overline{M}^\top t}{n} \\ &= S^{-1/2}U_0\frac{X_0^\top t}{n} = \widehat{w}_\infty.\end{aligned} \tag{B.88}$$

When $p_0=0$, the parameter class contains only zero; the zero estimator has risk zero and the same formulas hold with empty matrices. □

The analysis gives an exact risk identity, an attained worst-case mismatch bound, and a Gaussian minimax floor for the fixed-feature proxy. The closed-form Gaussian confidence bound is sufficient; optimizing the exact cumulant gives a sharper Chernoff bound. These results concern the proxy and its noise model. Optimality of the neural architecture, cross-entropy optimization, and hierarchy selection remains open.

C Experimental Details

C.1 Benchmarks, Splits, and Annotation Budgets

Dataset versions. BANKING77, CLINC150, and HWU64 are the shared benchmarks of recent few-shot intent classifiers (Park et al., 2025; Chen & Wang, 2025). We use the public few-shot intent-detection release used by CPFT and QAID (Zhang et al., 2021; Yehudai et al., 2023), including its BANKING77 validation split. Table 3 gives the source-pool sizes before few-shot subsampling. CLINC150 uses its in-scope labels only. MASSIVE-EN denotes the English `en-US` configuration with its official train/development/test partitions; slot annotations are unused. BANKING77 and MASSIVE also appear in the recent BTZSC evaluation (Aarab, 2026). For HINT3, we use the corrected v2 release, which matches the data distributed with dynamic label refinement (Park et al., 2025), and exclude `NO_NODES_DETECTED` test queries. POWERPLAY11 v2 has 57 training labels, unlike the 59-label v1 description. The inventory includes every training label, including labels absent from the test split. Release identifiers and split hashes identify each data version.

Table 3: **Dataset statistics before few-shot subsampling.** HINT3 test counts exclude out-of-scope queries; its full training pools are used without creating a target development set.

Dataset	Labels	Training pool	Development	Test
BANKING77	77	8,622	1,540	3,080
CLINC150	150	15,000	3,000	4,500
HWU64	64	8,954	1,076	1,076
MASSIVE-EN	60	11,514	2,033	2,974
HINT3: CUREKART	28	599	–	459
HINT3: POWERPLAY11	57	471	–	309
HINT3: SOFMATTRESS	21	328	–	253

Few-shot sampling. On the three primary datasets, each seed defines a random ordering within each training class. Prefixes of that ordering give nested b-shot subsets, with $n = bm$ distinct training annotations. The primary comparison uses $b = 10$ for the methods in Table 1. The label-efficiency sweep uses $b \in \{1, 2, 4, 8, 16\}$ for LoRA-SFT, paired SFT, transported KL, and SIGMA only. The backbone panel uses $b \in \{5, 10\}$ for LoRA-SFT, paired SFT, and SIGMA only. Methods compared at the same budget share the same subsets. MASSIVE-EN uses $b = 10$ with capped budgets $n_z = \min(10, N_z)$, where N_z is the available training count for class z. HINT3 uses its full released support sets, because some classes have fewer than five examples. HINT3 hyperparameters are frozen using the primary datasets' development results before HINT3 evaluation. The annotation budget counts unique examples; HINT3 uses its naturally unbalanced full support sets.

What the budget counts. Guidance edits reuse semantic annotations. Curves indexed by b measure distinct training-label use conditional on a fixed development set; edited views add computation but no independent labels. Total target annotation cost is $n_{\text{train}} + n_{\text{dev}}$, including development labels consulted during prompt selection. Unused target training examples remain inaccessible in the controlled comparison. Table 8 reports processed-token costs; the token-matched SFT control addresses additional training computation.

C.2 Guidance and Held-Out Interventions

Schema construction. We build root–domain–intent trees from published domain/scenario metadata for CLINC150, HWU64, and MASSIVE. For BANKING77 and each HINT3 task, we construct label-only groups using frozen `all-mpnet-base-v2` embeddings (Reimers & Gurevych, 2019), cosine distance, and average-linkage clustering cut into $\lceil \sqrt{m} \rceil$ groups. Sorted label identifiers resolve ties; group descriptions list their member intent names. These label-derived groups define an experimental hierarchy. Their assignments and eligible sibling groups are fixed before model evaluation, using only label metadata. Leaves retain their semantic descriptions when output bindings change.

Outputs and training views. The default output inventory contains the original intent strings, with identity bindings under g_0. The verbalizer control in Table 10 replaces them with fixed opaque identifiers for LoRA-SFT, paired SFT, transported KL, SIGMA, and SIGMA-F only, on the primary backbone at 10 shots on the three primary datasets. Within each run, the complete inventory and its order remain fixed. Training follows Appendix A.1: a uniformly sampled eligible non-root parent and leaf pair on odd optimizer steps, and reversal of all child lists on even steps. Empty edit sets and invalid constructions use the specified identity fallback without resampling. The construction and fallback rule are shared by paired SFT and SIGMA.

Test guidance. Each test instance is evaluated under g_0, the reversal, and ten label-independent binding draws shared across methods. We additionally generate ten sequences of $k \in \{2, 4, 8\}$ sibling swaps per instance, evaluating their final guidance both with and without reversal. Transfer evaluation retains only final documents outside $\mathcal{G}_{\text{train}}$, including the full support of single edits, as defined in Appendix A.2; it rejects returning or cancelling compositions and logs acceptance rates. These documents are excluded from model selection. The additional correctness diagnostic in Table 12 uses these same label-independent draws and separates changed and unchanged targets. Its sample counts and composition coverage are specified in Table 14.

C.3 Metrics and Statistical Reporting

Task performance. Accuracy measures exact correctness of the prescribed output $a_g(z)$, the primary metric in the intent-classification comparisons (Park et al., 2025; Chen & Wang, 2025). Macro-F1 gives equal weight to semantic classes with test support, consistent with the emphasis on class balance in BTZSC (Aarab, 2026). For aggregation across different guidance versions, predictions are expressed as $a_g^{-1}(\widehat{y}_g)$ for evaluation only; this bookkeeping does not alter model predictions. Accuracy is reported under Base/Comp. in Table 1 and Figure 4; Table 6 additionally reports macro-F1. Table 7 restricts the four-view macro-F1 and composed-guidance rankings to its five listed methods on the primary datasets. For a gold-output rank r_i, its ranking metrics are $\text{Hit@}K_{\text{top}} = N^{-1} \sum_i \mathbf{1}[r_i \le K_{\text{top}}]$, $K_{\text{top}} \in \{4, 8\}$, and $\text{MRR} = N^{-1} \sum_i r_i^{-1}$; Hit@1 equals accuracy on identical views and instances. Repeated views are averaged within each original instance first. Ranking diagnostics use the full inventory.

Following guidance changes. For a base/edited pair, let $\widehat{y}_g$ denote the top-ranked output, $y = a_g(z)$, and $y' = \pi_e(y)$. We report

$$\text{PairAcc} = \mathbb{E}[\mathbf{1}[\widehat{y}_g = y \ \wedge \ \widehat{y}_{g'} = y']], \tag{C.1}$$

$$\text{EquivAcc} = \mathbb{E}[\mathbf{1}[\widehat{y}_{g'} = \pi_e(\widehat{y}_g)]], \tag{C.2}$$

$$\text{TV}_{\text{eq}} = \mathbb{E}\big[\tfrac{1}{2} \|\mathbf{q}_{g'} - P_e \mathbf{q}_g\|_1\big]. \tag{C.3}$$

For compositions, π_e and P_e denote the corresponding ordered products. Table 2 reports pooled binding and composition diagnostics for its listed controls. Table 12 separates composition accuracy and PairAcc on $y' \neq y$ and $y' = y$ for paired SFT, transported KL, SIGMA, and SIGMA-F on each primary dataset; Table 14 specifies the corresponding counts and label coverage. Binding diagnostics use the original test population; composition diagnostics use the accepted composition pairs and their aggregation weights. Under reversal, EquivAcc becomes prediction invariance, which is paired with reversal accuracy on the original test population. Standalone reversal metrics use the original test weights. SIGMA-F's Base/Comp. comparison in Table 1 uses matched composition weights for both columns. PairAcc measures correctness, while EquivAcc and TV measure consistency. Distributional diagnostics are restricted to methods exposing comparable complete candidate distributions, without applying test-time score transport to SIGMA's predictions.

Efficiency and uncertainty. We summarize label-efficiency curves by the trapezoidal area under accuracy versus $\log_2 b$ on $b \in \{1, 2, 4, 8, 16\}$, divided by four. This normalized area is reported separately for base and composed guidance on the primary datasets. Seeds $\{13, 21, 42, 87, 100\}$ jointly control subset sampling, initialization, and edits. Table 5 reports mean and standard deviation. The additional controls report five-seed mean/standard deviation and paired 95% intervals for the pre-specified contrasts in Table 13. These intervals use 10,000 bootstrap resamples of seeds and original instances, retaining all views and budget evaluations of an instance together. Each dataset is

analyzed separately before any equal-weight average. MASSIVE complements HWU64; their shared source provenance (FitzGerald et al., 2023) precludes treating them as independent replications.

C.4 Baseline Adaptations and Fair Comparisons

Controlled parameter adaptation. Frozen prompting uses the unadapted backbone with the same guidance and candidate scorer. Single-view SFT receives the same guidance and candidate-ranking cross-entropy as SIGMA, under g_0 only. Paired SFT uses precisely the SIGMA pairs with $\lambda = 0$, sharing initialization, minibatches, optimizer steps, and LoRA configuration. The transported-KL control replaces the score penalty by $\frac{\beta}{2}\{\mathrm{KL}(P_e\mathbf{q}_g\|\mathbf{q}_{g'}) + \mathrm{KL}(\mathbf{q}_{g'}\|P_e\mathbf{q}_g)\}$. This control adapts R-Drop's symmetric distribution matching (Liang et al., 2021) to guidance edits, with dropout disabled. A computation-matched SFT run repeats original views until it reaches SIGMA's measured token-processing budget. Binding-only and representation-only ablations retain the same two-view budget.

ICL and label refinement. Retrieval-ICL (Milios et al., 2023) and dynamic label refinement (Park et al., 2025) retrieve 20 demonstrations using frozen `all-mpnet-base-v2`, ordering them from least to most similar. Only the shared labeled training subset enters the index, and demonstration outputs follow the current binding map. Their controlled variants, denoted ICL-Full and DLR-Full, retain every candidate and the common guidance. DLR can add generated semantic descriptions associated with immutable label identifiers while preserving executable bindings. Its refinement call contributes to its cost.

Prompt optimization. MIPROv2 (Opsahl-Ong et al., 2024) and GEPA (Agrawal et al., 2026) optimize the instruction wrapper and, where applicable, demonstrations while retaining the executable schema. Both access the same labeled training views, the same optimizer backbone, and a cap of 5,000 scored instance-view evaluations, including validation evaluations. Proposals may use training traces and feedback; development access exposes aggregate selection scores only. Table 8 separates auxiliary and processed tokens for the reported primary configuration. MetaTuner-J (Bo et al., 2026) receives the current guidance as a read-only input to its prompt/parameter generation pipeline, and the actor scores the full inventory. Its auxiliary network, warm-start training, and rollouts contribute to system costs. The learned generators are frozen at test time; generated adapters remain part of its inference computation. GEPA and MetaTuner are ICLR 2026 methods adapted here to intent classification.

