VERNIER: DIAGNOSING AND MITIGATING LEXICAL SENSITIVITY IN CAUSAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Lexical renaming exposes a gap between specifying a causal problem and eliciting a stable answer from a language model. We study this gap through Vernier, a paired-view LoRA update that combines supervision on original and placeholder prompts with answer-token consistency. The analysis separates three outcomes: improved task performance, agreement across lexical views, and the representation measurements used to explain them. Recorded multi-model evaluations show substantial gains in both views for several Qwen and Llama models, with stronger transfer on CRASS than on e-CARE. An independently traceable 900-question evaluation confirms large paired-view accuracy gains, while also showing that such gains need not reduce the absolute lexical gap. Probes measure the transfer of lexical-content decoders, and donor-controlled patching demonstrates strong control of answer identity at an intermediate layer. Together, the results locate a useful target for adaptation at the answer interface and establish why task improvement, lexical consistency, and mechanistic explanation must be evaluated separately. Code is available at https://anonymous.4open.science/r/Vernier.

1 INTRODUCTION

Instruction-tuned language models can give different answers to the same causal-reasoning question when its variable names change. Replacing surface forms such as *smoking*, *husband*, and *alarm clock* with typed placeholders X_1, X_2, X_3 in a structurally identical prompt produces a different answer distribution, even though the structural causal model, the joint distribution, the query, and the gold answer are all preserved. Yu & Zhou (2026) report this lexical gap across model families and find that direct prompting, chain-of-thought prompting, and scaffolded prompting leave it visible. Related 2025–2026 evaluations report similar brittleness under meaning-preserving rewrites, context shifts, and anti-semantic controls in causal or mathematical reasoning settings (Hao et al., 2025; Lee et al., 2025; Wang et al., 2026). The failure is therefore not just that the model needs a better prompt.

The scientific question is which aspects of a lexical failure a weight update repairs. One possibility is reduced accessibility of task-relevant content under the placeholder view. Another is a change in how available content is read out as an answer. These possibilities connect lexical sensitivity to broader work on spurious-signal reliance and prompt sensitivity (Webson & Pavlick, 2022; Sclar et al., 2024). They need not be mutually exclusive: altered representations, answer preferences, and decision boundaries can all affect the observed gap. We therefore examine performance and representation measurements jointly, while distinguishing the evidence supplied by each.

We study this question with a targeted weight update. Vernier trains a LoRA adapter on paired lexical views (Hu et al., 2022): both views receive answer supervision, while a symmetric KL term encourages agreement over candidate answer tokens (Xie et al., 2020; Liang et al., 2021). This provides an explicit intervention at the training level and a compact interface for comparing the two views. The recorded reference configurations suggest that paired supervision is an important contributor to task performance. Comparisons involving the KL term are interpreted separately from that broader gain. Recent benchmark work also motivates looking beyond prompt-side variation when evaluating stable reasoning behaviour (Sclar et al., 2024; Chi et al., 2024).

The analysis follows a sequence of increasingly specific questions. Do both views improve, or does a small gap hide regression on the original view? Does a decoder trained on one view transfer to the other? Does patching require the matched item's representation, or can a donor with the same

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.

predicted answer produce similar behaviour? We combine behavioural evaluation with probes and logit-lens diagnostics (Belrose et al., 2023), activation patching (Zhang & Nanda, 2024), capability checks (Hendrycks et al., 2021), and model-family comparisons. This design makes agreement informative only when accompanied by correctness, and intervention outcomes informative only when accompanied by suitable donor controls. Our **contributions** are as follows:

- **We separate lexical task improvement from evidence of invariance.** Paired-view accuracy, signed gaps, and answer agreement reveal distinct outcomes of adaptation. A traceable 900-question comparison anchors the behavioural result, while the broader result archive describes variation across models and tasks.
- **We study paired lexical training at the answer interface.** Vernier combines full-vocabulary supervision on both views with consistency over candidate answer tokens. The reference configurations compare this intervention with original-view training, dropout consistency, and paired supervision without KL.
- **We test the specificity of representation-based explanations.** Lexical-content probes and donor-controlled patching distinguish decoder transfer from information use, and answer-state transfer from evidence requiring an item-matched donor. Scale, capability, and transfer analyses characterize where the observed gains do and do not extend.

