WHERE DO HEAD-LEVEL ADAPTATION GAINS COME FROM? DESIGN-SPACE EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Head-level adaptation has an attribution gap: complete pipelines are evaluated, but individual stages receive credit. Existing methods typically fine-tune random or all heads, or localize important heads and update them directly, bundling design choices whose contributions remain unresolved. Direct updating further assumes that diagnostic importance implies intervention utility, leaving a consequential choice implicit: *Bridging*, the mapping from localized heads to intervention targets. On Qwen2.5 with GSM8K, adding bridging shifts a localization-guided configuration from 5.9 percentage points below a random-head baseline to 4.8 points above it. To disentangle these contributions, we formulate an Evaluatology-based self-contained adaptation design system (SCDS). SCDS represents localization, bridging, and fine-tuning as attribute designed objects (DOs), their composition into adaptation paths as a structural DO, and model, dataset, and run-level conditions as confounding objects (COs). A blocked factorial study with matched interventions examines 14 localization rules, three bridging rules, and four fine-tuning rules across four model–dataset blocks, alongside random-head and all-head baselines. Across all four blocks, the structural path, bridging, and fine-tuning exhibit significant effects, whereas localization has a small, non-significant main effect. Of the two bridging rules, representation bridging significantly outperforms both Localize–FT and Random–FT in three of four blocks, while information-flow bridging shows no significant difference from the best baseline in the fourth. Bridging can therefore help, but only when the rule matches the block. Generalization tests of the same adapted models on two new datasets broadly corroborate the structural and bridging-rule findings. SCDS thus turns pipeline comparison into evidence of which design choices contribute to adaptation gains and how their effects depend on composition and evaluation conditions.

1 INTRODUCTION

Head-level adaptation has emerged as a promising approach to targeted and parameter-efficient language-model adaptation (Zhang et al., 2024; Yin et al., 2024). Research in this area has developed along two closely related directions. One asks *how* attention heads should be updated, leading to selective head fine-tuning, selective head LoRA, head output bias, and head output remapping (Zhang et al., 2024; Chen et al., 2024; Li et al., 2023; Yin et al., 2024; Wang et al., 2025; Su et al., 2026). The other asks *where* adaptation should be applied, using intervention, representation or parameter-based signals to localize heads associated with a target capability (Meng et al., 2022; Li et al., 2023; Nikankin et al., 2025; Zhang et al., 2024; Wang et al., 2025; Zhang et al., 2026; Chen et al., 2025; Yin et al., 2024; Su et al., 2026). Combining these directions gives the prevailing localize-then-fine-tune paradigm: diagnose important heads and restrict an update operator to those heads.

These advances expose a fundamental attribution gap. An adaptation outcome jointly depends on head selection, the update rule, their composition, and the conditions under which the model is trained and evaluated. Random-head baselines and component ablations provide useful conditional comparisons, but do not by themselves establish whether a gain persists across other rules or settings. For example, on Llama-3.1 with ARC-Challenge, configurations using the same ablation localization lie 10.4 percentage points above a random-head reference with selective head fine-tuning, but 25.2 points below it with head output bias. A reported gain therefore cannot be attributed to any one choice: it may come from localization, fine-tuning, their interaction, or other conditions.

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.

Beyond this attribution gap, localize-then-fine-tune pipelines often retain a further implicit assumption: heads identified as important for a capability or behavior should also be the targets for changing it. Yet diagnosing a head's role in the current model does not establish the benefit of updating it. A similar mismatch has been observed at the MLP-layer level in factual knowledge editing, where causal-tracing scores poorly predict edit success (Hase et al., 2023). We make the mapping from localized heads to intervention targets explicit as *Bridging* and study its contribution jointly with localization, fine-tuning, and their composition. This choice can materially change the conclusion drawn from a localization-guided pipeline. For example, on Qwen2.5-1.5B-Instruct with GSM8K, path patching with head output bias lies 5.9 points below the random-head reference (69.7%) when the localized heads are updated directly, but 4.8 points above it after bridging. On Llama-3.1-8B-Instruct with ARC-Challenge, logit attribution localization with head output remapping lies 1.8 points below the random-head reference (70.5%) under direct update, but 9.3 points above it after bridging.(Appendix A.7) Whether localization helps therefore depends on how diagnosed heads are mapped to intervention targets, not only on which heads are diagnosed.

To attribute gains across this design space, we formulate head-level adaptation as a self-contained adaptation design system (SCDS; Figure 1), based on Evaluatology (Zhan et al., 2024; 2025). SCDS organizes the pipeline into designed objects (DOs)—the choices whose effects are to be attributed—and confounding objects (COs)—the non-target conditions under which those choices are trained and evaluated. The attribute DOs are localization, bridging, and fine-tuning; a structural DO specifies the adaptation path, including which attribute DOs are active and how the intervention heads are constructed. It has four paths: All–FT, Random–FT, Localize–FT, and Localize–Bridge–FT. Setting-level COs specify the model and dataset; run-level COs specify the applicable sample budgets, head budget, and optimization hyperparameters. We identify their effects with a blocked factorial design and controlled object interventions. Legal DO combinations are fully enumerated within each model–dataset block, while run-level COs are repeatedly sampled and reused across blocks. Factorial analysis then estimates structural, attribute, and CO effects in the domains where they are defined; matched interventions vary one target DO while holding the remaining DOs and COs fixed.

Our results show that adaptation structure, bridging, and fine-tuning significantly affect performance across all four blocks, whereas the localization rule has a small, non-significant main effect. Fine-tuning has the largest marginal effect among attribute DOs, and bridging interacts with it in every block. Of the two bridging rules, representation bridging significantly outperforms both Localize–FT and Random–FT in three of four blocks, while information-flow bridging shows no significant difference from the best baseline in the fourth. Bridging can therefore help, but only when the rule matches the block. The preferred configurations vary across settings, while the main structural and bridging findings broadly extend to two new datasets without further adaptation. These results provide a systematic account of adaptation gains: the value of each design choice depends on how it is composed with other choices and under which conditions it is evaluated.

2 Related Work

Localization. Component localization uses activation patching (Nikankin et al., 2025; Zhang et al., 2024; Chen et al., 2024), ablation (Zhou et al., 2025; Wang et al., 2025), loss sensitivity (Michel et al., 2019), logit relevance, and attention confidence (Voita et al., 2019), parameter differences (Chen et al., 2025), concept similarity (Su et al., 2026), information richness and correlation (Gui & Xiao, 2023), weight norms and attention entropy (Yu et al., 2025), and entropy tracing (Zhang et al., 2026). These methods characterize component function or redundancy. Their diagnostic criteria need not directly identify effective adaptation targets: at the MLP-layer level in factual knowledge editing, Hase et al. (2023) find that causal-tracing scores poorly predict edit success. Our focus is task-level attention-head adaptation and the design choices connecting diagnosis to model updates.

Head-level adaptation. Path patching is paired with in-place head training in Precise SFT (Zhang et al., 2024) and with LoRA in SPT (Chen et al., 2024). FocalLoRA selects heads sensitive to instruction conflicts for targeted LoRA updates (Shi et al., 2026). ALPS uses parameter differences (Chen et al., 2025), while HiFi uses information richness and inter-head dependence (Gui & Xiao, 2023), to select heads for fine-tuning. Entropy tracing targets persistently uncertain heads, linking selection to their adaptation potential (Zhang et al., 2026). ITI combines probing with inference-time acti-

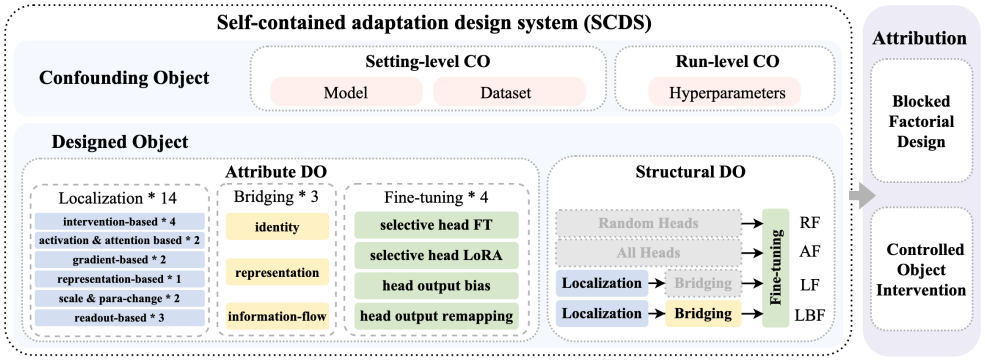

Figure 1: The Self-contained adaptation design system (SCDS) Framework

vation shifts (Li et al., 2023); LoFiT learns selection scales followed by additive offsets (Yin et al., 2024). HeadMap pairs layer-conditioned localization with low-rank output remapping (Wang et al., 2025), and SAMD/SAMI combines concept-aligned selection with scalar modulation (Su et al., 2026). JoLA jointly learns head-selection gates and intervention vectors (Lai et al., 2025). These approaches develop specific selection–update strategies, including strategies optimized jointly.

Design attribution. Prior work primarily develops better head-selection and update mechanisms. Such method-centered studies assess particular adaptation strategies, leaving the contributions and interactions of alternative design choices insufficiently resolved. To our knowledge, we present the first systematic design-attribution study of head-level adaptation that jointly examines localization, bridging, fine-tuning, and adaptation paths under controlled evaluation conditions.

3 METHOD

3.1 SELF-CONTAINED ADAPTATION DESIGN SYSTEM

We formulate a self-contained adaptation design system (SCDS) as an object-based representation of head-level adaptation for design attribution. SCDS distinguishes designed objects (DOs), which encode the adaptation choices under study, from confounding objects (COs), which describe non-target conditions that may also affect the outcome. Note that COs are conditions in SCDS, not SCM/causal-inference confounders. DOs are organized hierarchically: attribute DOs specify the rules used by individual adaptation operations, whereas a structural DO specifies how those operations form a complete adaptation path. Together, this organization yields a closed, controllable, and compositional design space in which attribution is traceable and object-isolated contrasts admit conditional causal interpretation.