Semantic classifiers and additional resources. CPFT (Zhang et al., 2021), QAID (Yehudai et al., 2023), and ICDA (Lin et al., 2023) retain their released encoder/generator architectures, with access to the same target-label subsets. Results using their additional unlabeled corpora or generated training examples appear in the additional-resource panel of Table 6. These comparisons include the additional resources of each system. FCSLM (Chen & Wang, 2025) retains its small-model routing, multi-round reasoning, and the named Qwen2.5-72B configuration. The frozen Qwen3-Reranker-8B (Zhang et al., 2025) scores the utterance against every semantic intent description. These semantic predictors break ties in semantic-label coordinates and use an exact lookup of a_g to obtain their output strings. They are identified as classifier-plus-mapping systems. Their exact binding compliance is an architectural property. Our same-backbone variant SIGMA-F in Appendix A.3 provides the corresponding matched-capacity control. All systems are evaluated end to end on the original test population, including routing and retrieval failures. Matched supervision and matched computation are reported separately; auxiliary models, generated data, candidate pruning, and symbolic mapping are identified in the corresponding tables. The resource comparison is restricted to the BANKING77 configuration in Table 8.

C.5 Backbones, Optimization, and Resource Accounting

Checkpoints and scoring. The primary checkpoint is `Qwen/Qwen2.5-7B-Instruct`; the other backbones are `meta-llama/Meta-Llama-3-8B-Instruct` and `Qwen/Qwen3-8B`. These cover the model families used by dynamic label refinement, MetaTuner, and GEPA. Qwen3 uses its non-thinking chat template. Figure 4 compares LoRA-SFT, paired SFT, and SIGMA at 5/10 shots on all three backbones. Only the four methods in Table 5 receive the 1/2/4/8/16-shot sweep on

the primary backbone. The remaining primary baselines are compared at 10 shots with Qwen2.5-7B-Instruct. All controlled methods use an 8,192-token context limit, the complete candidate list, and the mean answer-token log-likelihood including the end-of-answer token specified in Eq. (3). The candidate softmax temperature is one, and dropout is disabled. Label descriptions are shortened before freezing the schema when needed to fit the context limit. Every method retains the complete guidance and candidate inventory.

Optimizer settings. LoRA uses rank 16, scaling 32, and zero dropout on the attention query/key/value/output projections and MLP gate/up/down projections. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, gradient clipping at 1.0, cosine decay, and 3% warmup. The effective batch contains $B = 16$ original instances, with gradient accumulation as necessary. Uniform sampling with replacement runs for $T = 2\lceil 10n/(2B) \rceil$ optimizer steps, giving approximately ten passes and an even step count. The learning-rate grid is $\{10^{-5}, 5 \times 10^{-5}, 10^{-4}\}$, and SIGMA's coefficient grid is $\lambda \in \{0, 0.01, 0.1, 1, 10\}$; β uses the same grid for the transported-KL control. The nominal defaults are 5×10^{-5} and $\lambda = 1$, subject to development selection. Trainable baselines receive at most 15 configuration trials each.

Model selection. Checkpoints are evaluated at five approximately equally spaced even-numbered steps, including the final step. The fixed development pool contains one base view and one edit drawn from $\rho_{\mathcal{T}}$ per instance, identical across compared settings. Selection maximizes mean top-1 accuracy across these views; ties favor the smaller regularization coefficient, earlier checkpoint, and then smaller learning rate. Composed test guidance is excluded from hyperparameter, schema, and checkpoint selection.

Hardware and cost. We use four NVIDIA A100 80GB GPUs, bfloat16, gradient checkpointing, and candidate microbatching. To bound training memory, we compute all candidate scores without retaining model graphs, differentiate the loss with respect to these scores, and then recompute candidate microbatches while accumulating their weighted gradients. With fixed parameters and disabled dropout between passes, this preserves the full-candidate objective; the recomputation cost is included. Table 8 reports trainable parameters, peak memory, GPU-hours, total processed and auxiliary tokens, and batch-one inference latency for its BANKING77, primary-backbone, 10-shot configurations. Latency includes retrieval, refinement, routing, and symbolic lookup when used, with warm-up excluded and median/95th percentile reported on the same hardware. The token-matched comparison in Table 11 uses these measured training-token budgets. SIGMA's $O(m)$ alignment arithmetic is distinguished from the total cost of scoring all candidates under two guidance views.

C.6 Result Aggregation

Base denotes the original guidance. Binding and reversal results are reported separately. Comp. averages the six conditions formed by $k \in \{2, 4, 8\}$ swaps with and without reversal: we first average accepted views within each condition and instance, then aggregate conditions and instances using the metric's evaluation weights. Conditions without an eligible held-out composition are excluded and their coverage is reported. Dataset averages assign equal weight to each dataset after computing its metric. SIGMA-F's Base and Comp. entries in Table 1 use the composition-matched weights of Table 11. Composition PairAcc in Table 2 uses all eligible composition pairs. Binding diagnostics and standalone reversal use the original test population, so their aggregate accuracies need not equal the composition-weighted values. Changed-target accuracy uses the label-independent test draws restricted to $y' \neq y$. The additional composition stratification is specified in Table 12. The paired intervals in Table 13 use the five seeds specified in Appendix C.3.

C.7 Preliminary Study Protocol and Scoring

C.7.1 Diagnostic Protocol and Evaluation Cohorts

Single-view training and diagnostic edits. Single-view SFT uses only original guidance g_0 and its labeled examples, without edited training views, paired supervision, or a consistency loss. At budgets $b \in \mathcal{B} = \{1, 2, 4, 8, 16\}$ labels per intent, the model is frozen before evaluation. A binding revision exchanges two sibling-class output bindings while preserving descriptions and tree structure.

An order revision changes sibling-rule serialization without changing semantic relations or bindings. Both documents are constructed separately from g_0 for evaluation.

Evaluation records and edit coverage. Across five runs, we evaluate 1,500 inputs over 150 candidates at each of the five budgets. The random seeds are 13, 37, 73, 101, and 137. The reference distribution is computed before SFT and shared across budgets and runs. Each stored binding edit exchanges the instance's original gold output with a distinct partner, so its required answer always changes. Its response-error rates therefore describe target-changing edits; Section 4 uses label-independent sampling.

Different cohorts for the two panels. Write $y_{i,0}$ for the gold output under g_0 and $\widehat{c}_{j,b,g}(i)$ for the top-ranked output at run j, budget b, and guidance g. Panel (a) conditions on correctness at budget 8; panel (b) conditions on correctness at every budget:

$$\begin{aligned}\mathcal{A}_j &= \{i : \widehat{c}_{j,8,g_0}(i) = y_{i,0}\},\\ \mathcal{F}_j &= \bigcap_{b\in\mathcal{B}} \{i : \widehat{c}_{j,b,g_0}(i) = y_{i,0}\}.\end{aligned} \tag{C.4}$$

In seed-label order, $|\mathcal{A}_j|$ is $1341, 1277, 1248, 1274, 1240$, and $|\mathcal{F}_j|$ is $1167, 1121, 1059, 1118, 1059$. The latter totals 5,524 instance–run records per budget, including repeated inputs across runs. This fixed cohort controls for changes in original-guidance correctness; its error rates are conditional on cohort membership.

C.7.2 Aligned Candidate-Rank Transitions

Let $q_g(c \mid x)$ denote the normalized candidate score, suppressing run and budget indices. For the instance-specific output transposition π, alignment gives

$$\widetilde{q}_{g_\mathrm{b}}(c \mid x) = q_{g_\mathrm{b}}(\pi(c) \mid x), \qquad \widetilde{q}_{g_\mathrm{o}}(c \mid x) = q_{g_\mathrm{o}}(c \mid x). \tag{C.5}$$

Here g_b and g_o are the binding and order revisions. Candidates are sorted by descending score, with ties resolved by fixed candidate-index order. Aligned ranks $\widetilde{r}_{g_v}$ come from sorting the aligned scores. Each originally top-10 candidate contributes one destination rank per selected instance; destination ranks greater than 10 are grouped into > 10.

For each run, conditional frequencies are row-normalized before averaging:

$$T^{(v,j)}_{r,s} = \Pr_{i\in\mathcal{A}_j}[\widetilde{r}_{g_v}(c) = s \mid r_{g_0}(c) = r], \qquad T^{(v)}_{r,s} = \frac{1}{5}\sum_{j=1}^{5} T^{(v,j)}_{r,s}, \tag{C.6}$$

where $v \in \{\mathrm{b}, \mathrm{o}\}$ and $r = 1, \ldots, 10$. All five cohorts are nonempty. Panel (a) therefore gives equal weight to each run. Each row sums to one, including overflow. The two rank-retention summaries are

$$R^{(v)}_{\mathrm{exact}} = \frac{1}{10}\sum_{r=1}^{10} T^{(v)}_{r,r}, \qquad R^{(v)}_{\pm 1,\le 10} = \frac{1}{10}\sum_{r=1}^{10}\sum_{\substack{1\le s\le 10\\ |s-r|\le 1}} T^{(v)}_{r,s}. \tag{C.7}$$

These summaries average empirical rank-transition frequencies. Binding and order revisions yield exact retention of 27.2566% and 26.8845%, and local retention of 67.7026% and 51.8782%. The overflow bin hides how many original rank-10 candidates move exactly to rank 11. Unrestricted local retention is therefore bounded by $[67.7026, 73.2401]\%$ and $[51.8782, 58.7220]\%$, respectively. Rank changes frequently exceed one position, including potentially near-tied candidates with small score differences. The first diagonal entries, corresponding to Top-1 retention, are 55.9557% and 84.2290%.

C.7.3 Response Errors, Uncertainty Bins, and Interpolation

Reference uncertainty. Reference uncertainty is the normalized entropy of the candidate distribution. Percentiles use all 1,500 distinct instance identifiers before cohort filtering:

$$H_i = -\frac{1}{\log 150}\sum_{c=1}^{150} q_{\mathrm{ref}}(c \mid x_i)\log q_{\mathrm{ref}}(c \mid x_i), \qquad u_i = \frac{100}{1500}\sum_{\ell=1}^{1500}\mathbf{1}[H_\ell \le H_i]. \tag{C.8}$$

Normalizing entropy by $\log 150$ does not change percentile order. Ties receive their upper empirical percentile. Twelve equal-width bins partition $[0, 100]$; internal boundaries enter the bin to their right, and percentile 100 enters the last bin. Their centers range from 4.1667 to 95.8333.