2 Related Work

2.1 Lexical robustness and causal reasoning in LLMs

Yu & Zhou (2026) study lexical anonymization in causal QA through SCM-based renaming on CLadder (Jin et al., 2023) and natural-language perturbations on CRASS (Frohberg & Binder, 2022) and e-CARE (Du et al., 2022). SCM renaming preserves the specified causal query, whereas natural-language replacement requires a separate semantic-preservation assumption: retaining the answer key does not guarantee that meaning is unchanged. Related studies document reliance on syntactic heuristics (McCoy et al., 2019) and sensitivity to instruction wording and prompt formatting (Webson & Pavlick, 2022; Sclar et al., 2024). Building on lexical robustness evaluation, we examine paired training as an intervention and combine behavioural measurements with representation diagnostics to investigate how responses to the two views change.

2.2 Invariance, consistency training, and counterfactual augmentation

Consistency training encourages agreement across related inputs (Xie et al., 2020). R-Drop (Liang et al., 2021) applies symmetric KL to dropout-sampled predictions for the same input, while SimCSE (Gao et al., 2021) uses dropout views for contrastive representation learning. Vernier instead pairs an original prompt with its lexical perturbation and applies symmetric KL over candidate answer-token IDs, alongside full-vocabulary task supervision on both views. This restriction targets agreement at the answer interface without requiring identical distributions over all continuations. Counterfactual data augmentation (Kaushik et al., 2020) provides a related approach to reducing reliance on spurious features. Our paired-supervision configuration ($\beta = 0$) separates the use of both lexical views from the additional consistency penalty. The contribution lies in this task-specific formulation and its empirical analysis, rather than a new general consistency objective.

2.3 Mechanistic interpretability of hidden representations

Our logit-lens-style diagnostic projects intermediate states through the unembedding matrix; the tuned lens (Belrose et al., 2023) instead learns layer-specific transformations for latent prediction. Studies of factual recall examine how information is represented and used across layers (Meng et al., 2022; Geva et al., 2023), while activation addition and representation engineering investigate behavioural control through hidden-state interventions (Turner et al., 2023; Zou et al., 2023). We combine lexical-family probes with activation patching and donor controls. These measurements address distinct questions: probes assess recoverability of a specified lexical proxy, patching tests whether an injected state changes the answer, and donor controls test whether that effect requires the matched question. This distinction guides our interpretation of the diagnostics without equating decodability or answer control with recovery of causal information.

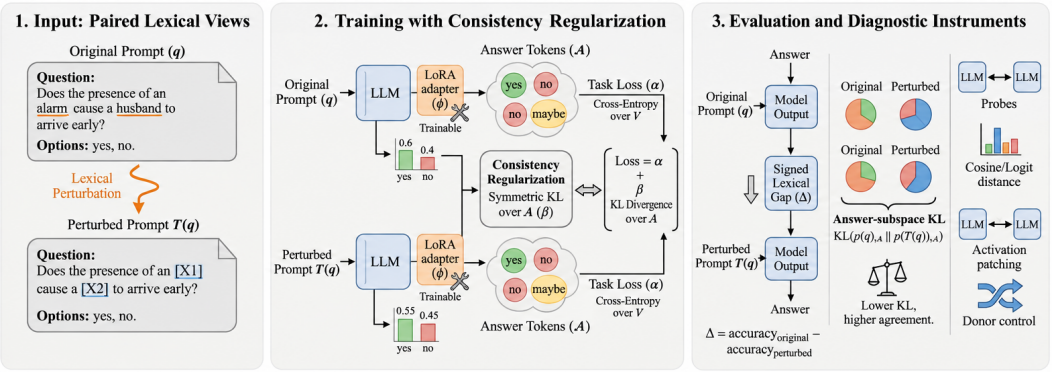

Figure 1: Vernier training schematic. The original prompt q and its structure-preserving perturbation $T(q)$ are forwarded through the same LoRA-adapted model. The loss combines per-view answer supervision with an answer-subspace symmetric KL term.