Attribute DOs We identify three attribute DOs corresponding to three adaptation operations. Localization (L) identifies a diagnosed set of important heads I from the base model M_0. Bridging (B) maps I to an intervention head set H. Fine-tuning (F) modifies the model through H to produce the adapted model M_1, from which the outcome Y is measured.

LOCALIZATION (L) We evaluate 14 localization rules from six representative families: (i) intervention-based—path patching, activation patching, ablation, and layer-conditioned ablation (HeadMap); (ii) activation-and attention-based—attention scoring, and activation scoring; (iii) readout-based—logit lens, entropy tracing, linear probing; (iv) representation-based—Scalable Attention Module Discovery (SAMD); (v) gradient-based—loss-gradient importance, logit attribution; and (vi) learned-scale and parameter-change—scale-based selection (LoFiT) and parameter-difference localization (ALPS). Together, they span complementary intervention, representation, activation, gradient, and parameter-based diagnostic signals.

BRIDGING (B) We evaluate three bridging rules and make explicit whether the heads identified by diagnosis should also be the heads modified during adaptation. Identity rule (B^{id}) means no bridging, and the diagnosed heads are intervention targets directly.

- *Representation remapping* (B^{rep}). Let $z_{\ell,h}$ be the pre-projection attention output of head (ℓ, h) and $W_O^{(\ell,h)}$ its output matrix, so that $o_{\ell,h} = z_{\ell,h} W_O^{(\ell,h)}$ is the residual write of that head. The writes of all heads in I are pooled as separate residual vectors, and a centered principal subspace $\mathcal{U}_I$ is estimated from them. Its rank is the number of leading components needed to reach 90% cumulative variance, capped at 32. For each candidate head j, one gradient step on its output matrix supplies the update direction $\Delta o_j = z_j(-\nabla_{W_{O,j}}\text{loss})$, where loss is the continuation loss on the same examples. With $\Pi_{\mathcal{U}_I}$ the orthogonal projector onto $\mathcal{U}_I$, the score is

$$\mathbb{E}\left[\frac{\|\Pi_{\mathcal{U}_I}\Delta o_j\|_2^2}{\|\Delta o_j\|_2^2}\right]. \tag{1}$$

  H is the set of $|I|$ heads with the largest score. The intervention set therefore has the same cardinality as I and is selected by residual-update alignment.
- *Information-flow expansion* (B^{flow}). Each diagnosed head is a seed. For a seed A_l in layer ℓ, every head A_{l-1} in layer $\ell - 1$ is scored by the mean norm of its residual write after weighting by A_l's attention pattern, $\|P_{A_l} o_{A_{l-1}}\|$, and every head A_{l+1} in layer $\ell + 1$ is scored by the mean norm of A_l's write after weighting by A_{l+1}'s attention, $\|P_{A_{l+1}} o_{A_l}\|$. Here o is the residual write above, and P_A weights a residual write by head A's attention pattern. The highest-scoring upstream head and the highest-scoring downstream head are kept. The first layer admits only a downstream neighbor, and the last layer only an upstream neighbor. H is the deduplicated union of the seeds and these neighbors, and therefore contains I.

FINE-TUNING (F) We evaluate four fine-tuning rules. Selective head fine-tuning directly updates the parameters associated with H; selective head LoRA applies low-rank updates to those parameters; head output bias learns an additive modification to head outputs; and head output remapping learns a transformation of those outputs. The four rules therefore differ in the form of the update applied through the same intervention head set.

Structural DO (S) The structural DO specifies a complete adaptation path, including how the intervention head set H is constructed and which attribute DOs are active. We consider four paths:

$$\begin{aligned}
\text{All–FT (AF):} \quad & M_0 \xrightarrow{\text{all-head selection}} H \xrightarrow{F} M_1 \longrightarrow Y, \\
\text{Random–FT (RF):} \quad & M_0 \xrightarrow{\text{random-head selection}} H \xrightarrow{F} M_1 \longrightarrow Y, \\
\text{Localize–FT (LF):} \quad & M_0 \xrightarrow{L} \underbrace{I = H}_{\text{direct intervention}=B^{id}} \xrightarrow{F} M_1 \longrightarrow Y, \\
\text{Localize–Bridge–FT (LBF):} \quad & M_0 \xrightarrow{L} I \xrightarrow{B^{\text{rep}}/B^{\text{flow}}} H \xrightarrow{F} M_1 \longrightarrow Y.
\end{aligned} \tag{2}$$

The four paths are distinguished by how they construct H: All–FT uses all heads, Random–FT a budget-matched random set, Localize–FT the diagnosed set itself ($H = I$), and Localize–Bridge–FT a bridging transformation of I. Accordingly, identity bridging (B^{id}) encodes Localize–FT, whereas representation (B^{rep}) and flow (B^{flow}) bridging are exclusive to Localize–Bridge–FT; only rules associated with operations active on a path are admissible.

Confounding Objects COs are conditions whose variation may change the observed outcome even when the DO assignment remains unchanged. We distinguish setting-level and run-level COs. The setting-level COs are the base model M_0 and dataset DS, while the run-level COs are the hyperparameter configurations. A CO configuration is

$$(M_0, DS, h), \qquad h = \left(\underbrace{n_L, n_B, n_F}_{\text{sample budgets}}, \underbrace{k}_{\text{head budget}}, \underbrace{lr, e, r}_{\text{optimization settings}}\right), \tag{3}$$

where n_L, n_B, and n_F are the numbers of samples used by Localization, Bridging, and Fine-tuning; k is the number of heads selected per layer; lr is the learning rate; e is the number of epochs; and r is the low-rank dimension. Head output bias does not use lr, e, or r; The applicable coordinates depend on the active DOs; Table 1a specifies their grids and applicability.

3.2 Factorial attribution and controlled intervention

Given the objects defined by SCDS, we use two complementary designs to identify their influence: a blocked factorial design for design-space attribution and matched interventions for object isolation. The former estimates how outcome variation is distributed across objects and their interactions; the latter isolates a target object by holding the remaining configuration fixed.

3.2.1 Blocked Factorial Design and Effect Estimation

We use a blocked factorial design, with each model–dataset pair defining a block. Within each block, we enumerate all $14 \times 3 \times 4$ combinations of L, B, and F. Each design cell is replicated with run-level COs sampled using independent random seeds, and the same sampled assignments are reused across blocks. Accordingly, complete enumeration prevents selective DO coverage, repeated CO sampling averages over run-level variation, and reusing the same draws across blocks enables aligned cross-setting comparisons. Random–FT enumerates all combinations of three head budgets and four fine-tuning rules, and All–FT enumerates all four fine-tuning rules, with run-level COs sampled from independent random seeds. Note that failed localization runs are omitted from statistics (less than 1% in our experiments). The number of CO samples depends on pilot-based simulation. Candidate seed counts are swept by doubling, and we retain the smallest count at which the 95% confidence-interval half-width of the mean accuracy falls below one percentage point on Localize–FT (shared with Localize–Bridge–FT) and Random–FT (shared with All–FT). We further assess robustness using leave-one-seed-out analysis for the $L \times B \times F$ factorial.

We fit separate ANOVA models for factors defined over different design domains rather than pooling all runs into a single model. The $L \times B \times F$ analysis estimates their main effects and interactions; the structural analysis compares the four paths; and each run-level CO is analyzed only where it is applicable. Model and dataset enter as setting-level factors, with their interactions used to assess cross-setting heterogeneity. We report effect sizes, significance, and bootstrap confidence intervals.

3.2.2 Controlled object intervention

To isolate a target DO d_j, we construct matched configurations that vary only its level while holding all other DOs and COs fixed, including run-level randomness. Across m matched configurations, the intervention effect is $\widehat{\tau}_j(a, a') = \frac{1}{m}\sum_{i=1}^{m}\left[Y(d_j = a; x_i) - Y(d_j = a'; x_i)\right]$, where x_i collects the fixed non-target components. Structural interventions also align the attributes shared by the compared legal paths. Confidence intervals are obtained by paired bootstrap over matched sets.

4 Experiments

4.1 Setup

We evaluate the SCDS on four model–dataset blocks: Llama-3.1-8B-Instruct(Grattafiori et al., 2024) and Qwen2.5-1.5B-Instruct(Yang et al., 2024), each paired with GSM8K(Cobbe et al., 2021) and ARC-Challenge(Clark et al., 2018). To keep the sample sets disjoint, from each test split we randomly hold out 200 samples as the fixed fine-tuning evaluation set, and use the remainder for localization. Bridging and fine-tuning draw their samples from the training split. We further validate generalization on new datasets—SVAMP and OpenBookQA (Section A.4). For Localize–FT (LF) and Localize–Bridge–FT (LBF), each block uses four stratified sampling seeds with 168 abstract plans per seed, covering all $14 \times 3 \times 4$ Localize–Bridging–Fine-tuning combinations once per seed while drawing the run-level CO h from Table 1a. All–FT and Random–FT (RF) baselines use the same fine-tuning rules and run-level CO grids: RF draws 20 joint replicates per (rule, k) cell, and All–FT draws 20 replicates per rule. These counts are the pilot estimate on Qwen–GSM8K, the smaller model on the noisier dataset: four seeds for $L \times B \times F$ (half-width 0.95 pp) and twenty replicates for Random–FT and All–FT (half-width 0.79 pp). Table 1b reports the sample allocation.

Table 1: Factorial Design. Rep. denotes independently seeded CO samples per design cell.

(a) Hyperparameter grids for h.

Coordinate	Grid	Notes
Localization samples	$\{64, 128, 192\}$	$\{300, 600, 900\}$ for probe/LoFiT scale.
Heads per layer	$\{1, 2, 3\}$	Budgets: 32/64/96 (Llama), 28/56/84 (Qwen).
Bridging samples	$\{64, 128, 192\}$	B^{rep} and B^{flow} only.
FT train samples	$\{256, 512, 768\}$	Shared across all F.
Learning rate	$\{10^{-5}, 10^{-4}\}$	Head FT, LoRA, and remapping.
Epochs	$\{1, 2\}$	Head FT, LoRA, and remapping.
Rank	$\{8, 16\}$	Head LoRA and remapping.