Union error and pooling. For each fixed-cohort record, define

$$E_{j,b}(i) = \mathbf{1}[\widehat{c}_{j,b,g_b}(i) \neq \pi_i(\widehat{c}_{j,b,g_0}(i)) \ \vee \ \widehat{c}_{j,b,g_o}(i) \neq \widehat{c}_{j,b,g_0}(i)]\,. \tag{C.9}$$

Since original-guidance predictions are correct on $\mathcal{F}_j$, this is equivalent to failure of either edited-view correctness check. Panel (b) reports their union, counting each record once even when both checks fail. For percentile bin I_h, counts and error rates pool records across runs:

$$\begin{aligned} N_h &= \sum_{j=1}^{5} \sum_{i \in \mathcal{F}_j} \mathbf{1}[u_i \in I_h], \\ p_{b,h} &= \frac{1}{N_h} \sum_{j=1}^{5} \sum_{i \in \mathcal{F}_j} \mathbf{1}[u_i \in I_h] E_{j,b}(i). \end{aligned} \tag{C.10}$$

Bin counts are $511, 499, 486, 458, 482, 445, 481, 449, 418, 446, 422, 427$ at every budget. Low and high uncertainty summarize $(p_{b,1} + p_{b,2})/2$ and $(p_{b,11} + p_{b,12})/2$. These are unweighted averages of bin rates, each covering one sixth of the percentile range. At budget 8 these two-bin averages equal 35.9657% and 68.9984%.

Rendering. Colors and contours bilinearly interpolate the five-by-twelve bin-rate grid in percentile and $\log_2 b$, on a 240×220 grid between the extreme bin centers. Values between evaluated budgets are interpolated for display. Markers locate original grid entries, with size based on bin count. Colors share a $[0, 1]$ scale; contour labels are rounded interpolated values.

Results and interpretation appear in Appendix D.5.

D Additional Experimental Results

D.1 Average Accuracy and Composition Gap

Table 4 takes the unweighted mean of the three dataset accuracies in Table 1. The drop is average Base minus average Comp., in percentage points; means and differences are computed before final rounding. Paired SFT and transported KL reduce the LoRA-SFT drop from 6.43 to 3.28 and 2.59, respectively; SIGMA reduces it to 1.91 while also improving Base accuracy.

Exact equivariance of SIGMA-F. SIGMA-F predicts a semantic label under a canonical representation and maps that prediction to the output space using the current binding function. Therefore, for any valid guidance composition, its semantic prediction is unchanged and its candidate distribution is transported exactly by the corresponding permutation. When Base and composition accuracy are evaluated on identical instances with identical aggregation weights, their correctness indicators coincide instance-wise. Consequently, SIGMA-F obtains a zero paired Base-to-Composition drop on all three primary datasets.

Table 4: **Average accuracy and the composition gap.** Unweighted averages across the three datasets in Table 1. Accuracy is in percent; the Base-to-Comp. drop is in percentage points. Best and second-best values are marked across all rows. SIGMA-F's zero drop follows from exact binding lookup on identical instances with identical aggregation weights (Appendix D.1).

Method	Avg. Base	Avg. Comp.	Base-to-Comp. drop
LoRA-SFT	90.80	84.37	6.43
Paired SFT	91.58	88.30	3.28
Transported KL	91.90	89.31	2.59
Our method and variants			
SIGMA	**92.91**	<u>91.00</u>	<u>1.91</u>
SIGMA-F	<u>92.22</u>	**92.22**	**0.00**

Separating task accuracy from composition sensitivity. SIGMA improves average Base accuracy over LoRA-SFT by 2.11 percentage points and average Comp. accuracy by 6.63 points. The smaller composition gap therefore accompanies higher accuracy in both views. Relative to paired SFT, its gains are 1.33 points on Base and 2.70 on Comp.; relative to transported KL, they are 1.01 and 1.69 points. Paired supervision already recovers much of the performance lost under composition, while SIGMA provides a further improvement over both paired baselines. These comparisons locate the larger empirical gain in the held-out compositions, although they do not isolate a unique internal mechanism.

Interpreting the factorized reference. SIGMA has the highest average Base accuracy in this table, whereas SIGMA-F has the highest Comp. accuracy. SIGMA exceeds SIGMA-F by 0.69 points on Base but trails it by 1.22 points on Comp. Exact binding lookup can therefore preserve a weaker semantic classifier's predictions well enough to reverse their ordering under composition. Its zero gap expresses a structural guarantee under matched evaluation support; it does not imply perfect semantic classification. Conversely, SIGMA's remaining 1.91-point gap shows that learned alignment does not attain exact transport in this evaluation. Reporting both absolute accuracies and their difference separates semantic performance from sensitivity to changes in guidance.

Scope of the summary. Equal dataset weights prevent the largest test set from dominating the average, but can conceal differences across datasets. The aggregate ordering should therefore be read alongside the dataset-level results in Table 1. These differences summarize the reported means; paired uncertainty estimates for the specified contrasts are given in Table 13.

D.2 Label Budgets

Table 5 reports the complete budget sweep under the same nested labeled subsets and fixed development protocol. It complements Figure 5 by showing Base accuracy, composition accuracy, and variation across seeds at every tested budget.

Table 5: **Label efficiency across nested annotation budgets.** Qwen2.5-7B-Instruct; mean ± standard deviation over five seeds. Accuracy and normalized log-budget AULC use the percentage scale. Development annotations are counted separately. Best and second-best means are marked within each dataset and view.

Method	View	$b = 1$	2	4	8	16	AULC
BANKING77							
LoRA-SFT	Base	72.84 ± 2.27	79.65 ± 1.59	84.72 ± 1.21	87.63 ± 1.01	89.46 ± 0.67	83.29 ± 0.64
	Comp.	59.71 ± 2.63	66.82 ± 1.83	73.54 ± 1.10	79.83 ± 1.28	82.61 ± 0.85	72.84 ± 0.57
Paired SFT	Base	75.91 ± 1.66	82.34 ± 1.40	86.71 ± 1.47	88.66 ± 0.78	90.11 ± 0.84	85.18 ± 0.90
	Comp.	67.83 ± 1.94	75.46 ± 1.48	81.94 ± 1.17	84.79 ± 1.06	86.48 ± 1.05	79.84 ± 0.53
Transported KL	Base	76.88 ± 2.12	83.12 ± 1.39	87.29 ± 1.19	89.11 ± 0.91	90.38 ± 0.66	85.79 ± 0.73
	Comp.	69.52 ± 2.19	77.04 ± 1.94	83.06 ± 1.30	86.15 ± 0.97	87.74 ± 0.85	81.22 ± 0.73
SIGMA	Base	**79.64** ± 1.81	**85.07** ± 1.39	**88.52** ± 1.19	**90.18** ± 0.80	**91.27** ± 0.55	**87.31** ± 0.60
	Comp.	**73.81** ± 2.13	**81.26** ± 1.75	**85.72** ± 1.13	**88.01** ± 0.96	**89.47** ± 0.76	**84.16** ± 0.55
CLINC150							
LoRA-SFT	Base	86.42 ± 2.02	90.34 ± 1.79	93.18 ± 1.17	94.83 ± 0.98	96.02 ± 0.53	92.39 ± 0.68
	Comp.	77.93 ± 2.68	83.86 ± 2.03	87.92 ± 1.60	90.12 ± 1.28	91.54 ± 0.85	86.66 ± 0.79
Paired SFT	Base	87.91 ± 1.96	91.82 ± 1.44	94.21 ± 1.26	95.46 ± 0.92	96.31 ± 0.68	93.40 ± 0.51
	Comp.	81.76 ± 2.22	87.12 ± 1.73	90.66 ± 1.52	93.06 ± 1.02	94.11 ± 0.74	89.69 ± 0.60
Transported KL	Base	88.54 ± 1.67	92.33 ± 1.33	94.62 ± 1.17	95.73 ± 0.81	96.42 ± 0.66	93.79 ± 0.55
	Comp.	82.94 ± 2.18	88.18 ± 1.75	91.47 ± 1.40	93.81 ± 1.00	94.72 ± 0.72	90.57 ± 0.51
SIGMA	Base	**89.76** ± 1.59	**93.51** ± 1.25	**95.43** ± 1.11	**96.31** ± 0.78	**97.01** ± 0.55	**94.66** ± 0.52
	Comp.	**85.47** ± 1.96	**90.31** ± 1.75	**93.42** ± 1.12	**95.01** ± 0.98	**95.81** ± 0.74	**92.34** ± 0.56
HWU64							
LoRA-SFT	Base	70.93 ± 2.22	78.84 ± 1.75	84.17 ± 1.46	87.96 ± 0.91	89.77 ± 0.77	82.83 ± 0.70
	Comp.	57.88 ± 2.66	65.93 ± 1.84	73.19 ± 1.54	80.37 ± 1.26	83.14 ± 0.94	72.50 ± 0.76
Paired SFT	Base	74.12 ± 2.10	81.36 ± 1.49	86.29 ± 1.19	89.09 ± 0.80	90.42 ± 0.61	84.75 ± 0.54
	Comp.	66.47 ± 2.02	74.98 ± 1.85	81.35 ± 1.59	85.63 ± 1.11	87.02 ± 0.81	79.68 ± 0.58
Transported KL	Base	75.26 ± 1.80	82.08 ± 1.53	86.94 ± 1.29	89.52 ± 0.88	90.69 ± 0.60	85.38 ± 0.51
	Comp.	68.13 ± 2.27	76.22 ± 1.73	82.44 ± 1.33	86.51 ± 1.16	87.89 ± 0.80	80.80 ± 0.58
SIGMA	Base	**78.31** ± 1.65	**84.07** ± 1.32	**88.22** ± 1.06	**90.66** ± 0.80	**91.81** ± 0.55	**87.00** ± 0.44
	Comp.	**72.65** ± 1.97	**80.41** ± 1.60	**85.28** ± 1.30	**88.37** ± 1.03	**89.91** ± 0.79	**83.83** ± 0.67

Gains across the budget range. SIGMA has the highest reported mean for both views at every tested budget on all three datasets. Its composition advantage over LoRA-SFT is largest at one label per intent: 14.10, 7.54, and 14.77 percentage points on BANKING77, CLINC150, and HWU64, respectively. At 16 labels per intent, the corresponding gains remain 6.86, 4.27, and 6.77 points. The advantage thus persists as supervision increases, while its magnitude narrows between the lowest and highest budgets. Gains over transported KL also remain positive throughout the sweep, showing that the ordering is not confined to the comparison with single-view training.

Matching accuracy with fewer training labels. At four labels per intent, SIGMA reaches Comp. accuracies of 85.72%, 93.42%, and 85.28%, exceeding 16-shot LoRA-SFT by 3.11, 1.88, and 2.14 points. Two-shot SIGMA remains below the same reference on every dataset, so four shots is the smallest tested budget that exceeds it. This comparison gives a 75% reduction in original training annotations at these measured operating points. The development set is unchanged, and the comparison does not imply the same reduction in total annotation cost or training computation. Intermediate budgets were not evaluated.

Integrated performance and uncertainty. Composition AULC is 84.16, 92.34, and 83.83 for SIGMA, compared with 81.22, 90.57, and 80.80 for transported KL. The integrated gains of 2.94, 1.77, and 3.03 points support the same ordering across the sampled log-budget range. AULC summarizes the observed curve rather than an unmeasured sample-complexity threshold. The displayed standard deviations describe variation over five seeds; the table alone does not establish statistical significance

for every cellwise comparison. These experiments vary distinct training utterances, whereas the repeated-response budget analysis in Appendix B.7 holds the design points fixed.