3 Method: Vernier

The method has three parts (see Figure 1): construction of paired lexical views, supervised training on both views, and consistency regularization over candidate answer tokens. We distinguish full-vocabulary task supervision from the restricted normalization used for consistency. A margin-dependent bound connects distributional agreement to prediction agreement. The empirical analyses then assess task performance, representation measurements, and donor specificity on their respective evaluation populations.

3.1 Lexical perturbation and gap

A causal-QA item is $(q, y^\star, M)$, where q is the prompt, $y^\star \in \mathcal{Y}$ is the gold answer, and M is the structural causal model. The lexical perturbation T replaces each SCM variable name and its variants with a typed placeholder $X_1, X_2, \ldots$, applied consistently throughout the item. The implementation assigns placeholders by sorted SCM identifiers. A worked prompt pair is provided in Appendix F (Table A24). It preserves M, the query, and $y^\star$, so an SCM-grounded reasoner should satisfy $R(\cdot \mid q) = R(\cdot \mid T(q))$. We score a model by the signed lexical gap

$$\Delta(\theta; \mathcal{D}) = \mathbb{E}_{(q,y^\star)\sim\mathcal{D}}[\mathbf{1}\{\hat{y}_\theta(q) = y^\star\} - \mathbf{1}\{\hat{y}_\theta(T(q)) = y^\star\}], \tag{1}$$

where predictions use the answer-token distribution at the last non-padding prompt position.

On CLadder, the perturbation is conservative. It does not paraphrase the story, change numbers, alter the causal graph, or change the requested answer type. It only removes the English variable names that can act as lexical anchors. This matters for interpretation: a remaining gap cannot be attributed to a changed SCM or a changed label, while a closed gap cannot by itself prove deeper causal understanding. The method therefore asks a narrower question: given two views that should be equivalent for the task, can a small weight update make the model read out the same answer-relevant information from both?

3.2 Vernier objective

Vernier trains a LoRA adapter on paired views $(P_0, P_1) = (q, T(q))$. For CLadder, the two prompts specify the same structural causal query under different variable names. Let $z^{(v)} \in \mathbb{R}^{|\mathcal{V}|}$ denote the next-token logits at the last non-padding prompt position. With gold answer token $t^\star$, the task loss is

$$\mathcal{L}_{\text{task}}(P_v) = -\log \text{softmax}(z^{(v)})_{t^\star},$$

normalized over the full vocabulary $\mathcal{V}$. Paired supervision applies this loss to both views; the consistency term is normalized separately over candidate answer-token IDs $\mathcal{A}$:

$$\mathcal{L} = \alpha[\mathcal{L}_{\text{task}}(P_0) + \mathcal{L}_{\text{task}}(P_1)] + \beta\,\mathcal{L}_{\text{cons}}(P_0, P_1), \quad \mathcal{L}_{\text{cons}} = \tfrac{1}{2}\text{KL}(p^{\mathcal{A}}_{P_0} \| p^{\mathcal{A}}_{P_1}) + \tfrac{1}{2}\text{KL}(p^{\mathcal{A}}_{P_1} \| p^{\mathcal{A}}_{P_0}). \tag{2}$$

Here $p^{\mathcal{A}}_{P_0}$ and $p^{\mathcal{A}}_{P_1}$ are softmax distributions over candidate token IDs. The implementation collects the first token of each answer string with and without a leading space, then deduplicates IDs. The gold token uses the space-prefixed form when available. Evaluation takes the highest-scoring candidate token and maps it to its answer label, rather than summing probability across token variants. We use $\alpha = \beta = 1$, summing the two task losses. Only LoRA parameters are updated. Right padding makes the attention-mask sum minus one the last non-padding prompt position; both views are truncated at 1,024 tokens. The training update is summarized in Appendix G (Algorithm 1).

The answer-subspace restriction is a design choice, not a computational shortcut. A full-vocabulary KL would require two prompts with different surface forms to agree on every next-token continuation, including continuations that restate the variable names. That target is stronger than the task invariance and can penalise harmless lexical differences. By contrast, the answer-subspace KL matches the evaluation interface: if two views encode the same causal problem, the distribution over admissible answers should agree even if the unrestricted language-model distribution does not.