(b) Total samples. HB: head budget

Stratum	Factor grid	Seeds	Blocks	Total
$L \times B \times F$	$14 \times 3 \times 4$	4	2×2	2688
Random–FT	$3_{\text{HB}} \times 4$	20	2×2	960
All–FT	4	20	2×2	320
Total				**3968**

Table 2: One-way η^2 of the structural DO, Localization, Bridging, Fine-tuning, and run-level COs, with ANOVA p-values. Each column is a separate one-way ANOVA. S uses all structural DOs (Localize–FT, Localize–Bridge–FT, Random–FT, and All–FT). All other factors use Localize–FT / Localize–Bridge–FT (guided) runs; n_L drops probe / LoFiT; n_B only if $B \neq \text{id}$; lr and e on SGD Fine-tuning; r on selective head LoRA/head output remapping.

Block	η^2_S	η^2_L	η^2_B	η^2_F	$\eta^2_{n_L}$	η^2_k	$\eta^2_{n_B}$	$\eta^2_{n_F}$	η^2_{lr}	η^2_e	η^2_r
Llama–GSM8K	0.126 $p<0.001$	0.034 $p=0.058$	0.049 $p<0.001$	0.191 $p<0.001$	0.003 $p=0.458$	0.009 $p=0.051$	0.003 $p=0.519$	0.008 $p=0.075$	0.254 $p<0.001$	0.021 $p=0.001$	0.002 $p=0.443$
Llama–ARC	0.045 $p<0.001$	0.017 $p=0.595$	0.109 $p<0.001$	0.413 $p<0.001$	0.004 $p=0.326$	0.001 $p=0.664$	0.001 $p=0.848$	0.009 $p=0.054$	0.238 $p<0.001$	0.017 $p=0.003$	0.009 $p=0.086$
Qwen–GSM8K	0.115 $p<0.001$	0.022 $p=0.304$	0.053 $p<0.001$	0.180 $p<0.001$	0.005 $p=0.260$	0.006 $p=0.113$	0.009 $p=0.140$	0.005 $p=0.167$	0.369 $p<0.001$	0.031 $p<0.001$	0.013 $p=0.034$
Qwen–ARC	0.011 $p=0.013$	0.017 $p=0.564$	0.036 $p<0.001$	0.164 $p<0.001$	0.002 $p=0.590$	0.001 $p=0.844$	0.019 $p=0.015$	0.007 $p=0.100$	0.005 $p=0.125$	0.000 $p=0.635$	0.001 $p=0.492$

A leave-one-seed-out analysis corroborates the four-seed count: dropping any single sampling seed, the order of all ten factors stays highly correlated with the full sample, with both average/median Spearman's ρ of 0.96 on the η^2 order and 0.93 on the p-value order (Table 14). Each run uses a single NVIDIA A800 80GB GPU under PyTorch 2.5.1 and CUDA 12.4.

4.2 Overview of variance attribution

We report main and interaction effects of DOs and COs from separate ANOVA models in their applicable design domains, within each model–dataset block. Table 2 reports one-way η^2 for the structural DO, the three attribute DOs, and the sampled run-level COs; Table 6 reports the corresponding two-way interactions.

Main Effects. *(1) Attribute DOs:* On the $L \times B \times F$ runs, fine-tuning dominates ($\eta^2_F = 0.16$–0.41, $p < 0.001$ on every block). Bridging is smaller but still significant ($\eta^2_B = 0.036$–0.109, $p < 0.001$). Localization remains small and non-significant ($\eta^2_L = 0.017$–0.034). *(2) Structural DO:* S is significant on every block ($\eta^2_S = 0.011$–0.126) and largest on the two GSM8K blocks. These estimates are not commensurate with the attribute-DO η^2 values, which are computed on the $L \times B \times F$ runs only. *(3) Run-level COs.* Most sampled hyperparameters remain small. Learning rate dominates on three blocks ($\eta^2_{lr} = 0.238$–0.369, $p < 0.001$) and is negligible on Qwen–ARC ($\eta^2 = 0.005$, $p = 0.125$); epochs are significant on the same three blocks ($\eta^2_e = 0.017$–0.031).

Interactions. Where both factors have significant main effects, the interactions remain detectable: $B \times F$ is significant on every block (Table 6). Although the learning-rate main effect is significant on three blocks, its interaction with fine-tuning is significant on all four. By contrast, interactions that pair a significant factor with a non-significant one are sporadic. $L \times F$ and $L \times B$ each reach significance in only one block, consistent with the weak main effect of localization.

4.3 Structural DO comparison and intervention

Figure 2 reports mean accuracy for Random–FT (RF), All–FT (AF), Localize–FT (LF), and four Localize–Bridge–FT (LBF) summaries. $\text{LB}_\text{a}\text{F}$ averages B^{rep} and B^{flow}. $\text{LB}_\text{r}\text{F}$ and $\text{LB}_\text{f}\text{F}$ apply representation or information-flow bridging rule to all samples, respectively. $\text{LB}_\text{m}\text{F}$ uses the better B^{rep} or B^{flow} per Localize–FT pair. Table 3 further reports pairwise differences among these structural summaries, with 95% bootstrap CIs and two-sided p-values.

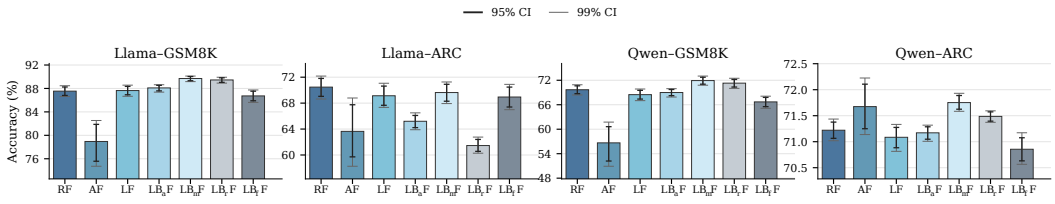

Figure 2: Structural DO mean accuracy (%). Random–FT (RF), All–FT (AF), Localize–FT (LF), and Localize–Bridge–FT (LBF): $\mathrm{LB_aF}$ (average of B^{rep} and B^{flow}), $\mathrm{LB_mF}$ (better B^{rep} or B^{flow} per Localize–FT pair), $\mathrm{LB_rF}$ (using B^{rep}), and $\mathrm{LB_fF}$ (using B^{flow}). Error bars are percentile bootstrap CIs (thick inner: 95%, thin outer: 99%); y-axis limits are block-specific.

Table 3: Accuracy differences in percentage points (95% bootstrap CIs). Columns are LF−RF (Localize–FT minus Random–FT); $\mathrm{LB_fF}$−LF and $\mathrm{LB_rF}$−LF ($\mathrm{LB_fF}$: flow Bridging; $\mathrm{LB_rF}$: representation Bridging; both minus Localize–FT); $\mathrm{LB_aF}$−LF ($\mathrm{LB_aF}$: average of both Bridging rules); $\mathrm{LB_mF}$−LF ($\mathrm{LB_mF}$: better B^{rep} or B^{flow} per Localize–FT pair); $\mathrm{LB_fF}$−RF and $\mathrm{LB_rF}$−RF; $\mathrm{LB_aF}$−RF; and $\mathrm{LB_mF}$−RF. Positive values favor the first term. p is two-sided bootstrap for a zero mean difference ($\mathrm{LB_mF}$ columns unadjusted for rule selection).

Block	LF−RF	$\mathrm{LB_fF}$−LF	$\mathrm{LB_rF}$−LF	$\mathrm{LB_aF}$−LF	$\mathrm{LB_mF}$−LF	$\mathrm{LB_fF}$−RF	$\mathrm{LB_rF}$−RF	$\mathrm{LB_aF}$−RF	$\mathrm{LB_mF}$−RF
Llama–GSM8K	+0.1 [-0.9, +1.1] p=0.868	-0.9 [-2.0, +0.1] p=0.099	+1.8 [+1.0, +2.7] p<0.001	+0.4 [-0.4, +1.3] p=0.321	+2.1 [+1.2, +2.9] p<0.001	-0.8 [-1.9, +0.2] p=0.148	+1.9 [+1.1, +2.8] p<0.001	+0.6 [-0.3, +1.4] p=0.210	+2.2 [+1.4, +2.9] p<0.001
Llama–ARC	-1.3 [-3.3, +0.7] p=0.198	-0.2 [-2.4, +1.9] p=0.851	-7.7 [-9.4, -5.9] p<0.001	-3.9 [-5.7, -2.2] p<0.001	+0.5 [-1.4, +2.5] p=0.654	-1.5 [-3.6, +0.5] p=0.148	-9.0 [-10.7, -7.5] p<0.001	-5.3 [-7.0, -3.6] p<0.001	-0.8 [-2.7, +1.1] p=0.388
Qwen–GSM8K	-1.2 [-2.6, +0.2] p=0.100	-1.8 [-3.3, -0.2] p=0.028	+2.8 [+1.4, +4.3] p<0.001	+0.5 [-0.8, +1.8] p=0.446	+3.4 [+2.1, +4.8] p<0.001	-3.0 [-4.5, -1.4] p<0.001	+1.6 [+0.2, +3.0] p=0.022	-0.7 [-1.9, +0.5] p=0.263	+2.2 [+0.9, +3.5] p=0.001
Qwen–ARC	-0.1 [-0.4, +0.1] p=0.308	-0.2 [-0.5, +0.04] p=0.097	+0.4 [+0.2, +0.6] p<0.001	+0.1 [-0.1, +0.3] p=0.497	+0.7 [+0.4, +0.9] p<0.001	-0.4 [-0.6, -0.1] p=0.002	+0.3 [+0.1, +0.5] p=0.004	-0.1 [-0.3, +0.1] p=0.635	+0.5 [+0.3, +0.8] p<0.001

Table 4: Bridging intervention on seed 0, in percentage points (95% paired bootstrap CIs). Identity and representation Bridging reuse the run-level CO h of the matched flow bridging run. Columns are $\mathrm{LB_fF}$−LF (flow bridging B^{flow}), $\mathrm{LB_rF}$−LF (representation bridging B^{rep}), $\mathrm{LB_aF}$−LF ($\mathrm{LB_aF}$: average of B^{rep} and B^{flow}), and $\mathrm{LB_mF}$−LF ($\mathrm{LB_mF}$: better B^{rep} or B^{flow} per Localize-FT pair). p is two-sided bootstrap for a zero mean difference.