D.3 Additional Benchmarks and Metrics

Table 6 extends evaluation to MASSIVE-EN and three HINT3 tasks, while Table 7 examines classification and candidate ranking on the primary datasets. These two views test how broadly the observed gains extend across benchmarks and evaluation measures.

Table 6: **Generalization to additional intent benchmarks.** Each cell reports accuracy / macro-F1 (%). MASSIVE-EN uses 10 shots and all 60 intents; CURE, POWER, and SOFA are HINT3 v2 tasks with full support sets (28/57/21 intents). The common backbone is Qwen2.5-7B-Instruct. Additional-resource baselines retain their named architectures; the frozen reranker uses no labeled examples. The final block contains SIGMA and its factorized variant, SIGMA-F.

Method	View	MASSIVE	CURE	POWER	SOFA
LoRA-SFT	Base	89.14 / 88.62	81.73 / 80.96	78.42 / 77.61	84.37 / 83.88
	Comp.	83.06 / 82.39	72.84 / 71.92	68.71 / 67.53	75.26 / 74.61
Paired SFT	Base	90.03 / 89.51	82.46 / 81.74	79.37 / 78.66	85.29 / 84.81
	Comp.	87.12 / 86.54	77.39 / 76.52	74.86 / 73.91	80.63 / 79.92
ICL-Full	Base	87.56 / 86.91	79.18 / 78.22	76.34 / 75.41	82.11 / 81.47
	Comp.	80.47 / 79.68	69.26 / 68.17	65.92 / 64.63	71.83 / 70.94
DLR-Full	Base	89.61 / 89.08	83.17 / 82.34	80.09 / 79.21	85.91 / 85.29
	Comp.	84.72 / 84.05	74.91 / 73.88	71.66 / 70.42	77.84 / 76.96
GEPA	Base	90.14 / 89.67	83.52 / 82.86	80.74 / 79.93	86.18 / 85.74
	Comp.	86.49 / 85.88	77.18 / 76.31	73.95 / 72.83	79.86 / 79.14
MetaTuner-J	Base	90.67 / 90.16	84.41 / 83.72	81.83 / 80.96	87.29 / 86.77
	Comp.	87.33 / 86.72	78.82 / 77.95	75.91 / 74.88	81.37 / 80.61
Additional-resource baselines with exact mapping					
CPFT (Zhang et al., 2021)	Base	88.72 / 88.21	80.64 / 79.71	77.18 / 76.24	83.06 / 82.41
	Comp.	88.69 / 88.17	80.61 / 79.67	77.14 / 76.20	83.02 / 82.37
QAID-large (Yehudai et al., 2023)	Base	90.06 / 89.55	83.91 / 83.12	80.73 / 79.86	86.44 / 85.91
	Comp.	90.02 / 89.51	83.87 / 83.08	80.70 / 79.82	86.39 / 85.86
ICDA-XL (Lin et al., 2023)	Base	89.83 / 89.31	82.74 / 81.88	79.68 / 78.79	85.52 / 84.93
	Comp.	89.79 / 89.27	82.70 / 81.84	79.65 / 78.76	85.47 / 84.88
FCSLM (72B) (Chen & Wang, 2025)	Base	91.02 / 90.61	85.13 / 84.45	82.11 / 81.27	87.83 / 87.26
	Comp.	90.98 / 90.57	85.09 / 84.41	82.07 / 81.23	87.79 / 87.22
Qwen3-Reranker-8B (Zhang et al., 2025)	Base	88.96 / 88.44	82.37 / 81.55	79.49 / 78.58	84.76 / 84.12
	Comp.	88.92 / 88.40	82.33 / 81.51	79.45 / 78.54	84.72 / 84.08
Our method and variants					
SIGMA-F	Base	90.31 / 89.86	84.96 / 84.27	82.14 / 81.36	87.67 / 87.09
	Comp.	90.27 / 89.81	84.91 / 84.21	82.18 / 81.39	87.62 / 87.03
SIGMA	Base	91.24 / 90.79	85.36 / 84.69	82.57 / 81.84	88.14 / 87.62
	Comp.	89.16 / 88.71	82.08 / 81.31	79.54 / 78.62	84.63 / 83.91

Transfer to additional benchmarks. Compared with paired SFT, SIGMA improves Base accuracy by 1.21, 2.90, 3.20, and 2.85 percentage points on MASSIVE, CURE, POWER, and SOFA, respectively. Its composition gains are larger: 2.04, 4.69, 4.68, and 4.00 points. Composition macro-F1 improves by 2.17, 4.79, 4.71, and 3.99 points, so the accuracy trend is accompanied by gains under equal class weighting. SIGMA nevertheless retains Base-to-Comp. accuracy gaps of 2.08–3.51 points across these benchmarks; improved transfer does not eliminate sensitivity to composed guidance.

Comparison scope. SIGMA exceeds the other methods in the common-backbone, text-guidance block on both accuracy and macro-F1 in every displayed view. The ranking changes when exact-mapping systems are included: SIGMA-F, for example, achieves higher composition accuracy than

SIGMA on all four benchmarks, by 1.11–2.99 points. Several additional-resource baselines also exceed SIGMA on composition accuracy. Their separate block retains the named architectures, auxiliary resources, and mapping interfaces, so these results do not constitute an isolated comparison of training objectives. MASSIVE uses 10-shot training, whereas the HINT3 tasks use full support sets; their absolute scores should be interpreted within each benchmark.

Table 7: **Complementary classification and ranking metrics.** Qwen2.5-7B-Instruct, 10 shots; the same five seeds as Table 1. Bind.: label-independent binding swaps; Rev.: reversal. Hit@4/8 are percentages; MRR is on $[0, 1]$. Ranking uses composed guidance and the full candidate inventory. Best and second-best values are marked within each dataset.

	Macro-F1 (%)				Composed-guidance ranking		
Method	**Base**	**Bind.**	**Rev.**	**Comp.**	**Hit@4**	**Hit@8**	**MRR**
BANKING77							
LoRA-SFT	88.06	82.17	87.42	80.51	92.84	96.31	0.855
Paired SFT	88.84	86.07	88.92	84.96	95.28	97.64	0.892
DLR-Full	87.71	83.19	87.63	81.89	93.51	96.72	0.867
GEPA	88.47	84.89	88.54	83.73	94.72	97.18	0.883
SIGMA	**90.37**	**89.18**	**90.44**	**88.17**	**97.06**	**98.54**	**0.922**
CLINC150							
LoRA-SFT	95.11	92.08	94.93	90.36	97.72	98.86	0.932
Paired SFT	95.69	94.21	95.62	93.17	98.46	99.17	0.956
DLR-Full	95.47	92.91	95.31	91.87	98.11	99.02	0.945
GEPA	95.76	94.02	95.73	92.84	98.39	99.12	0.953
SIGMA	**96.55**	**95.79**	**96.49**	**95.12**	**99.08**	**99.54**	**0.972**
HWU64							
LoRA-SFT	88.31	82.56	87.92	81.07	93.47	96.62	0.861
Paired SFT	89.29	86.79	89.31	85.67	95.91	97.91	0.901
DLR-Full	88.54	83.77	88.62	82.28	94.23	97.04	0.876
GEPA	89.08	85.43	89.16	84.31	95.17	97.48	0.892
SIGMA	**90.91**	**89.64**	**90.96**	**88.61**	**97.46**	**98.78**	**0.929**

Classification under different guidance edits. Among the methods in Table 7, SIGMA has the highest macro-F1 for Base, binding swaps, reversal, and compositions on all three datasets. Relative to paired SFT, its composition macro-F1 gains are 3.21, 1.95, and 2.94 points on BANKING77, CLINC150, and HWU64; its Base gains are smaller at 1.53, 0.86, and 1.62 points. Reversal macro-F1 stays close to Base for SIGMA, with absolute differences of at most 0.07 points. Aggregate similarity, however, does not establish instancewise invariance: different predictions can yield similar macro-F1. The direct consistency diagnostics in Table 2 address that separate question.

Candidate ranking under compositions. SIGMA also leads the displayed methods in Hit@4, Hit@8, and MRR. Over paired SFT, its Hit@4 gains are 1.78, 0.62, and 1.55 percentage points, and its MRR gains are 0.030, 0.016, and 0.028 on the three datasets. The corresponding Hit@8 gains are smaller at 0.90, 0.37, and 0.87 points, where scores are already close to the metric's ceiling. Together, these measures indicate that the improvements extend to the placement of the correct candidate within the ranked list. They measure ranking quality under composed guidance, rather than whether the entire score vector is transported equivariantly; that property requires the aligned-score diagnostics and their stated scope.

D.4 Computational Cost

Table 8 reports training and inference costs on BANKING77 under the hardware and accounting protocol in Appendix C.5. Parameter counts, processed tokens, and measured GPU-hours capture different sources of overhead.

Table 8: **Computational cost and inference latency.** BANKING77, Qwen2.5-7B-Instruct, 10 shots; hardware as in Appendix C.5. GPU-hours include auxiliary training and optimization. Token counts include candidate rescoring and prompt search. Batch-one latency includes retrieval, refinement, routing, and mapping. SIGMA is shaded.

	Trainable	Compute	Peak	Processed	Auxiliary	Latency (ms)	
Method	**params (M)**	**GPU-h**	**GiB**	**tokens (M)**	**tokens (M)**	**p50**	**p95**
LoRA-SFT	40.42	7.84	58.60	107.84	0.00	684	793
SFT (token-matched)	40.42	13.51	58.80	199.42	0.00	687	799
Paired SFT	40.42	13.18	59.40	196.87	0.00	690	804
Transported KL	40.42	13.46	59.70	197.11	0.00	691	806
ICL-Full	0.00	0.61	33.80	41.28	0.00	1018	1267
DLR-Full	0.00	1.93	38.40	46.17	18.62	1186	1479
MIPROv2	0.00	17.62	70.40	271.86	93.41	721	853
GEPA	0.00	21.47	71.00	314.58	127.63	735	871
MetaTuner-J	112.64	26.84	74.30	232.47	156.28	789	1016
Our method and variants							
SIGMA-F	40.42	8.11	57.90	123.08	0.00	675	786
SIGMA	40.42	13.73	59.80	197.33	0.00	688	797

The marginal cost of alignment. SIGMA, LoRA-SFT, paired SFT, and transported KL each optimize 40.42 million parameters. SIGMA takes 13.73 GPU-hours, compared with 13.18 for paired SFT and 13.46 for transported KL: increases of 0.55 GPU-hours (4.2%) and 0.27 GPU-hours (2.0%), respectively. Their processed-token counts are also close, at 197.33, 196.87, and 197.11 million, and peak memory ranges from 59.40 to 59.80 GiB. Under this paired-training protocol, centered-score alignment therefore adds a small measured increment to the cost of scoring two guidance views, without adding trainable parameters or auxiliary training tokens.

Paired supervision and the training budget. Against ordinary LoRA-SFT, SIGMA requires 5.89 additional GPU-hours, about 75.1% more, and processes 197.33 rather than 107.84 million tokens. This larger comparison includes the cost of paired training as well as the alignment objective. Token-matched SFT takes 13.51 GPU-hours and processes 199.42 million tokens, providing a closer computational reference for the accuracy comparisons in Table 11. Fewer original training labels do not necessarily imply less computation, because each label can support multiple scored guidance views.