3.3 Distinguishing misalignment from information loss

A paired update can improve performance through several routes: better question-conditioned representations, improved access to existing content, or a change in answer preferences. Lexical-content probes measure one observable consequence of these changes: whether a decoder fitted on the original view transfers to the placeholder view. Increased transfer is compatible with more similar coding of that proxy. It does not by itself establish preservation of every answer-relevant fact or identify the computation used to answer.

We therefore interpret the measurements jointly. Behavioural improvement requires examining both component accuracies, not just their difference. Cosine and logit-lens measurements describe changes at the decision interface; probes describe decoder performance for a specified label. Patching tests whether an injected state can control the answer. The donor controls in Figure A2 then test whether that control specifically requires the matched question. These measurements constrain an explanation without treating any one diagnostic as a complete mechanism test.

3.4 Answer-subspace consistency

The answer-token KL measures distributional disagreement on $\mathcal{A}$. Let D_{SKL} be the symmetric KL and let

$$G(\theta; \mathcal{D}) = \mathbb{E}_{q \sim \mathcal{D}}[D_{\mathrm{SKL}}(p_\theta(\cdot \mid q)_{\mathcal{A}} \| p_\theta(\cdot \mid T(q))_{\mathcal{A}})]. \tag{3}$$

For a fixed model and batches sampled from $\mathcal{D}$, the batch mean of $\mathcal{L}_{\mathrm{cons}}$ estimates G. Since symmetric KL is positive definite on the finite simplex, $G = 0$ iff the two answer-token distributions agree almost everywhere. Distributional agreement then implies equal arg-max predictions, but away from zero this only controls prediction flips under a margin condition. If $p_\theta(\cdot \mid q)_{\mathcal{A}}$ has token-level top-two margin at least $\gamma > 0$ almost everywhere, with fixed tie-breaking, then a flip requires total variation at least $\gamma/2$, and Pinsker plus Jensen gives

$$\Pr[\hat{y}_\theta(q) \neq \hat{y}_\theta(T(q))] \leq \frac{\sqrt{2G(\theta; \mathcal{D})}}{\gamma}. \tag{4}$$

A label change requires a token-argmax change under the fixed token-to-label map, so the bound also applies to answer labels. If the margin assumption fails, the low-margin probability must be accounted for separately. The bound concerns agreement rather than accuracy: a constant answer can satisfy consistency while remaining wrong. Task supervision supplies a distinct learning signal, but does not guarantee avoidance of such solutions.

3.5 Identification limits

The intervention tests whether paired lexical training improves the measured answers and which diagnostics change alongside it. A small gap caused by original-view regression is distinct from improvement on both views. Likewise, lexical-proxy decoding differs from recovery of causal facts, and answer transfer under patching differs from recovery of information already present in the recipient. The outcome map in Table A21 makes these distinctions explicit. Changes in representations, decoding, and answer thresholds remain compatible explanations unless an additional comparison separates them.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. CLadder (Jin et al., 2023) is the in-distribution benchmark. The legacy result archive specifies training on the first 1,500 entries of the file labelled `commonsensical` and evaluation on a 200-item slice; CRASS (Frohberg & Binder, 2022) and e-CARE (Du et al., 2022) are the recorded transfer benchmarks. These historical tables summarize saved evaluations; the common split description alone does not establish training membership for every checkpoint. We therefore separately verify a complete execution record for a Qwen-7B paired update: 1,500 training question IDs and 900 evaluation IDs with zero intersection. Under the corrected condition mapping, training contains 1,224 commonsense and 276 anti-commonsense items; evaluation has 150 items in each rung crossed with anti-commonsense or nonsensical naming. This is question-level separation, without a template-disjoint guarantee. General-capability screens use 200 items each from MMLU (Hendrycks et al., 2021), HellaSwag (Zellers et al., 2019), and GSM8K (Cobbe et al., 2021).