Block	$\mathrm{LB_fF}$−LF	$\mathrm{LB_rF}$−LF	$\mathrm{LB_aF}$−LF	$\mathrm{LB_mF}$−LF
Llama–GSM8K	-0.6 [-1.2, +0.0] p=0.062	+1.9 [+0.9, +3.1] p<0.001	+0.6 [+0.0, +1.4] p=0.047	+2.3 [+1.4, +3.5] p<0.001
Llama–ARC	+0.5 [-0.3, +1.3] p=0.256	-7.0 [-9.3, -4.8] p<0.001	-3.2 [-4.5, -2.1] p<0.001	+1.6 [+0.9, +2.4] p<0.001
Qwen–GSM8K	-2.1 [-3.0, -1.3] p<0.001	+2.5 [+1.4, +3.7] p<0.001	+0.2 [-0.5, +0.9] p=0.572	+2.8 [+1.7, +3.9] p<0.001
Qwen–ARC	-0.3 [-0.6, +0.1] p=0.114	+0.4 [+0.1, +0.8] p=0.025	+0.1 [-0.2, +0.4] p=0.533	+0.7 [+0.3, +1.0] p<0.001

According to Figure 2 and Table 3, LBF is the strongest structural DO once bridging is chosen well, but the choice of rule is first-order. $\mathrm{LB_mF}$ is significantly better than both LF and RF on three blocks; on Llama–ARC neither contrast is significant, and $\mathrm{LB_mF}$ is statistically indistinguishable from the best bar (RF). The two fixed bridging rules do not share this pattern. Relative to LF, $\mathrm{LB_rF}$ is significantly better on three blocks and significantly worse on Llama–ARC, whereas $\mathrm{LB_fF}$ is significantly worse on one block and not significant on the other three. LF and RF do not differ significantly on any block, which agrees with the non-significant localization main effect in Section 4.2. AF ranks last or second-to-last on three blocks, and approaches $\mathrm{LB_mF}$ only on Qwen–ARC.

Figure 4 stratifies the structural DOs by head budget $k \in \{1, 2, 3\}$. The ranking is stable across k, so the structural pattern is not an artifact of a single budget.

To isolate bridging, we run a paired intervention on seed 0. We randomly select B^{flow} as the anchor, so each such run supplies a matched assignment (L, F, h). B^{id} and B^{rep} are then executed under that same assignment, completing a matched triplet that differs only in B.

Table 4 reports the paired contrasts. We find that the intervention results yield nearly the same qualitative pattern as the descriptive contrasts in Table 3. Specifically, $\text{LB}_{\text{m}}\text{F}$ minus LF is positive and significant on all four blocks. Relative to LF, B^{rep} is again significantly better on three blocks and significantly worse only on Llama–ARC, whereas B^{flow} is significantly worse on one block (Qwen-GSM8K) and not significantly different on the other three. The isolated effect of B is therefore asymmetric: a significant loss is confined to a single block and rule, while a well-chosen bridging rule is significantly better in most settings. Bridging thus raises the chance of a gain without a systematic, significant drop relative to LF when the right rule is selected.

4.4 Attribute DO comparison and intervention

Section 4.2 ranked the attribute DOs by η^2: Fine-tuning dominates, Bridging is significant but smaller, and Localization is weak. Here we compare rules within each attribute DO.

Localization. Figure 5 ranks the 14 localization rules. The ranking is block-dependent and weakly separated. No rule leads on every block, and pairwise Kendall τ of the 14-way ranking is near zero (-0.30 to $+0.19$). The spread among rule means is modest, and 95% CIs overlap for most pairs. The leading rule differs across blocks. For example, ablation is last or near-last on three blocks, yet second on Llama–ARC. Averaged across blocks, entropy tracing and SAMD rank highest and ablation lowest.

Fine-tuning. Figure 6 ranks the four fine-tuning rules. The ranking is block-dependent but much more separated. A rule that leads on one dataset may rank last on another. For example, selective head fine-tuning leads on both ARC blocks and is last on both GSM8K blocks. Head output bias is the reverse: last on both ARC blocks, but first on Llama–GSM8K. Selective head LoRA and head output remapping are middling in most rankings. On Qwen–ARC, changing the fine-tuning rule barely changes accuracy. The four rules differ by only about 1.5 pp; on the other three blocks the gap is 5–19 pp, several times the localization spread.

Bridging. Figure 3 splits the runs by bridging rule, with RF as a reference. B^{rep} is the strongest bar on three of four blocks; Llama–ARC is the exception, where it is the weakest. Table 5 isolates $B^{\text{rep}} - B^{\text{flow}}$. On every block the 95% CI excludes zero and $p<0.001$. The contrast is positive on three blocks and reverses on Llama–ARC.

To further establish the contrast, Table 5 estimates $B^{\text{rep}} - B^{\text{flow}}$ as a controlled intervention on seed 0 with paired h: B^{rep} reuses the run-level CO of the matched B^{flow} run. The intervention estimates match the factorial results in sign, and the magnitudes remain close. Once h is held fixed, the gap remains significant on every block and its sign stays block-dependent.

4.5 Setting- and run-level COs

The preceding ANOVA treats each model–dataset pair as a block. We now enter the two setting-level COs as factors. Pooling models within each dataset (Table 8), the model effect is significant on both datasets and largest on GSM8K ($\eta^2 = 0.659$). Pooling datasets within each model (Table 9), the dataset effect is significant on both models and largest on Llama ($\eta^2 = 0.612$). The ranking reverses: Llama leads Qwen by 19.1 pp on GSM8K (87.9% vs. 68.8%) and trails it by 4.6 pp on ARC (66.5% vs. 71.1%).

Each $L \times B \times F$ cell samples h once from Table 1a rather than enumerating the grid, so Tables 2 and 6 report one-way and interaction η^2 among these draws. Learning rate dominates on three blocks ($\eta^2_{lr} = 0.24$–0.37); Qwen–ARC is insensitive. The preferred rate flips: 10^{-4} raises SGD Fine-tuning on Llama–ARC from 65.0% to 74.6%, whereas 10^{-5} is better on both GSM8K blocks. Epochs are significant on the same three blocks; n_L, k, n_B, n_F, and r remain small.

Figure 9 reports mean accuracy by Fine-tuning rule and learning rate. Within a block, selective head fine-tuning, selective head LoRA, and head output remapping usually share the sign of the learning-rate change, but no rule keeps that sign across blocks.

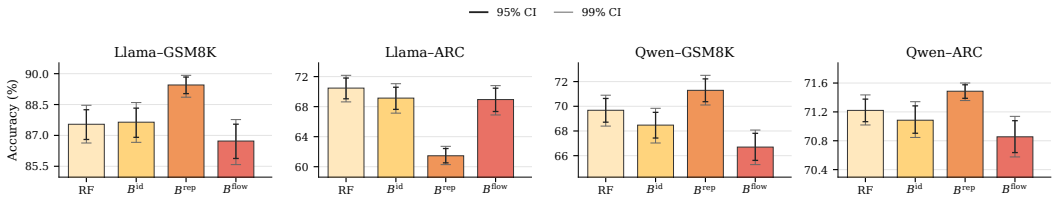

Figure 3: Mean accuracy (%) by bridging rule compared to random baseline. Error bars are percentile bootstrap CIs (thick inner: 95%, thin outer: 99%); y-axis limits are block-specific.

Table 5: Representation minus flow bridging ($B^{\text{rep}} - B^{\text{flow}}$), in percentage points. Brackets give 95% bootstrap confidence intervals. Positive values favor representation bridging. p is two-sided bootstrap for a zero mean difference.

	Llama–GSM8K	Llama–ARC	Qwen–GSM8K	Qwen–ARC
Factorial	+2.7 [+1.8, +3.6] $p<0.001$	-7.5 [-9.3, -5.6] $p<0.001$	+4.6 [+3.2, +6.0] $p<0.001$	+0.6 [+0.4, +0.9] $p<0.001$
Intervention	+2.5 [+1.4, +3.6] $p<0.001$	-7.4 [-10.1, -5.0] $p<0.001$	+4.6 [+3.1, +6.2] $p<0.001$	+0.7 [+0.4, +1.1] $p<0.001$

4.6 Is there one recipe that beats the baseline and transfers?

The factor-level results above describe averages, not the question a practitioner asks: whether a single $L \times B \times F$ configuration beats Random–FT and remains strong across settings. A winner exists within each block: the best seed-averaged combination exceeds block-wise RF by $+2.5$ to $+10.5$ pp. However, those winners are not the same recipe. Llama–GSM8K prefers logit attribution localization with flow Bridging and selective head LoRA (92.3% vs. 87.5%); Llama–ARC prefers ALPS with flow Bridging and selective head fine-tuning (81.0% vs. 70.5%). Among combinations observed on all four blocks, none stays in the top 40. The only configuration that beats RF everywhere, path patching with flow Bridging and head output remapping, gains at most $+5.6$ pp. There is therefore no single recipe that is both clearly better than the baseline and transferable.