Inference and alternative pipelines. SIGMA's batch-one p50/p95 latencies are 688/797 ms, close to LoRA-SFT's 684/793 ms in this measurement. The alignment term is used during training; inference scores candidates under one current guidance view. SIGMA-F is cheaper to train at 8.11 GPU-hours and has latencies of 675/786 ms, while requiring the explicit semantic-classification and binding-lookup interface described in Appendix A.3. Zero trainable parameters also need not mean zero optimization cost: MIPROv2 and GEPA consume 17.62 and 21.47 GPU-hours, respectively, with prompt search included. Conversely, ICL-Full and DLR-Full require less recorded preparation compute but have higher measured inference latencies. These trade-offs depend on the workload, hardware, and accounted pipeline stages; the table does not establish a universal speed ordering across deployment settings.

D.5 Preliminary Study Results

Table 9 reports response errors for the fixed cohort defined in Appendix C.7.

The overall rate decreases and a broad uncertainty gradient persists, but neither trend is pointwise monotone. For example, the bin centered at percentile 87.5 rises from 64.93% at budget 8 to 66.35% at budget 16. These conditional frequencies describe an association between uncertainty and response error, without confidence intervals or significance tests. The ranking changes and uncertainty gradient leave causal mechanisms unresolved.

Table 9: **Response error in the fixed cohort (%).** Overall error weights bins by count; low and high uncertainty each average two bin rates.

Labels per intent	Overall	Low uncertainty	High uncertainty
1	57.91	41.30	78.56
2	55.18	38.93	75.97
4	53.39	38.83	74.44
8	51.34	35.97	69.00
16	48.41	34.88	66.78

D.6 Controls for Guidance Use and Label Efficiency

Tables 10–14 report additional evidence for guidance use and training-label efficiency. All controls use the three primary datasets, Qwen2.5-7B-Instruct, and the five seeds in Appendix C.3. Training uses 10 shots per intent unless a label-budget contrast is explicitly indicated.

Removing output-name cues. For Table 10, output strings are replaced by unique opaque identifiers, assigned by a fixed label-independent permutation shared across methods. Semantic descriptions, tree structure, labeled subsets, and edit definitions are unchanged. The identifier inventory and its order remain fixed within each run. LoRA-SFT, paired SFT, transported KL, SIGMA, and SIGMA-F are retrained with this inventory using the same development-selection rule. This control tests the effect of removing semantic cues from output strings while retaining semantic descriptions in the guidance.

Matched computation and correctness. Table 11 reports task accuracy for the original-string models, including token-matched SFT. Within each composition condition, Base is evaluated on exactly the same accepted instances, with the same weights, as Comp. The reported drop is their paired difference. Matching support isolates accuracy changes under composed guidance. SIGMA-F's zero drops follow from exact mapping (Appendix A.3) and match Table 1. Table 12 separates changed and unchanged composition targets and pairs reversal invariance with reversal correctness. Both strata use the same label-independent test draws.

Uncertainty and evaluation support. Table 13 reports the pre-specified paired comparisons for composition accuracy, the four-versus-sixteen-shot comparison, and composition AULC. The 10-shot contrasts use the shared support in Table 11; cross-budget contrasts also share the evaluation support and preserve each instance's budget trajectory. Each contrast is computed per seed before aggregation. Bootstrap resampling preserves method pairing, all views, and nested budgets. Table 14 records the accepted compositions and the sample sizes behind the target-change diagnostics. Accuracy is reported only for strata with observations.

Table 10: **Opaque-output control.** Qwen2.5-7B-Instruct, 10 shots; accuracy in percent, five-seed mean ± SD. Only output strings change; semantic descriptions remain. All five methods are retrained with the same opaque inventory. Base and Comp. use matched instances and weights within each composition condition. **Bold**/underline: highest/second-highest mean in each column.

	BANKING77		CLINC150		HWU64	
Method	**Base**	**Comp.**	**Base**	**Comp.**	**Base**	**Comp.**
LoRA-SFT	86.08 ± 1.02	76.67 ± 1.29	93.74 ± 0.84	87.80 ± 1.17	86.24 ± 0.99	77.35 ± 1.24
Paired SFT	87.72 ± 0.93	82.79 ± 1.16	94.86 ± 0.78	91.63 ± 1.00	88.05 ± 0.90	83.62 ± 1.10
Transported KL	88.39 ± 0.88	84.54 ± 1.05	95.22 ± 0.74	92.57 ± 0.91	88.63 ± 0.86	84.67 ± 1.01
SIGMA-F	89.91 ± 0.72	**89.91** ± 0.72	**96.25** ± 0.56	**96.25** ± 0.56	**90.49** ± 0.68	**90.49** ± 0.68
SIGMA	**89.92** ± 0.82	87.21 ± 0.95	96.05 ± 0.70	94.49 ± 0.84	90.39 ± 0.79	87.69 ± 0.92

Performance without output-name cues. SIGMA exceeds paired SFT in composition accuracy by 4.42, 2.86, and 4.07 percentage points on BANKING77, CLINC150, and HWU64, respectively; its

gains over transported KL are 2.67, 1.92, and 3.02 points. The advantage therefore persists when the output strings no longer provide recognizable class names. Semantic descriptions remain available, so this control supports robustness to output renaming rather than learning without semantic information. SIGMA-F achieves the highest composition accuracy on all three datasets, with equal Base and Comp. accuracy under matched support. Its explicit binding interface remains a distinct structural advantage even when output identifiers are opaque. The comparison concerns models retrained with the opaque inventory; it does not measure a frozen model's adaptation to previously unseen output tokens.

Table 11: **Accuracy with matched evaluation support.** Original output strings; primary backbone, 10 shots. Accuracy is in percent; drop is in percentage points. Each entry is the five-seed mean ± SD. For every composition condition, Base reuses the accepted Comp. instances and aggregation weights. The Base-minus-Comp. drop is computed within each seed before its SD. Token-matched SFT uses the measured SIGMA training-token budget. **Bold**/underline: best/second-best mean within each dataset; higher accuracy and lower drop are preferred.

Method	Matched Base	Comp.	Paired drop
BANKING77			
LoRA-SFT	88.39 ± 0.86	80.96 ± 1.12	7.43 ± 0.71
SFT (token-matched)	88.72 ± 0.79	80.43 ± 1.17	8.29 ± 0.76
Paired SFT	89.16 ± 0.79	85.37 ± 1.05	3.79 ± 0.61
Transported KL	89.60 ± 0.81	86.83 ± 0.91	2.77 ± 0.52
SIGMA-F	<u>89.91</u> ± 0.72	**89.91** ± 0.72	**0.00** ± 0.00
SIGMA	**90.70** ± 0.73	<u>88.56</u> ± 0.87	<u>2.14</u> ± 0.48
CLINC150			
LoRA-SFT	95.26 ± 0.74	90.61 ± 1.08	4.65 ± 0.49
SFT (token-matched)	95.39 ± 0.68	90.22 ± 1.11	5.17 ± 0.55
Paired SFT	95.83 ± 0.72	93.42 ± 0.91	2.41 ± 0.38
Transported KL	96.02 ± 0.70	94.17 ± 0.86	1.85 ± 0.31
SIGMA-F	<u>96.25</u> ± 0.56	**96.25** ± 0.56	**0.00** ± 0.00
SIGMA	**96.67** ± 0.67	<u>95.37</u> ± 0.82	<u>1.30</u> ± 0.27
HWU64			
LoRA-SFT	88.71 ± 0.82	81.54 ± 1.10	7.17 ± 0.68
SFT (token-matched)	88.96 ± 0.77	80.88 ± 1.15	8.08 ± 0.74
Paired SFT	89.69 ± 0.75	86.11 ± 0.98	3.58 ± 0.57
Transported KL	90.04 ± 0.73	86.94 ± 0.90	3.10 ± 0.51
SIGMA-F	<u>90.49</u> ± 0.68	**90.49** ± 0.68	**0.00** ± 0.00
SIGMA	**91.32** ± 0.69	<u>89.08</u> ± 0.86	<u>2.24</u> ± 0.46

Additional computation does not reproduce the composition gain. Token-matched SFT slightly improves Base accuracy over ordinary LoRA-SFT on each dataset, but its composition accuracy is lower by 0.53, 0.39, and 0.66 points, respectively. Under the measured token budget, extending single-view supervision therefore does not recover the gains from paired training and alignment. SIGMA exceeds this control by 8.13, 5.15, and 8.20 points on compositions. This result addresses the extra-token explanation within the specified optimization protocol; it does not exclude other single-view training schedules or objectives.

Separating matched support from exact transport. With identical evaluation instances and weights, SIGMA's paired drops are 2.14, 1.30, and 2.24 points, compared with 3.79, 2.41, and 3.58 for paired SFT. Thus the reduced gap persists after controlling the population on which Base and composition accuracy are measured. SIGMA-F's zero drops are consistent with its exact lookup construction, while its composition accuracies exceed SIGMA's by 1.35, 0.88, and 1.41 points. The table distinguishes two empirical objectives: improving the semantic predictions made under Base guidance and preserving those predictions correctly as bindings change. SIGMA improves both relative to the learned paired baselines, but retains a measurable gap to exact transport.

Required answer changes drive the larger gains. SIGMA's changed-target accuracy exceeds paired SFT by 13.48, 9.57, and 14.24 points across the three datasets. Its unchanged-target gains are smaller, at 1.90, 1.47, and 1.25 points. This stratification shows that the aggregate improvement includes

Table 12: **Correct responses to guidance edits.** Original output strings; primary backbone, 10 shots; five-seed mean ± SD in percent. Changed/unchanged strata use the accepted label-independent compositions, conditioned on $y' \neq y$ / $y' = y$. PairAcc requires both Base and edited answers to be correct. Reversal is evaluated on the original test population; Inv. is prediction invariance under reversal. Composition counts are given in Table 14. **Bold**/underline: highest/second-highest mean within each dataset.