Baselines. The Qwen2.5-7B reference configurations compare Vernier ($\beta = 1$, answer-token KL) with the following settings. Their recorded results are descriptive comparisons; a common checkpoint-step label does not establish matched training schedules. **Base.** The pretrained instruction-tuned model with no fine-tuning. **P_0-only SFT.** Supervised fine-tuning on the original prompt alone, with no perturbed view. **R-Drop** (Liang et al., 2021). Symmetric KL between two dropout-sampled forward passes of P_0, with no perturbation T. **Twin task** ($\beta = 0$). Both views routed through the task loss, with no consistency term. **Full-vocabulary KL** ($\beta = 1$). The same training as Vernier but with the KL computed over the full vocabulary rather than the answer subspace.

Metrics. For each item we score $(q, T(q))$ by taking the softmax over the answer-token subset at the last non-padding prompt position. The primary metric is the signed lexical gap $P_0 - P_1$ in percentage points, reported alongside P_0 and P_1 accuracy so that closure driven by P_1 improvement can be distinguished from closure driven by P_0 regression. A gap near zero is interpreted together with the component accuracies. A negative gap is an inversion in which P_1 exceeds P_0, which we report as such rather than as a larger success. For OOD benchmarks we also give the relative gap reduction $\Delta_{gap} = (gap_{Vernier} - gap_{base})/gap_{base}$ as a convenience, but we caution that this ratio is unstable when the base gap is small and can exceed 100 percent when the gap crosses zero, so we anchor all claims on the signed gap in percentage points rather than on this ratio. On general-capability benchmarks we report top-1 accuracy.

Implementation. We attach a LoRA (Hu et al., 2022) adapter of rank $r = 16$ to the query, key, value, and output projections of every attention layer, with LoRA alpha 32 and dropout 0.05. The reference recipe uses AdamW (Loshchilov & Hutter, 2019) at learning rate 1×10^{-4}, weight decay 0, batch size 4 pairs per micro-step, gradient accumulation 4, linear warmup over 50 steps, cosine decay thereafter, and a total budget of 500 steps. We evaluate nine instruction-tuned models spanning five families, namely Phi-3-mini-4k-instruct (3.8B) (Abdin et al., 2024), Qwen2.5 at 1.5B, 3B, 7B, 14B, and 32B (Yang et al., 2024), Llama-3.1-8B-Instruct (Grattafiori et al., 2024), Mistral-7B-Instruct-v0.3 (Jiang et al., 2023), and Gemma-2-2B-it (Gemma Team, 2024). The 1.5B to 8B models train in BF16 on a single 24 GB GPU. The 14B and 32B models use QLoRA (Dettmers et al., 2023) on a 40 GB GPU, the 32B with batch size 2 and gradient accumulation 8. For Phi-3, the reported attention adapter is checkpoint 70 from a completed 100-step run with peak learning rate 5×10^{-6}, 50 warmup steps, and cosine decay (Section D). On Mistral-7B the canonical learning rate produces an aggressive in-distribution closure with substantial downstream-capability degradation. A lower-rate Mistral trajectory (lr=5×10^{-6}, reported checkpoint 150) provides an additional operating point; both rows are included in the main tables. The hyperparameter trade-off is analysed with the failure modes. The recipe audit distinguishes saved checkpoint descriptions from verified execution settings (see Section C).

4.2 COMPARISON

The execution-verified Qwen-7B comparison provides a direct behavioural reference. With seed 42 and 500 total optimizer steps, original/anonymized accuracies rise from 52.6/49.9% to 78.6/82.8%

Table 1: Results on CLadder (ID), CRASS, and e-CARE (OOD), $n = 200$ each. P_0/P_1: original/perturbed accuracy; Gap: $100(P_0 - P_1)$; Rel. gap: percentage change from base. Uncertainty: three-run sample SD for adapted † entries, item-bootstrap SE otherwise. **Bold**/underline: best/second-best displayed values per dataset and column, including ties (accuracy ↑, |Gap| ↓, Rel. gap ↓); no significance implied. See Section C for provenance.