4.7 Generalization: Do the conclusions hold on a new, related dataset?

We rescore the same adapted models on related held-out datasets without further training. To match the in-domain evaluation size, we randomly sample 200 examples from SVAMP for GSM8K-trained models and from OpenBookQA for ARC-Challenge-trained models. $\text{LB}_\text{m}\text{F}$ retains the bridging rule selected for each localization–fine-tuning pair on the original evaluation set, with no reselection during generalization evaluation.

Structural rankings broadly persist (Figures 2 and 11; Spearman $\rho = 0.85$–0.96). $\text{LB}_\text{m}\text{F}$ ranks first on Qwen–GSM8K and second elsewhere, with no significant gap to the top-ranked alternative in two of the three remaining blocks. Localize–FT and Random–FT again show no significant difference (Table 10). A target-set oracle that reselects bridging using generalization accuracy yields similar rankings (Figure 12, Spearman $\rho = 0.85$–1).

The B^{rep}–B^{flow} comparison preserves the in-domain pattern in three model–dataset blocks (Figure 13; Table 11). B^{rep} performs significantly better on both GSM8K-to-SVAMP transfers and significantly worse for ARC-trained Llama models on OpenBookQA. For ARC-trained Qwen models on OpenBookQA, the two rules show no significant difference.

5 Conclusion

We formulate an Evaluatology-based SCDS for design attribution in head-level adaptation, separating designed objects from confounding objects and combining factorial analysis with matched interventions. The experiments support this attribution view: adaptation gains are tied to interacting design choices and to the conditions under which their effects are measured.

AI USE STATEMENT

In this work, we used generative AI tools for assisting with translation, polishing writing, commenting on drafts of the paper, including the results, and literature retrieval. We have not used generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing ingredients for or writing proofs, proposing or refining hypotheses, designing research methodology or experiments, implementing methods, or cleaning or reformatting datasets. Supporting qualitative and thematic data analysis is not applicable to this work. We have reviewed all AI-assisted work: comments suggested with AI assistance were checked against the experimental results by the authors, and literature retrieved or identified with AI assistance was read and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Hao Chen, Haoze Li, Zhiqing Xiao, Lirong Gao, Qi Zhang, Xiaomeng Hu, Ningtao Wang, Xing Fu, and Junbo Zhao. ALPS: Attention localization and pruning strategy for efficient adaptation of large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria, July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-acl.612/.

Wei Chen, Zhen Huang, Liang Xie, Binbin Lin, Houqiang Li, Le Lu, Xinmei Tian, Deng Cai, Yonggang Zhang, Wenxiao Wang, Xu Shen, and Jieping Ye. From yes-men to truth-tellers: Addressing sycophancy in large language models with pinpoint tuning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=d2vONO90Rw.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Karl Cobbe, Vineet Kosaraju, Mo Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *ArXiv*, abs/2110.14168, 2021. URL https://api.semanticscholar.org/CorpusID:239998651.

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.

Anchun Gui and Han Xiao. Hifi: High-information attention heads hold for parameter-efficient model adaptation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8521–8537, 2023.

Peter Hase, Mohit Bansal, Been Kim, and Asma Ghandeharioun. Does localization inform editing? surprising differences in causality-based localization vs. knowledge editing in language models. *Advances in Neural Information Processing Systems*, 36:17643–17668, 2023.

Wen Lai, Alexander Fraser, and Ivan Titov. Joint localization and activation editing for low-resource fine-tuning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 32206–32227. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/lai25a.html.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. *Advances in neural information processing systems*, 36:41451–41530, 2023.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372, 2022.

Paul Michel, Omer Levy, and Graham Neubig. Are sixteen heads really better than one? *Advances in neural information processing systems*, 32, 2019.

Yaniv Nikankin, Anja Reusch, Aaron Mueller, and Yonatan Belinkov. Arithmetic without algorithms: Language models solve math with a bag of heuristics. In *International Conference on Learning Representations*, volume 2025, pp. 55939–55965, 2025.

Zitong Shi, Frank Wan, Haixin Wang, Ruoyan Li, Zijie Huang, Wanjia Zhao, Yijia Xiao, Xiao Luo, Carl Yang, Yizhou Sun, et al. Don't forget the enjoin: Focallora for instruction hierarchical alignment in large language models. *Advances in Neural Information Processing Systems*, 38: 8277–8299, 2026.

Jingtong Su, Julia Kempe, and Karen Ullrich. From concepts to components: Concept-agnostic attention module discovery in transformers. In *International Conference on Learning Representations*, volume 2026, pp. 25595–25627, 2026.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 5797–5808, 2019.

Xuehao Wang, Liyuan Wang, Binghuai Lin, and Yu Zhang. Headmap: Locating and enhancing knowledge circuits in llms. In *International Conference on Learning Representations*, volume 2025, pp. 33509–33526, 2025.

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

Fangcong Yin, Xi Ye, and Greg Durrett. Lofit: Localized fine-tuning on llm representations. *Advances in Neural Information Processing Systems*, 37:9474–9506, 2024.

Fengming Yu, Qingyu Meng, Haiwei Pan, and Kejia Zhang. Efficient mathematical reasoning models via dynamic pruning and knowledge distillation. *arXiv preprint arXiv:2511.17577*, 2025.

Jianfeng Zhan, Lei Wang, Wanling Gao, Hongxiao Li, Chenxi Wang, Yunyou Huang, Yatao Li, Zhengxin Yang, Guoxin Kang, Chunjie Luo, et al. Evaluatology: The science and engineering of evaluation. *BenchCouncil Transactions on Benchmarks, Standards and Evaluations*, 4(1):100162, 2024.

Jianfeng Zhan, Lei Wang, Wanling Gao, Chenxi Wang, Hongxiao Li, Fanda Fan, and Guoxin Kang. *Evaluatology: The Science of Uncovering the Effects*. 11 2025. ISBN 978-988-71596-8-1. doi: 10.66834/fr30pm74.

Tianhe Zhang, Yonghong Deng, Ping Jian, Zhen Yang, Boyang Wang, and Xinyue Zhang. Where cot reasoning commits: Entropy traces identify interpretable attention heads. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2777–2795, 2026.

Wei Zhang, Chaoqun Wan, Yonggang Zhang, Yiu-ming Cheung, Xinmei Tian, Xu Shen, and Jieping Ye. Interpreting and improving large language models in arithmetic calculation. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, Kun Wang, Yang Liu, Junfeng Fang, and Yongbin Li. On the role of attention heads in large language model safety. In *International Conference on Learning Representations*, volume 2025, pp. 84042–84071, 2025.

A APPENDIX

A.1 STRUCTURAL DO

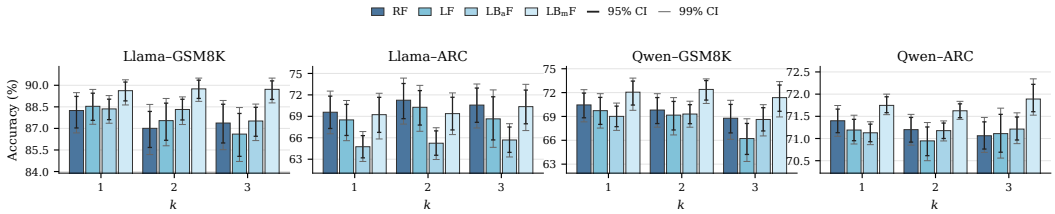

Figure 4: Path mean accuracy (%) by heads-per-layer $k \in \{1, 2, 3\}$ (realized budget $k \times n_{\text{layers}}$). FT-only with random heads (RF), Localize–FT (LF), $\text{LB}_\text{a}\text{F}$ (average of both Bridging rules), and $\text{LB}_\text{m}\text{F}$ (better Bridging rule per Localize–FT pair). Error bars are percentile bootstrap CIs (thick inner: 95%, thin outer: 99%); y-axis limits are block-specific.

A.2 ATTRIBUTE DO

Figure 7 shows how each Localization rule responds to Bridging. On both GSM8K blocks and Qwen–ARC, most rows are darkest under B^{rep} and lightest under B^{flow}; Llama–ARC inverts, with B^{rep} lightening every row. The mean span across the three rules is 2.8 pp on Llama–GSM8K, 8.9 pp on Llama–ARC, 4.9 pp on Qwen–GSM8K, and 0.9 pp on Qwen–ARC; the largest single-rule swing is logit lens on Llama–ARC (73.3% under B^{id} vs. 59.7% under B^{rep}).

Figure 8 shows how each Fine-tuning rule responds to Bridging. The ranking depends on the Bridging rule. Selective head fine-tuning is weak under B^{id} and B^{flow} on Llama–GSM8K but recovers under B^{rep}; Llama–ARC is the reverse, where B^{rep} lowers every method and selective head fine-tuning with B^{id} or B^{flow} is strongest. Head output bias is stable on both Llama blocks—high on GSM8K, poor on ARC—but swings 9.6 pp on Qwen–GSM8K; head output remapping is best under every rule there. Qwen–ARC is nearly flat (69.4–71.9%). Fine-tuning still moves accuracy more than Bridging on average, but the ranking is not independent of the rule.