	Changed target		Unchanged target		Reversal	
Method	**Acc.**	**PairAcc**	**Acc.**	**PairAcc**	**Acc.**	**Inv.**
BANKING77						
Paired SFT	69.70 ± 1.42	67.80 ± 1.51	87.34 ± 0.92	82.39 ± 1.06	89.26 ± 0.81	95.12 ± 0.74
Transported KL	74.38 ± 1.31	72.60 ± 1.38	88.39 ± 0.88	84.59 ± 0.97	89.67 ± 0.79	96.42 ± 0.61
SIGMA-F	**89.54** ± 0.77	**89.54** ± 0.77	**89.96** ± 0.71	**89.96** ± 0.71	89.87 ± 0.70	**100.00** ± 0.00
SIGMA	83.18 ± 1.02	81.95 ± 1.08	89.24 ± 0.80	87.50 ± 0.88	**90.78** ± 0.73	98.61 ± 0.34
CLINC150						
Paired SFT	80.85 ± 1.12	78.70 ± 1.18	94.22 ± 0.78	89.47 ± 0.91	95.77 ± 0.71	97.36 ± 0.52
Transported KL	84.27 ± 0.98	82.65 ± 1.03	94.80 ± 0.73	91.13 ± 0.81	96.07 ± 0.68	98.08 ± 0.43
SIGMA-F	**96.13** ± 0.57	**96.13** ± 0.57	**96.26** ± 0.55	**96.26** ± 0.55	96.22 ± 0.56	**100.00** ± 0.00
SIGMA	90.42 ± 0.82	89.35 ± 0.87	95.69 ± 0.66	94.15 ± 0.73	**96.64** ± 0.64	99.16 ± 0.23
HWU64						
Paired SFT	68.99 ± 1.55	66.60 ± 1.64	88.72 ± 0.95	83.87 ± 1.08	89.73 ± 0.79	94.84 ± 0.82
Transported KL	73.58 ± 1.42	71.80 ± 1.50	88.98 ± 0.89	85.08 ± 0.99	90.09 ± 0.76	96.11 ± 0.65
SIGMA-F	**90.60** ± 0.72	**90.60** ± 0.72	**90.47** ± 0.67	**90.47** ± 0.67	90.53 ± 0.68	**100.00** ± 0.00
SIGMA	83.23 ± 1.11	81.70 ± 1.18	89.97 ± 0.82	87.44 ± 0.91	**91.39** ± 0.70	98.47 ± 0.37

inputs whose required answer actually changes, so retaining the original prediction cannot explain the full result. Changed-target PairAcc rises by 14.15, 10.65, and 15.10 points, indicating more cases in which both the original answer and its edited counterpart are correct. These are conditional rates; their contributions to an overall score depend on the stratum sizes in Table 14.

Consistency and correctness remain distinct. SIGMA's reversal invariance ranges from 98.47% to 99.16%, while reversal accuracy ranges from 90.78% to 96.64%. A model can preserve an incorrect prediction, so high invariance alone is insufficient evidence of successful task performance. SIGMA-F makes this distinction explicit: it has 100% reversal invariance but imperfect accuracy. Conversely, PairAcc imposes correctness in both views and therefore cannot exceed the corresponding edited-view accuracy. Reading these columns jointly avoids treating stable outputs and correct use of a binding change as interchangeable outcomes.

Paired evidence for the main comparisons. All 24 displayed pointwise intervals lie above zero. In particular, the 10-shot composition gains over transported KL are 1.73, 1.20, and 2.14 points, with lower interval endpoints of 1.19, 0.79, and 1.57 points. The four-shot SIGMA versus 16-shot LoRA-SFT comparison also has positive intervals on all three datasets, complementing the mean-curve comparison in Figure 5. Positive AULC contrasts support an advantage integrated across the tested budgets, rather than only at one selected operating point.

The pairing preserves shared evaluation instances and training seeds, so uncertainty concerns differences between methods under this design. These are pointwise 95% intervals, not a simultaneous confidence statement for the entire table. They also do not incorporate uncertainty from alternative task hierarchies, prompt specifications, or benchmark selection. Their evidential scope is the reported contrasts under the fixed experimental protocol.

Table 13: **Paired differences for the central comparisons.** Each cell is SIGMA minus the comparator in percentage points, with its paired pointwise 95% bootstrap interval. Shot pairs are SIGMA versus comparator. AULC uses 1/2/4/8/16-shot composition accuracy on the normalized log-budget scale. Intervals use 10,000 resamples, preserving seed, instance, method, view, and budget pairing. Positive differences favor SIGMA; bold differences have intervals entirely above zero.

Comparator	Quantity / shots	BANKING77	CLINC150	HWU64
LoRA-SFT	Comp., 10 vs. 10	**+7.60** [6.91, 8.33]	**+4.76** [4.18, 5.37]	**+7.54** [6.78, 8.28]
SFT (token-matched)	Comp., 10 vs. 10	**+8.13** [7.36, 8.91]	**+5.15** [4.51, 5.79]	**+8.20** [7.39, 9.03]
Paired SFT	Comp., 10 vs. 10	**+3.19** [2.63, 3.76]	**+1.95** [1.48, 2.43]	**+2.97** [2.38, 3.59]
Transported KL	Comp., 10 vs. 10	**+1.73** [1.19, 2.25]	**+1.20** [0.79, 1.64]	**+2.14** [1.57, 2.72]
LoRA-SFT	Comp., 4 vs. 16	**+3.11** [2.22, 4.03]	**+1.88** [1.16, 2.63]	**+2.14** [1.20, 3.11]
LoRA-SFT	Comp. AULC	**+11.32** [10.77, 11.89]	**+5.68** [5.17, 6.20]	**+11.33** [10.75, 11.93]
Paired SFT	Comp. AULC	**+4.32** [3.83, 4.83]	**+2.65** [2.24, 3.06]	**+4.15** [3.63, 4.68]
Transported KL	Comp. AULC	**+2.94** [2.49, 3.41]	**+1.77** [1.40, 2.15]	**+3.03** [2.56, 3.52]

Table 14: **Held-out composition acceptance and coverage.** R denotes reversal. Accept gives the percentage and exact accepted/attempted final-document counts; 100.00% can reflect rounding. Final documents in $\mathcal{G}_{\text{train}}$ are excluded. Inst. is the fraction of test instance–seed pairs with at least one accepted view. Changed/unchanged count accepted instance–seed–view combinations over five seeds, including repeated views of each input. Labels is the number of distinct test-supported semantic labels represented among changed-target views / all test-supported labels. The same draws and acceptance mask are shared by methods and budgets.

Condition	Accept (%)	Inst. (%)	Changed	Unchanged	Labels
BANKING77					
$k = 2$	99.66 (153,482 / 154,000)	100.00	7,842	145,640	77 / 77
$k = 2$ + R	99.66 (153,482 / 154,000)	100.00	7,842	145,640	77 / 77
$k = 4$	100.00 (153,994 / 154,000)	100.00	15,173	138,821	77 / 77
$k = 4$ + R	100.00 (153,994 / 154,000)	100.00	15,173	138,821	77 / 77
$k = 8$	100.00 (153,999 / 154,000)	100.00	28,516	125,483	77 / 77
$k = 8$ + R	100.00 (153,999 / 154,000)	100.00	28,516	125,483	77 / 77
CLINC150					
$k = 2$	99.90 (224,782 / 225,000)	100.00	6,003	218,779	150 / 150
$k = 2$ + R	99.90 (224,782 / 225,000)	100.00	6,003	218,779	150 / 150
$k = 4$	100.00 (224,998 / 225,000)	100.00	11,894	213,104	150 / 150
$k = 4$ + R	100.00 (224,998 / 225,000)	100.00	11,894	213,104	150 / 150
$k = 8$	100.00 (224,999 / 225,000)	100.00	22,653	202,346	150 / 150
$k = 8$ + R	100.00 (224,999 / 225,000)	100.00	22,653	202,346	150 / 150
HWU64					
$k = 2$	98.95 (53,235 / 53,800)	100.00	3,238	49,997	64 / 64
$k = 2$ + R	98.95 (53,235 / 53,800)	100.00	3,238	49,997	64 / 64
$k = 4$	99.96 (53,780 / 53,800)	100.00	6,304	47,476	64 / 64
$k = 4$ + R	99.96 (53,780 / 53,800)	100.00	6,304	47,476	64 / 64
$k = 8$	100.00 (53,799 / 53,800)	100.00	11,744	42,055	64 / 64
$k = 8$ + R	100.00 (53,799 / 53,800)	100.00	11,744	42,055	64 / 64

High acceptance and heterogeneous target-change rates. At depth two, final-document acceptance is 99.66%, 99.90%, and 98.95% for BANKING77, CLINC150, and HWU64, respectively, and it

approaches 100% at larger depths. Every reported condition covers all test instance–seed pairs and all test-supported labels among changed-target views. Coverage therefore rules out complete omission of a test instance or semantic label in these summaries, although it does not imply equal numbers of accepted views per instance or label. Exact accepted/attempted counts retain information hidden by rounded percentages; a displayed 100.00% need not mean that no view was excluded.

Among accepted views, the changed-target share increases from 5.11% to 18.52% between depths two and eight on BANKING77, from 2.67% to 10.07% on CLINC150, and from 6.08% to 21.83% on HWU64. Unchanged targets remain the majority at every reported depth. Consequently, overall composition accuracy can mask a substantial weakness on required answer changes, motivating the separate strata in Table 12. Counts are unchanged when reversal is added at a fixed depth, consistent with reversal preserving output bindings. The shared draws and acceptance mask keep these evaluation populations aligned across methods and budgets; repeated views of an input remain dependent observations rather than new independent test examples.

D.7 Margin–Residual Diagnostic

Quantities and theoretical reference. For a fixed input and candidate inventory, consider a trajectory $g_0 \to g_1 \to \cdots \to g_k$, centered candidate-score vectors $\mathbf{u}_j = \mathbf{u}_{g_j}$, and edit permutations P_j. We compute

$$\varepsilon_j = \|\mathbf{u}_j - P_j \mathbf{u}_{j-1}\|_2, \qquad E_{\mathrm{local}} = \sum_{j=1}^{k} \varepsilon_j. \tag{D.1}$$

For the correct initial output y_0, define

$$\gamma_0 = \mathbf{u}_0[y_0] - \max_{c \neq y_0} \mathbf{u}_0[c], \qquad \mathcal{R}_{\mathrm{mr}} = \frac{\sqrt{2} E_{\mathrm{local}}}{\gamma_0} \quad (\gamma_0 > 0). \tag{D.2}$$

Corollary B.4 is a deterministic statement for quantities computed on the same path. If $\mathcal{R}_{\mathrm{mr}} < 1$, the transported correct target is the unique top-1 at the final step and every prefix, since prefix sums of nonnegative residuals cannot exceed the full sum. This condition is sufficient, not necessary. In Figure 6, the certified region is above $\gamma_0 = \sqrt{2} E_{\mathrm{local}}$.

Diagnostic scope and aggregation. Each record contains 12 candidate scores per step. The diagnostic contains 4,926 trajectories per method: 1,601 from BANKING77, 1,878 CLINC150, and 1,447 HWU64. These are restricted-candidate subsets, distinct from the complete 77-, 150-, and 64-candidate inventories in the main evaluation. The counts at depths 2, 4, and 8 are 1,791, 1,600, and 1,535 per method. All recorded initial targets are uniquely top-ranked. Thus all percentages in this diagnostic condition on initial correctness in the restricted inventory.

The five seed identifiers are $\{13, 21, 42, 87, 100\}$, with 988, 986, 985, 984, and 983 trajectories per method, respectively. Methods share dataset, seed, and trajectory identifiers, initial targets, and swap sequences. Pooled percentages weight trajectories equally, so datasets, depths, and seeds contribute in proportion to their trajectory counts. We report cohort summaries without confidence intervals. Their representativeness for full-inventory evaluation remains undetermined.

Reconstruction from scores and swaps. For every active step, we center the saved 12-dimensional score vector and apply the recorded index swap to the preceding vector. We then recompute ε_j using the unnormalized Euclidean norm. The same ordered swaps transport y_0 to y_j; each recorded intermediate target is checked against this reconstruction. Final correctness compares the top-ranked candidate with y_k; the final target margin is $\gamma_k = \mathbf{u}_k[y_k] - \max_{c \neq y_k} \mathbf{u}_k[c]$. The gold-target initial margin is computed directly from $\mathbf{u}_0$.