	Base			Vernier			
Model	P_0	P_1	**Gap**	P_0	P_1	**Gap**	**Rel. gap**
CLadder (ID)							
Phi-3-mini (3.8B)	0.585±0.035	0.540±0.035	+4.5±2.3	0.550±0.035	0.550±0.035	**0.0**±2.1	—
Qwen2.5-7B†	**0.625**±0.034	0.545±0.035	+8.0±2.6	0.868±0.003	0.893±0.010	-2.5±0.9	—
Qwen2.5-14B†	0.615±0.035	0.580±0.035	+3.5±2.7	0.955±0.033	0.950±0.018	+0.5±1.7	—
Qwen2.5-32B (QLoRA)	0.565±0.035	**0.585**±0.035	**−2.0**±2.1	0.970±0.012	0.965±0.013	+0.5±1.3	—
Llama-3.1-8B†	0.540±0.035	0.515±0.035	+2.5±3.6	0.945±0.005	0.948±0.010	−0.3±1.3	—
Mistral-7B (lr=10^{-4})	0.610±0.034	0.565±0.035	+4.5±3.9	**0.980**±0.010	**0.995**±0.005	-1.5±1.1	—
Mistral-7B (lr=5×10^{-6}, s150)	0.610±0.034	0.565±0.035	+4.5±3.9	0.605±0.035	0.600±0.035	+0.5±3.1	—
CRASS (OOD)							
Phi-3-mini (3.8B)	0.845±0.026	0.545±0.035	+30.0±3.9	0.840±0.026	0.550±0.035	+29.0±3.9	−3%
Qwen2.5-7B†	0.845±0.026	0.505±0.035	+34.0±3.9	0.870±0.013	0.622±0.012	+24.8±0.3	−27%
Qwen2.5-14B†	0.895±0.022	0.625±0.034	**+27.0**±3.5	0.918±0.008	0.722±0.013	**+19.7**±1.0	−27%
Qwen2.5-32B (QLoRA)	**0.910**±0.020	**0.635**±0.034	+27.5±3.6	**0.950**±0.015	**0.725**±0.031	+22.5±3.3	−18%
Llama-3.1-8B†	0.850±0.025	0.445±0.035	+40.5±3.8	0.872±0.010	0.558±0.040	+31.3±3.8	−23%
Mistral-7B (lr=10^{-4})	0.795±0.028	0.415±0.035	+38.0±3.9	0.710±0.032	0.500±0.036	+21.0±3.4	**−45**%
Mistral-7B (lr=5×10^{-6}, s150)	0.795±0.028	0.415±0.035	+38.0±3.9	0.735±0.031	0.365±0.034	+37.0±3.8	−3%
e-CARE (OOD)							
Phi-3-mini (3.8B)	0.795±0.029	0.590±0.035	+20.5±3.9	0.785±0.029	0.600±0.035	+18.5±3.9	−10%
Qwen2.5-7B†	0.815±0.028	0.585±0.035	+23.0±3.8	0.810±0.010	**0.622**±0.012	+18.8±1.5	−18%
Qwen2.5-14B†	**0.820**±0.027	0.605±0.035	+21.5±3.7	**0.823**±0.003	0.610±0.005	+21.3±0.6	−1%
Qwen2.5-32B (QLoRA)	**0.820**±0.027	**0.625**±0.034	**+19.5**±3.8	0.815±0.028	0.595±0.035	+22.0±3.9	+13%
Llama-3.1-8B†	0.790±0.029	0.550±0.035	+24.0±4.5	0.793±0.006	0.615±0.005	**+17.8**±0.8	**−26**%
Mistral-7B (lr=10^{-4})	0.790±0.029	0.555±0.035	+23.5±4.1	0.775±0.030	0.580±0.035	+19.5±4.2	−17%
Mistral-7B (lr=5×10^{-6}, s150)	0.790±0.029	0.555±0.035	+23.5±4.1	0.770±0.030	0.565±0.035	+20.5±3.6	−13%

on the same 900 questions. The paired gains are 26.0 and 32.9 pp, with 95% stratified item-bootstrap intervals of [22.0,30.0] and [29.0,36.7] pp. These intervals describe item sampling for the evaluated checkpoints. The signed gap changes from $+2.67$ to -4.22 pp: the gain in task accuracy is substantial, but does not establish reduced lexical disparity. The historical multi-model comparisons below provide complementary descriptive results on their own evaluation slices.