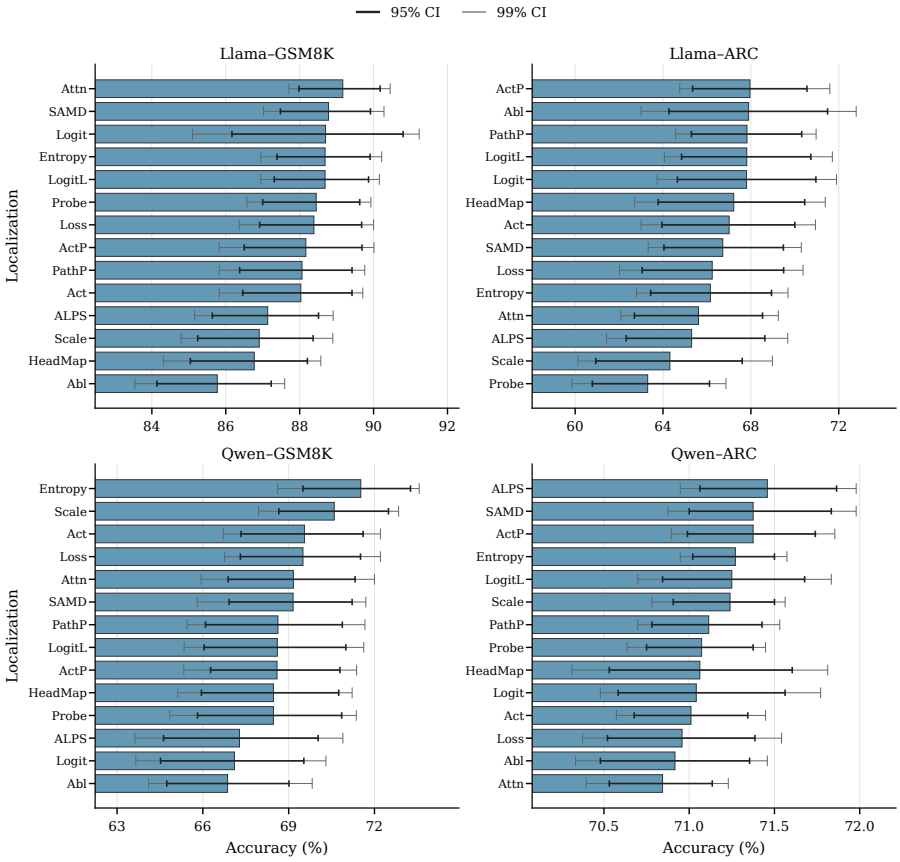

Figure 5: Mean accuracy (%) by Localization rule, sorted within each block. PathP = path patching, ActP = activation patching, Probe = linear probing, Entropy = entropy tracing, LogitL = logit lens, Attn = attention scoring, Act = activation scoring, Logit = logit attribution, Loss = loss-gradient importance, SAMD = Scalable Attention Module Discovery, Scale = scale-based selection (LoFiT), Abl = ablation, HeadMap = layer-conditioned ablation, ALPS = parameter-difference localization. Error bars are percentile bootstrap CIs (thick inner: 95%, thin outer: 99%); x-axis limits are block-specific.

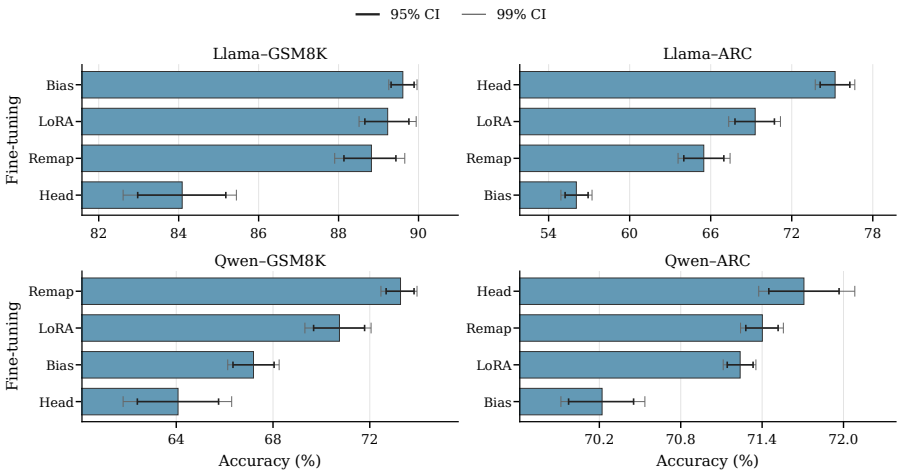

Figure 6: Mean accuracy (%) by Fine-tuning rule, sorted within each block. Head = selective head fine-tuning, LoRA = selective head LoRA, Bias = head output bias, Remap = head output remapping. Error bars are percentile bootstrap CIs (thick inner: 95%, thin outer: 99%); x-axis limits are block-specific.

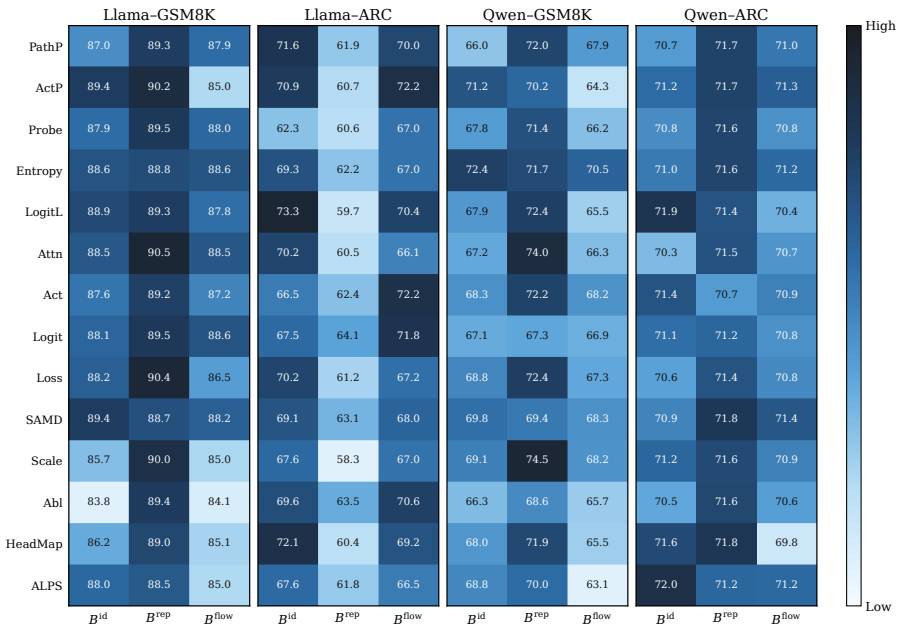

Figure 7: Cell-mean accuracy (%) for Localization × Bridging; colour scales are block-specific. Rule abbreviations match Figure 5.

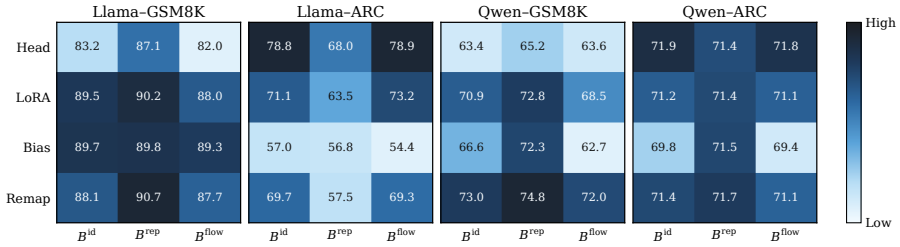

Figure 8: Cell-mean accuracy (%) for Fine-tuning × Bridging; colour scales are block-specific. Rule abbreviations match Figure 6.

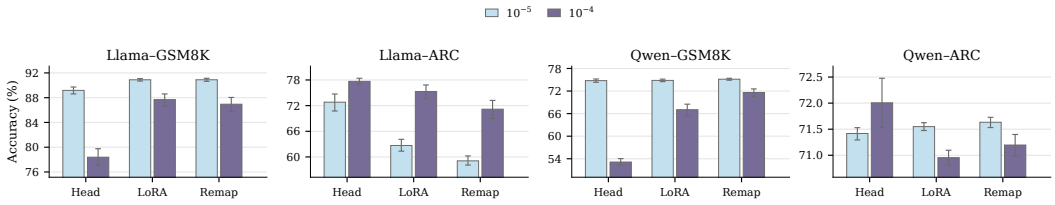

Figure 9: Mean accuracy (%) by SGD Fine-tuning rule and learning rate on guided runs (Head = selective head fine-tuning, LoRA = selective head LoRA, Remap = head output remapping). Error bars are 95% bootstrap CIs; y-axis limits are block-specific.

A.3 MAIN AND INTERACTION EFFECTS DETAILS

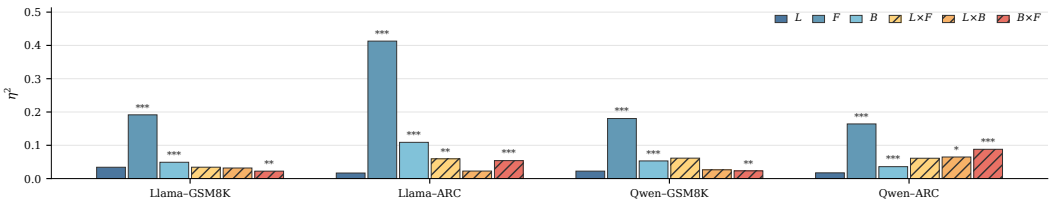

Figure 10: One-way and interaction η^2 of Localization, Bridging, and Fine-tuning on guided runs (stars mark ANOVA $p<0.05$, $p<0.01$, $p<0.001$).

Table 6: Two-way interaction η^2 from a cell-means ANOVA, with p-values. $L \times B$, $L \times F$, $B \times F$, and run-level-CO×attribute-DO terms use Localize–FT / Localize–Bridge–FT (guided) runs. $n_L \times L$ drops linear probing / scale-based selection; $n_B \times B$ is non-identity Bridging; $lr \times F$ and $e \times F$ are SGD Fine-tuning; $r \times F$ is selective head LoRA/head output remapping. $S \times n_F$, $S \times lr$, $S \times e$, and $S \times r$ use all four structural DOs (every structural DO has Fine-tuning; lr/e skip head output bias, r is selective head LoRA/head output remapping).