The saved scores have ten decimal places. Recomputed residuals, ratios, and margins agree with their stored summaries under absolute and relative tolerances of 10^{-8}; the largest ratio discrepancy is 4.13×10^{-9}. Target indices and correctness labels agree exactly. We check the accumulation bound and $\gamma_j \geq \gamma_0 - \sqrt{2} \sum_{t=1}^{j} \varepsilon_t$ at all 66,786 recorded prefixes. All checks pass. No trajectory lies within 10^{-8} of the final margin-bound threshold. All records contribute to the summaries using their recorded correctness labels. These checks verify that scores, targets, and permutations are numerically consistent.

Coverage and correctness. For Paired SFT, transported KL, and SIGMA, the certified subsets contain 3,247, 3,712, and 4,252 trajectories, corresponding to 65.9%, 75.4%, and 86.3% coverage. Their final and prefix correctness is 100%, with zero violations. Final correctness outside the certified region is 13.2%, 11.8%, and 3.4%, respectively; correct answers also occur when $\mathcal{R}_{\mathrm{mr}} \geq 1$. Pooled final correctness is 70.4%, 78.3%, and 86.8%. These results agree with the deterministic condition for this cohort.

Within-band comparison. For $\gamma_0 \in [1.4, 1.8]$, counts are 890, 906, and 912 and median cumulative residuals are 0.61653, 0.46866, and 0.35280 for the three methods. Relative to Paired SFT and transported KL, the SIGMA medians are lower by 42.8% and 24.7%, computed as $100(1 - \mathrm{median}(E_{\mathrm{local}}^{\mathrm{SIGMA}})/\,\mathrm{median}(E_{\mathrm{local}}^{\mathrm{reference}}))$. Band membership can differ across methods despite shared trajectory identifiers; these are descriptive, unpaired within-band summaries.

Rendering. All panels share coordinate limits, a 180×180 histogram grid, and Gaussian smoothing bandwidth of 2.5 grid cells. Correctness-weighted counts and total counts are smoothed separately within each side of $\mathcal{R}_{\mathrm{mr}} = 1$, then divided within that side. Colors therefore represent smoothed empirical correctness. Low-density cells below the eighth percentile of positive density for their side are masked, as are grid cells intersecting the boundary, preventing interpolation across it. White contours use the 35th, 55th, 72nd, 84th, 92nd, and 97th percentiles of the full-cohort smoothed density in each panel. Display limits are approximately $E_{\mathrm{local}} \in [0.073, 1.920]$ and $\gamma_0 \in [0.201, 4.100]$; all numerical summaries use every record. For scatter, we sort by dataset, seed, and trajectory identifier and select 350 evenly spaced records per method, producing the same identifier subset across methods. Coordinates are the recomputed values, with no jitter.

What the certificate adds to accuracy. Coverage and final accuracy answer different questions. Final accuracy counts correct endpoints, while coverage identifies trajectories whose cumulative discrepancies are small enough relative to their initial margin to guarantee the transported answer. The zero violations inside the certified region agree with this deterministic implication. Correct endpoints outside it show why the inequality should not be interpreted as a necessary condition: the sum of local residual norms is an upper bound and can exceed the final discrepancy when transported residuals partly cancel. Failure to certify therefore does not itself predict an incorrect answer.

SIGMA's higher coverage is accompanied by a smaller median residual within the reported initial-margin band. This supplies a complementary view of the aggregate pattern, but does not hold the exact inputs or initial margins fixed across methods within that band. The evidence supports lower observed discrepancies and broader certification in this restricted cohort. It does not establish that the same coverage applies to the full candidate inventories or that residual reduction alone causes the accuracy difference. The unsmoothed trajectory summaries, rather than interpolated colors in the plot, determine these conclusions.

D.8 Which Spectral Directions Does SIGMA Contract?

Question and protocol. We examine whether residual reduction is concentrated in output directions exposed by the sibling-binding intervention law. SIGMA and Paired SFT use Qwen2.5-7B-Instruct, 10 shots per intent, the same labeled inputs, paired guidance views, candidate inventory, optimizer, and LoRA configuration (Appendix C.5); the equivariance penalty is the controlled difference. Frozen models are evaluated on the same held-out input–binding-edit pairs, sampled independently of the correct label, with the complete candidate inventory. The five seeds are $\{13, 21, 42, 87, 100\}$. The records contain all $m - 1$ centered directions: 76/149/63 for BANKING77/CLINC150/HWU64, yielding 1,440 seed–direction observations, including 1,270 edit-exposed and 170 null-space observations.

Intervention spectrum. Let $P_\perp = I_m - m^{-1}\mathbf{1}\mathbf{1}^\top$ and restrict the following operators to $\mathcal{H} = \mathbf{1}^\perp$. As in Appendix B.6, ρ_T assigns probability $1/2$ to representation reversal ($P_e = I_m$) and $1/2$ to binding swaps, choosing one of J eligible sibling groups uniformly and then an unordered pair uniformly within that group. For b_v leaves in group G_v,

$$C_{\mathrm{out}} = \mathbb{E}_{e\sim\rho_T}\left[\tfrac{1}{4}(P_e - I_m)^\top(P_e - I_m)\right] = \sum_{v=1}^{J} c_v \Pi_v, \qquad c_v = \frac{1}{2J(b_v - 1)}, \tag{D.3}$$

where Π_v projects onto the zero-sum contrasts within G_v. The factor $1/2$ in c_v retains the training mixture even though the diagnostic evaluates binding edits. Positive eigenvalues κ_j identify directions exposed by output permutations. The zero eigenspace is the null space of this operator, not necessarily of the full nonlinear equivariance loss.

To resolve repeated eigenvalues reproducibly, we use the fixed candidate ordering $G_v = \{g_{v,1}, \ldots, g_{v,b_v}\}$ and Helmert contrasts

$$\mathbf{v}_{v,r} = \frac{\sum_{\ell=1}^{r} \mathbf{e}_{g_{v,\ell}} - r\mathbf{e}_{g_{v,r+1}}}{\sqrt{r(r+1)}}, \qquad r = 1, \ldots, b_v - 1. \tag{D.4}$$

These have eigenvalue c_v; the remaining centered directions span $\ker(C_{\text{out}}) \cap \mathcal{H}$. The group counts are $J = 9, 10, 17$, with 68/140/46 exposed directions and 8/9/17 null directions, respectively. HWU64 has one candidate outside the eligible sibling groups.

Directional residual energy. For method M, write the aligned centered-score residual as $\boldsymbol{\delta}_{i,e}^{(M)} = \mathbf{u}_{g'}^{(M)} - P_e \mathbf{u}_g^{(M)}$. For the common evaluation pairs $\mathcal{Q}_s$ in seed s, define

$$E_{j,s}^{(M)} = \frac{1}{|\mathcal{Q}_s|} \sum_{(i,e)\in\mathcal{Q}_s} \left(\mathbf{v}_j^\top \boldsymbol{\delta}_{i,e}^{(M)}\right)^2, \qquad R_{j,s} = \log_2 \frac{E_{j,s}^{(\text{SIGMA})} + \epsilon}{E_{j,s}^{(\text{Paired})} + \epsilon}, \quad \epsilon = 10^{-12}. \tag{D.5}$$

All recorded energies are positive, and ϵ has no effect at reported precision. Values $R = -1, -2, -3$ correspond to approximately one half, one quarter, and one eighth of the reference energy. We compute pooled medians over seed–direction pairs and Spearman correlation over the 254 exposed directions after averaging $R_{j,s}$ over seeds. The 95% interval uses 2,000 sibling-group bootstrap replicates, keeping within-group directions together.

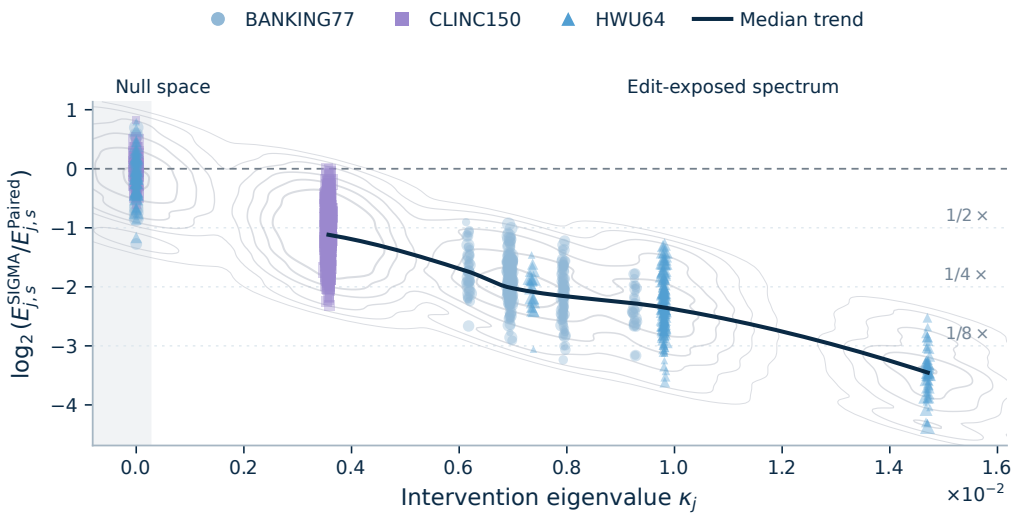


Figure 7: **Directional residual contraction under structured interventions.** Each marker is one seed–direction observation, with dataset encoded by color and shape. Negative log-ratios indicate lower residual energy under SIGMA than Paired SFT. Marker area is an affine function of baseline energy, clipped at its pooled 5th and 95th percentiles. Gray contours show pooled density; the dark curve interpolates medians from ten equal-width bins of positive spectral strength, omitting empty bins. The shaded strip marks the operator's null space; all null observations have $\kappa_j = 0$. Display jitter separates overlapping non-null points, while all statistics, contours, and trend medians use raw coordinates. Horizontal guides indicate energy ratios of $1/2$, $1/4$, and $1/8$.

Findings. Figure 7 shows nonuniform residual contraction. The pooled median is $R = \mathbf{-0.122}$ in the operator null space (91.9% retained energy), compared with $R = \mathbf{-1.539}$ in exposed directions

(34.4%). The pooled Spearman correlation is $\rho = -\mathbf{0.878}$, with a group-bootstrap 95% interval of $[-0.913, -0.778]$. Using the lower and upper quartile cutoffs of positive κ_j, the weak-strength subset has median $R = -1.118$ (46.1% retained), compared with $R = -2.403$ (18.9%) in the strong-strength subset. Including ties at the cutoffs of 0.003571 and 0.007937 selects 700 and 345 observations, respectively, from the 1,270 exposed observations.

Dataset-specific comparisons qualify the pooled trend. For BANKING77, CLINC150, and HWU64, median R is respectively -0.274, 0.004, and -0.130 in the null space, and -2.001, -1.118, and -2.527 in exposed directions. Within-dataset Spearman correlations are -0.585 for BANKING77 and -0.764 for HWU64. All CLINC150 exposed directions have the same eigenvalue, so a within-dataset correlation is undefined. The pooled correlation therefore also reflects differences between datasets and their hierarchy spectra.