Block	$\eta^2_{S\times n_F}$	$\eta^2_{S\times \mathrm{lr}}$	$\eta^2_{S\times e}$	$\eta^2_{S\times r}$	$\eta^2_{L\times B}$	$\eta^2_{L\times F}$	$\eta^2_{B\times F}$	$\eta^2_{n_L\times L}$	$\eta^2_{n_B\times B}$	$\eta^2_{n_F\times F}$	$\eta^2_{\mathrm{lr}\times F}$	$\eta^2_{e\times F}$	$\eta^2_{r\times F}$	$\eta^2_{k\times L}$	$\eta^2_{k\times B}$	$\eta^2_{k\times F}$
	0.036	**0.105**	0.000	0.007	0.032	0.034	0.022	0.038	0.000	0.014	**0.110**	0.004	0.000	0.051	0.009	0.025
Llama–GSM8K	$p<0.001$	$p<0.001$	$p=1$	$p=0.320$	$p=0.715$	$p=0.923$	$p=0.004$	$p=0.408$	$p=1$	$p=0.093$	$p<0.001$	$p=0.299$	$p=0.845$	$p=0.144$	$p=0.194$	$p=0.003$
	0.015	**0.069**	0.004	0.003	0.023	0.059	0.054	0.041	0.000	0.014	0.045	0.000	0.001	0.049	0.005	0.013
Llama–ARC	$p=0.018$	$p<0.001$	$p=0.414$	$p=0.685$	$p=0.913$	$p=0.002$	$p<0.001$	$p=0.376$	$p=0.922$	$p=0.010$	$p<0.001$	$p=1$	$p=0.577$	$p=0.157$	$p=0.432$	$p=0.024$
	0.006	0.001	0.000	0.002	0.026	0.061	0.023	0.039	0.004	0.028	**0.208**	0.000	0.002	0.039	0.009	0.007
Qwen–GSM8K	$p=0.360$	$p=0.636$	$p=1$	$p=0.750$	$p=0.854$	$p=0.097$	$p=0.002$	$p=0.441$	$p=0.425$	$p<0.001$	$p<0.001$	$p=1$	$p=0.416$	$p=0.463$	$p=0.163$	$p=0.448$
	0.011	0.007	0.003	0.029	0.065	0.061	**0.088**	0.017	0.004	0.055	0.051	0.021	0.004	0.035	0.002	0.011
Qwen–ARC	$p=0.078$	$p=0.159$	$p=0.527$	$p=0.002$	$p=0.010$	$p=0.126$	$p<0.001$	$p=0.987$	$p=0.392$	$p<0.001$	$p<0.001$	$p=0.004$	$p=0.252$	$p=0.625$	$p=0.837$	$p=0.172$

Table 7: One-way η^2 of Localization, Bridging, Fine-tuning, and run-level COs on Localize–FT and Localize–Bridge–FT runs, with ANOVA p-values. Each row is one block of the $L \times B \times F$ design (4 seeds $\times 14 \times 3 \times 4$). n_L, n_B, and n_F are Localization, Bridging, and Fine-tuning sample sizes (n_B only if $B \neq \mathrm{id}$; n_L drops linear probing / scale-based selection, which use $\{300, 600, 900\}$ rather than $\{64, 128, 192\}$); k is heads per layer (k and n_F use all guided runs); lr and e (epochs) are computed on SGD Fine-tuning (selective head fine-tuning, selective head LoRA, head output remapping); r is the selective head LoRA/head output remapping rank.

	Attribute DO			run-level CO						
Block	η^2_L	η^2_F	η^2_B	$\eta^2_{n_L}$	η^2_k	$\eta^2_{n_B}$	$\eta^2_{n_F}$	η^2_{lr}	η^2_e	η^2_r
	0.034	0.191	0.049	0.003	0.009	0.003	0.008	0.254	0.021	0.002
Llama–GSM8K	$p=0.058$	$p<0.001$	$p<0.001$	$p=0.458$	$p=0.051$	$p=0.519$	$p=0.075$	$p<0.001$	$p=0.001$	$p=0.443$
	0.017	0.413	0.109	0.004	0.001	0.001	0.009	0.238	0.017	0.009
Llama–ARC	$p=0.595$	$p<0.001$	$p<0.001$	$p=0.326$	$p=0.664$	$p=0.848$	$p=0.054$	$p<0.001$	$p=0.003$	$p=0.086$
	0.022	0.180	0.053	0.005	0.006	0.009	0.005	0.369	0.031	0.013
Qwen–GSM8K	$p=0.304$	$p<0.001$	$p<0.001$	$p=0.260$	$p=0.113$	$p=0.140$	$p=0.167$	$p<0.001$	$p<0.001$	$p=0.034$
	0.017	0.164	0.036	0.002	0.001	0.019	0.007	0.005	0.000	0.001
Qwen–ARC	$p=0.564$	$p<0.001$	$p<0.001$	$p=0.590$	$p=0.844$	$p=0.015$	$p=0.100$	$p=0.125$	$p=0.635$	$p=0.492$

Table 8: One-way η^2 of model, Localization, Bridging, Fine-tuning, and run-level COs on Localize–FT and Localize–Bridge–FT runs, pooled across the two models within each dataset, with ANOVA p-values. Each column is a separate one-way ANOVA. n_L drops linear probing / scale-based selection; n_B only if $B \neq$ id; lr and e are computed on SGD Fine-tuning (selective head fine-tuning, selective head LoRA, head output remapping); r is the selective head LoRA/head output remapping rank.

Dataset	η^2_{model}	η^2_L	η^2_B	η^2_F	$\eta^2_{n_L}$	η^2_k	$\eta^2_{n_B}$	$\eta^2_{n_F}$	η^2_{lr}	η^2_e	η^2_r
GSM8K	0.659 $p<0.001$	0.018 $p=0.030$	0.017 $p<0.001$	0.052 $p<0.001$	0.002 $p=0.342$	0.003 $p=0.160$	0.002 $p=0.432$	0.002 $p=0.265$	0.131 $p<0.001$	0.010 $p=0.002$	0.001 $p=0.382$
ARC	0.082 $p<0.001$	0.008 $p=0.681$	0.043 $p<0.001$	0.216 $p<0.001$	0.002 $p=0.404$	0.001 $p=0.705$	0.000 $p=0.839$	0.005 $p=0.039$	0.112 $p<0.001$	0.009 $p=0.003$	0.004 $p=0.109$

Table 9: One-way η^2 of dataset, Localization, Bridging, Fine-tuning, and run-level COs on Localize–FT and Localize–Bridge–FT runs, pooled across the two datasets within each model, with ANOVA p-values. Each column is a separate one-way ANOVA. n_L drops linear probing / scale-based selection; n_B only if $B \neq$ id; lr and e are computed on SGD Fine-tuning (selective head fine-tuning, selective head LoRA, head output remapping); r is the selective head LoRA/head output remapping rank.

Model	η^2_{dataset}	η^2_L	η^2_B	η^2_F	$\eta^2_{n_L}$	η^2_k	$\eta^2_{n_B}$	$\eta^2_{n_F}$	η^2_{lr}	η^2_e	η^2_r
Llama	0.612 $p<0.001$	0.008 $p=0.662$	0.009 $p=0.002$	0.043 $p<0.001$	0.001 $p=0.683$	0.000 $p=0.804$	0.000 $p=0.990$	0.000 $p=0.731$	0.007 $p=0.011$	0.001 $p=0.456$	0.001 $p=0.331$
Qwen	0.037 $p<0.001$	0.011 $p=0.312$	0.032 $p<0.001$	0.087 $p<0.001$	0.003 $p=0.215$	0.003 $p=0.150$	0.006 $p=0.066$	0.002 $p=0.328$	0.182 $p<0.001$	0.014 $p<0.001$	0.007 $p=0.029$

A.4 GENERALIZATION

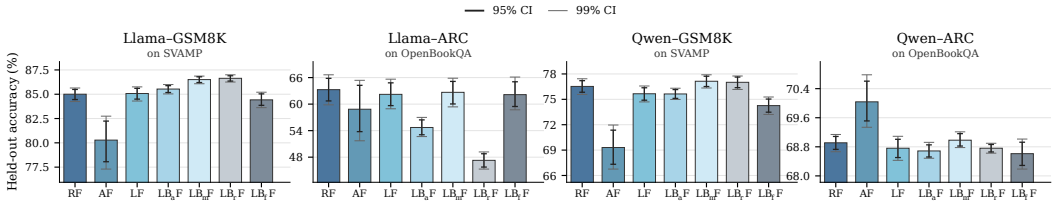

Figure 11: Generalization accuracy (%). Models fine-tuned on GSM8K are scored on SVAMP; models fine-tuned on ARC-Challenge are scored on OpenBookQA. Panel titles name the trained model–dataset block. Bars are Random–FT (RF), All–FT (AF), Localize–FT (LF), and Localize–Bridge–FT: LB$_a$F (average of B^{rep} and B^{flow}) and LB$_m$F retains the bridging rule selected for each localize–FT pair on the original evaluation set (Figure 2), keeping this choice fixed throughout generalization evaluation. Error bars are percentile bootstrap CIs (thick inner: 95%, thin outer: 99%); y-axis limits are block-specific.

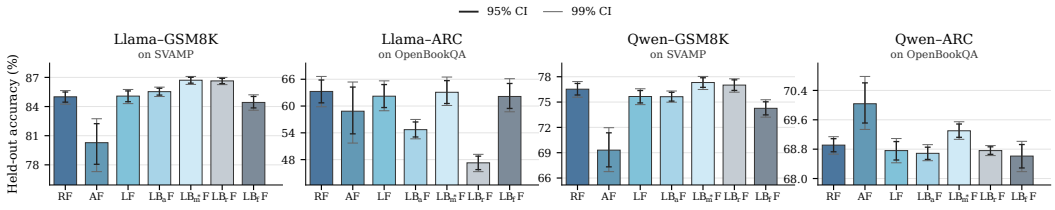

Figure 12: Generalization accuracy (%). Models fine-tuned on GSM8K are scored on SVAMP; models fine-tuned on ARC-Challenge are scored on OpenBookQA. Panel titles name the trained model–dataset block. Bars match Figure 11, with LB$_{m^*}$F in place of LB$_m$F. LB$_{m^*}$F is the better Bridging rule per Localize–FT pair, selected on the new dataset. Error bars are percentile bootstrap CIs (thick inner: 95%, thin outer: 99%); y-axis limits are block-specific.

Table 10: Generalization accuracy differences in percentage points (95% bootstrap CIs). Columns are LF−RF, LB$_a$F−LF, LB$_m$F−LF, LB$_{m^*}$F−LF, LB$_a$F−RF, LB$_m$F−RF, and LB$_{m^*}$F−RF. Positive values favor the first term. LB$_m$F uses, for each Localize–FT pair, the Bridging rule that was better in the main experiment; the difference is that rule's generalization accuracy. LB$_{m^*}$F is the better Bridging rule per Localize–FT pair, selected on the generalization accuracy. p is two-sided bootstrap for a zero mean difference (LB$_m$F and LB$_{m^*}$F columns unadjusted for rule selection).

Block	LF−RF	LB$_a$F−LF	LB$_m$F−LF	LB$_{m^*}$F−LF	LB$_a$F−RF	LB$_m$F−RF	LB$_{m^*}$F−RF
Llama–GSM8K	+0.1	+0.5	+1.4	+1.6	+0.5	+1.5	+1.7
on SVAMP	[-0.7, +0.8]	[-0.2, +1.1]	[+0.8, +2.1]	[+1.0, +2.3]	[-0.1, +1.1]	[+0.9, +2.1]	[+1.1, +2.3]
	p=0.855	p=0.170	p<0.001	p<0.001	p=0.103	p<0.001	p<0.001
Llama–ARC	-1.1	-7.5	+0.5	+0.9	-8.6	-0.6	-0.2
on OpenBookQA	[-4.6, +2.6]	[-10.6, -4.6]	[-3.3, +4.3]	[-2.7, +4.5]	[-11.6, -5.4]	[-4.3, +3.0]	[-3.7, +3.5]
	p=0.550	p<0.001	p=0.778	p=0.625	p<0.001	p=0.773	p=0.898
Qwen–GSM8K	-0.9	-0.02	+1.5	+1.7	-0.9	+0.6	+0.8
on SVAMP	[-1.8, +0.1]	[-0.8, +0.9]	[+0.5, +2.3]	[+0.7, +2.6]	[-1.8, -0.04]	[-0.3, +1.5]	[-0.1, +1.7]
	p=0.089	p=0.952	p=0.001	p<0.001	p=0.042	p=0.197	p=0.072
Qwen–ARC	-0.1	-0.1	+0.2	+0.5	-0.2	+0.1	+0.4
on OpenBookQA	[-0.4, +0.2]	[-0.4, +0.2]	[-0.1, +0.5]	[+0.2, +0.9]	[-0.5, +0.03]	[-0.2, +0.3]	[+0.2, +0.6]
	p=0.353	p=0.623	p=0.142	p=0.001	p=0.074	p=0.586	p=0.001

Table 11: Generalization Bridging-rule difference in percentage points (95% bootstrap CIs): representation Bridging minus flow Bridging. Positive values favor representation Bridging. p is two-sided bootstrap for a zero mean difference.

Llama–GSM8K on SVAMP	Llama–ARC on OpenBookQA	Qwen–GSM8K on SVAMP	Qwen–ARC on OpenBookQA
+2.2 [+1.6, +2.9] $p<0.001$	-14.9 [-17.9, -11.8] $p<0.001$	+2.8 [+1.8, +3.7] $p<0.001$	+0.1 [-0.2, +0.5] $p=0.364$

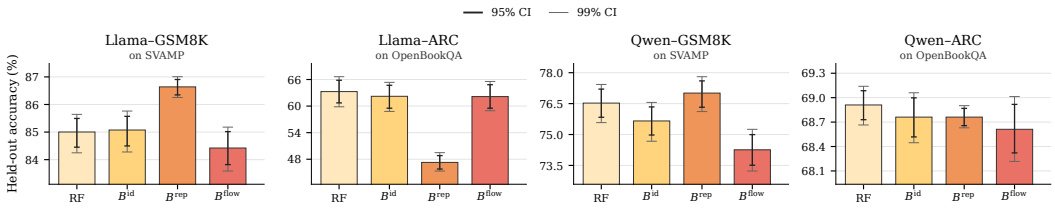

Figure 13: Generalization accuracy (%) by Bridging rule. Models fine-tuned on GSM8K are scored on SVAMP; models fine-tuned on ARC-Challenge are scored on OpenBookQA. Random–FT (RF), identity Bridging B^{id} (Localize–FT), representation Bridging B^{rep}, and flow Bridging B^{flow}. Error bars are percentile bootstrap CIs (thick inner: 95%, thin outer: 99%); y-axis limits are block-specific.

A.5 COMPUTE COST

Each run occupies one NVIDIA A800 80GB GPU. Table 12 sums that GPU time. The $L \times B \times F$ row is Localize–FT and Localize–Bridge–FT: four stratified sampling seeds, with 168 plans per seed in each block. Random–FT and All–FT are sampled separately. When the same Localization or Bridging result is reused, its time is counted once, so the totals are time actually spent.

Table 12: GPU-hours (single A800 wall time) by block.

Sample	Llama–GSM8K	Llama–ARC	Qwen–GSM8K	Qwen–ARC	Total
$L \times B \times F$	456.6	159.2	341.9	57.8	1015.5
Random–FT	58.0	8.0	52.9	7.1	126.0
All–FT	13.4	6.9	11.0	4.9	36.2
Total	528.0	174.1	405.8	69.8	1177.7

Table 13 reports the mean time of each rule, in minutes. Localization and Bridging means use runs on which that rule is executed; identity Bridging has no Bridging time and is omitted. Fine-tuning means use the Localize–FT and Localize–Bridge–FT runs. Evaluation time is excluded.

Table 13: Mean time (minutes) of the 14 Localization rules, the two non-identity Bridging rules, and the four Fine-tuning rules. PathP = path patching, ActP = activation patching, Abl = ablation, HeadMap = layer-conditioned ablation, Probe = linear probing, Entropy = entropy tracing, Attn = attention scoring, Act = activation scoring, LogitL = logit lens, Logit = logit attribution, Loss = loss-gradient importance, SAMD = Scalable Attention Module Discovery, Scale = scale-based selection, ALPS = parameter-difference localization. Head = selective head fine-tuning, LoRA = selective head LoRA, Bias = head output bias, Remap = head output remapping.

Attribute DO	Rule	Llama–GSM8K	Llama–ARC	Qwen–GSM8K	Qwen–ARC
L	PathP	82.7	52.8	22.2	9.6
	ActP	80.9	73.0	19.3	16.5
	Abl	46.4	46.8	7.5	7.9
	HeadMap	40.5	45.5	8.9	8.7
	Probe	1.3	1.4	0.8	0.8
	Entropy	26.9	1.3	20.5	0.8
	Attn	26.4	1.1	19.9	0.7
	Act	24.4	1.1	18.5	0.7
	LogitL	33.0	1.8	66.8	0.9
	Logit	27.8	1.7	49.9	1.2
	SAMD	0.6	0.7	0.3	0.3
	Loss	0.8	0.8	0.5	0.5
	Scale	4.2	4.1	3.8	3.5
	ALPS	0.9	0.6	0.1	0.1
B	B^{rep}	4.3	4.4	0.9	0.9
	B^{flow}	27.0	1.2	20.4	0.7
F	Head	1.1	1.0	1.0	0.9
	LoRA	1.2	1.2	1.2	1.1
	Bias	28.6	4.0	26.0	5.7
	Remap	1.1	1.1	1.1	0.9

A.6 PILOT EXPERIMENTS

Table 14: Spearman ρ of the full ten-factor order after each dropped sampling plan. Row k removes plan k. ρ correlates the order of all ten factors on *all* with their order after plan k is removed.

Block	Dropped	p-order ρ	η^2-order ρ
Llama–GSM8K	0	.90	.99
Llama–GSM8K	1	.83	.94
Llama–GSM8K	2	.99	1
Llama–GSM8K	3	.99	.96
Llama–ARC	0	.96	.96
Llama–ARC	1	.98	.99
Llama–ARC	2	.89	.94
Llama–ARC	3	.99	.95
Qwen–GSM8K	0	.96	.96
Qwen–GSM8K	1	.89	.95
Qwen–GSM8K	2	.92	.92
Qwen–GSM8K	3	.94	.96
Qwen–ARC	0	.92	.99
Qwen–ARC	1	.88	.95
Qwen–ARC	2	.84	.96
Qwen–ARC	3	.98	.99

A.7 INTRODUCTION COMPARISONS DATA

Table 15 reports the cell means behind the examples in Section 1. The ablation rows on Llama–ARC hold bridging at identity and differ only in the fine-tuning rule. For the bridging examples, direct update is identity bridging; the other row is representation bridging on Qwen–GSM8K and flow bridging on Llama–ARC.

Table 15: Mean accuracy of the examples in Section 1, and the difference from Random–FT. Qwen–GSM8K is Qwen2.5-1.5B-Instruct with GSM8K; Llama–ARC is Llama-3.1-8B-Instruct with ARC-Challenge. Random–FT is the mean of all successful random-head runs in the block (n=240), pooled across the four fine-tuning rules. Δ is the cell mean minus that Random–FT mean, in percentage points.

Block	Configuration	Bridging	Seeds (%)	Mean (%)	Δ (pp)
Qwen–GSM8K	Random–FT	—	—	69.68	—
	path patching, head output bias	B^{id}	61.5, 60.0, 64.0, 69.5	63.75	-5.93
	path patching, head output bias	B^{rep}	75.0, 73.5, 73.5, 76.0	74.50	$+4.82$
Llama–ARC	Random–FT	—	—	70.47	—
	ablation, selective head fine-tuning	B^{id}	82.0, 81.0, 79.5, 81.0	80.88	$+10.41$
	ablation, head output bias	B^{id}	43.0, 43.0, 44.5, 50.5	45.25	-25.22
	logit attribution, head output remapping	B^{id}	78.5, 80.5, 53.5, 62.0	68.63	-1.84
	logit attribution, head output remapping	B^{flow}	81.0, 80.5, 78.0, 79.5	79.75	$+9.28$