Interpretation and scope. SIGMA shows larger residual-energy reductions in edit-exposed directions than in the operator null space, consistent with the regularizer's spectral interpretation. This diagnostic measures aligned residual energy; overall candidate-score variance and a causal link to composition gains remain unmeasured. Guidance dependence can produce residuals in the output-permutation null space. Within repeated eigenspaces, directional energies depend on the chosen basis, while total subspace energy is basis invariant. The available data contain directional energies and bootstrap correlations. Reconstructing them requires candidate-score vectors, basis matrices, evaluation-pair counts, and bootstrap indices, which are unavailable.

Relating the spectrum to the empirical constraints. The contrast between exposed and null directions clarifies what the chosen edit family can emphasize. A sibling swap directly acts on within-group output contrasts, whereas a group-constant component is unchanged by that permutation. The observed separation between these subspaces is compatible with the explicit model in Appendix B.6. It does not imply that every score component in the neural model obeys the linear proxy: the scores can also change because the guidance text itself changes. In particular, the small null-space energy change is an observed comparison, not a guarantee that training leaves those directions untouched.

The pooled association also should not be read as a response curve obtained by experimentally varying one eigenvalue. Spectral strengths are fixed by the task hierarchies, and the constant exposed eigenvalue within CLINC150 contributes no within-dataset ordering. Together with the negative correlations in BANKING77 and HWU64, the results support a structured, nonuniform contraction pattern across the observed tasks. They do not identify an optimal sibling-group size or justify changing the hierarchy solely to increase spectral strength.

D.9 Where Does Guidance-Aware Behavior Reside?

We localize the dependence of composition performance on the learned LoRA updates by removing one projection's update at a time. This intervention measures the effect of each removal in an otherwise intact SIGMA checkpoint. Comparing its effect on Base and composition accuracy separates their differential sensitivity to the knockout from the composition drop alone.

Single-module knockout. We use the primary Qwen2.5-7B-Instruct backbone at 10 shots per intent, with the LoRA configuration in Appendix C.5: rank $r = 16$, scaling $\alpha = 32$, and zero dropout. At layer ℓ and projection p, the effective weight is

$$W_{\ell,p}^{\text{eff}} = W_{\ell,p}^{(0)} + \frac{\alpha}{r} B_{\ell,p} A_{\ell,p}, \qquad W_{\ell,p}^{-\ell,p} = W_{\ell,p}^{(0)}. \tag{D.6}$$

The intervention removes only this low-rank update; the pretrained projection and all other LoRA updates remain unchanged. There is no retraining after removal. We test the attention query, key, value, and output projections (Q/K/V/O) and the MLP Gate/Up/Down projections in all 28 layers, indexed 0–27. This gives 196 single-module interventions per checkpoint. Across three datasets and five seeds $\{13, 21, 42, 87, 100\}$, the analysis includes 2,940 records, with 15 observations for every layer–projection cell.

Within each dataset and seed, the intact and knockout models are evaluated on the same Base inputs and held-out guidance compositions. The full candidate inventory, guidance edits, and evaluation settings are fixed across knockouts. Each intervention therefore has its own paired comparison with the recorded intact model.

Differential accuracy and equivariance effects. Let A_v^{full} and $A_v^{-\ell,p}$ denote intact and knockout accuracy for view $v \in \{\text{base}, \text{comp}\}$, in percent. For each dataset and seed, define

$$\begin{aligned} D_{\text{base}}(\ell, p) &= A_{\text{base}}^{\text{full}} - A_{\text{base}}^{-\ell,p}, \\ D_{\text{comp}}(\ell, p) &= A_{\text{comp}}^{\text{full}} - A_{\text{comp}}^{-\ell,p}, \\ I_{\text{guide}}(\ell, p) &= D_{\text{comp}}(\ell, p) - D_{\text{base}}(\ell, p). \end{aligned} \tag{D.7}$$

Positive I_{guide} means a larger accuracy loss for composition than for Base, measured in percentage points (pp). It measures the differential effect of a single-module knockout. Using the distributional metric in Appendix C.3, define

$$I_{\text{eq}}(\ell, p) = \text{TV}_{\text{eq}}^{-\ell,p} - \text{TV}_{\text{eq}}^{\text{full}}. \tag{D.8}$$

Positive I_{eq} denotes worse equivariance and is an absolute TV difference, without percentage scaling. We compute effects before averaging, assigning equal weight to each of the three datasets and five seeds. Layer-band summaries also weight their modules and layers equally.

Intact-model reference. The five-seed intact Base/Comp. means in the knockout evaluation are 90.698/88.408 for BANKING77, 96.567/95.357 for CLINC150, and 91.389/89.064 for HWU64. Their small differences from Table 1 remain unexplained in the available records. Each knockout effect is computed against its paired intact reference, and the conclusions apply to these comparisons.

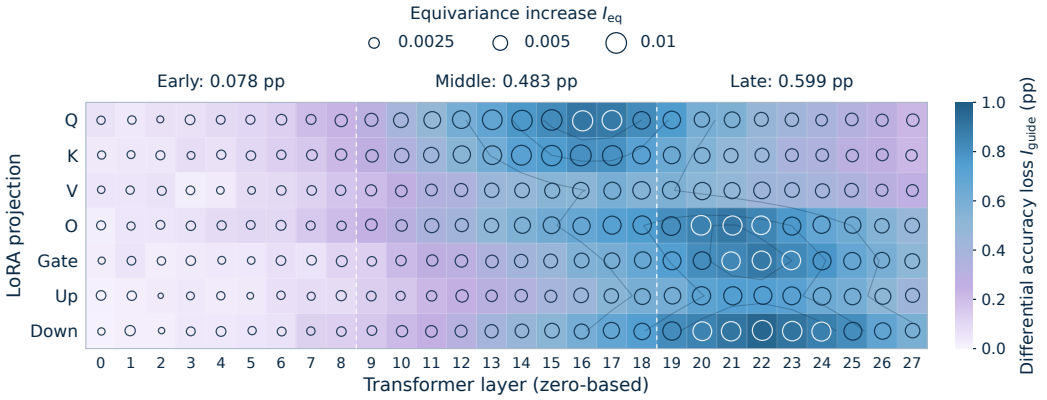


Figure 8: **LoRA module knockout effects on guidance composition.** Each cell removes one low-rank projection update from SIGMA. Color shows the additional composition-accuracy loss beyond the Base loss, I_{guide} (pp); hollow circles encode the increase in equivariance error, I_{eq}, on an absolute TV scale. Circle area is an affine function of I_{eq} over the observed cell-mean range, with the same mapping in the legend. Cells average three datasets and five seeds equally. Band labels report mean I_{guide} over layers 0–8, 9–18, and 19–27. Heatmap cells show unsmoothed means; contours are a smoothed visual guide. Circle offsets are confined to their cells and used only for display.

Layer- and projection-specific effects. Figure 8 shows a nonuniform knockout profile. Mean I_{guide} rises from **0.078 pp** in layers 0–8 to **0.483 pp** in layers 9–18 and **0.599 pp** in layers 19–27. The same early-to-late ordering holds separately on all three datasets. Q and K peak at layer 16, with effects of 0.899 and 0.804 pp. The O projection peaks at layer 21 (0.893 pp), while Gate and Down peak at layer 22 (0.889 and 0.963 pp). For the largest effect, layer-22 Down, the Base and composition drops are 0.378 and 1.340 pp; their difference is 0.963 pp.

The two response metrics agree spatially: across the 196 cell means, the descriptive Spearman correlation between I_{guide} and I_{eq} is **0.965**. Layer-16 K has the largest I_{eq} (0.01026), while layer-22 Down increases it by 0.00946. The two metrics describe coupled responses to the same knockouts: larger differential composition losses accompany larger equivariance errors.

Interpretation and limits. Under this single-update removal, the adapted model depends more strongly on middle-layer Q/K updates and later O/MLP updates for composition performance relative

to Base prediction. Their depth ordering is compatible with guidance conditioning followed by candidate-evidence formation; this proposed sequence and the semantic roles of individual projections remain hypotheses. Knockout effects can depend on update magnitude, parameter count, redundancy, downstream compensation, and task difficulty beyond the observed Base-accuracy loss. Necessity, sufficiency, and interactions remain unresolved because the study includes only single-module removals, with no resampling or restoration. The available aggregate metrics lack per-instance outcomes and checkpoint identifiers. Localization is therefore descriptive across the 15 dataset–seed comparisons, without inferential intervals over the 196 module positions.

Reading a knockout effect. The heatmap reports a difference of accuracy losses. A positive cell means that removing the update harms composition accuracy more than Base accuracy; it does not mean that the update exclusively implements guidance use. For example, layer-22 Down affects both views, with the larger loss under composition producing its high differential score. Subtracting the Base loss helps separate this asymmetry from a general accuracy decline, but cannot control all differences in difficulty or redundancy between the two evaluations. A small differential score can also coexist with appreciable losses in both views.

Distributed dependence across depth. The early, middle, and late band means summarize different sets of individual removals. Their ordering indicates where composition performance is more sensitive to this intervention on average, while the Q/K and O/MLP peaks show that projection type matters within a band. These effects cannot be added to estimate the damage from removing an entire band: simultaneous removals could interact or reveal compensation that is hidden in a single-module experiment. Likewise, a low-impact update need not be dispensable after other updates are removed. The present map supports prioritizing specific modules for follow-up tests, rather than prescribing a smaller adapter architecture without retraining and evaluation.

Connecting output geometry and parameter interventions. The spectral analysis and the knockout map describe different levels of the same trained system. The former locates residual-energy changes in output directions exposed by the edit law; the latter measures performance and equivariance changes after removing individual parameter updates. Their agreement with improved guidance use is informative, but the available records do not connect a particular projection to a particular spectral direction. Establishing that link would require scores from the same knockout runs projected onto the specified basis, together with matched reference and restoration conditions. Such measurements are not part of the reported results.

Comparability across diagnostic units. The three diagnostics use different units of analysis. The margin study weights initially correct, 12-candidate trajectories equally; the spectral summary pools seed–direction observations from the full candidate inventories; and the knockout map gives equal weight to each dataset and seed when averaging a module's effect. A dataset with more exposed directions consequently contributes more observations to the pooled spectral median, but receives no extra weight in a knockout cell mean. Similarly, certification coverage is conditional on the margin study's restricted cohort, whereas the knockout accuracy losses use their own paired intact references. Their numerical magnitudes therefore should not be combined into a single effect estimate. Each analysis supplies a different check: whether a score trajectory satisfies a sufficient condition, where aligned output discrepancies are reduced, and which parameter removals disproportionately harm composition performance. Agreement across these checks is useful as complementary evidence, while direct mediation between them remains untested.

The combined evidence therefore supports a bounded conclusion: SIGMA improves behavioral responses to guidance changes, shows structured reductions in aligned residual energy, and depends unevenly on its learned LoRA updates. It does not establish a unique circuit or a complete causal sequence from an edit through hidden representations to the final answer. This distinction preserves the role of the mathematical guarantees, which concern specified scores and edit sequences, while keeping the empirical localization tied to the actual interventions and cohorts evaluated here.