In-distribution gap closure is reported in Table 1. Vernier drives the gap to at or near zero on the listed model configurations. On Qwen-7B, Qwen-14B, and Llama-3.1-8B both P_0 and P_1 rise by 24 to 44 pp from base, so the closure is driven by joint accuracy gains rather than by a regression on P_0. Mistral-7B shows the same joint gains under the default learning rate but at a steep capability cost, so we report its rescued low-rate row as the operating point. On Phi-3 the closure has the opposite signature, with P_0 falling by 3.5 pp and P_1 rising by only 1.0 pp. This model-specific failure is analysed with the mechanism and scale diagnostics. The Qwen2.5-32B base model already has a negative gap of -2.0 pp on this slice, so the gap-closure metric is not directly comparable, but both P_0 and P_1 rise by roughly 40 pp after Vernier. The in-distribution base gaps are small on larger held-out CLadder slices, and their paired confidence intervals cross zero for the base models (see Table A15). We therefore treat in-distribution closure as a sanity check and rest the behavioural case on the out-of-distribution transfer below.

Zero-shot transfer to CRASS and e-CARE is reported in Table 1. The recorded CLadder-adapter evaluations show CRASS gap reductions of 18 to 27 percent across the three Qwen scales, with the Qwen-7B value averaged over three recorded runs. The larger-OOD check strengthens this CRASS result rather than weakening it (see Table A14). On all 274 prepared CRASS items, the gap falls from 35.0 to 23.7 pp on Qwen-7B and from 27.4 to 16.8 pp on Qwen-14B. The e-CARE picture on Qwen is weaker and non-monotonic in scale. At $n = 200$, Qwen-7B reduces e-CARE by 18 percent, Qwen-14B shows a small point-estimate reduction (about 1 percent), and Qwen-32B widens the e-CARE gap by 2.5 pp despite closing CLadder and CRASS. At $n = 1000$, the e-CARE

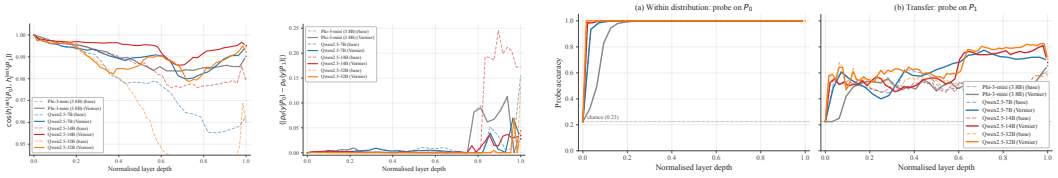

Figure 2: Layer-wise mechanism diagnostics on held-out CLadder. (a) Last-token hidden-state cosine, (b) answer-token logit-lens disagreement, (c) and (d) variable-name probe accuracy. Dashed curves are base models, and solid curves are Vernier adapters.

reduction remains small, from 20.7 to 18.7 pp on Qwen-7B and from 23.2 to 22.4 pp on Qwen-14B (see Table A14). Llama-3.1-8B reduces the CRASS gap by 23 percent and the e-CARE gap by 26 percent as a three-run mean. The rescued Mistral-7B (lr=5×10^{-6}, step 150) shows a 13 percent e-CARE reduction but only 3 percent on CRASS. Phi-3 shows small observed transfer changes on both benchmarks, alongside small changes in the reported diagnostics.

4.3 Ablation Study

The saved Qwen-7B reference configurations are reported in Table A1. Original-view SFT has an in-distribution gap of 9.5 pp, compared with 8.0 pp for the base model. R-Drop has a gap of 18.5 pp, with $P_0 = 0.965$ and $P_1 = 0.780$. Paired supervision without KL ($\beta = 0$) instead has a 1.0 pp gap, with recorded CRASS and e-CARE reductions of 21 and 13 percent. This pattern motivates training on both lexical views rather than relying on dropout consistency alone. On the larger CRASS evaluation ($n = 274$), the reported Qwen-7B gap is 18.2 pp for paired supervision and 23.4 pp for the full loss, against 34.3 pp for the base comparison; Qwen-14B shows similar reductions for the two recipes (Table A2). Full-vocabulary KL exhibits a different transfer trade-off. These rows do not establish a consistent accuracy advantage for answer-token KL. Because the historical summaries do not certify identical training membership and optimization schedules across recipes, their differences are not isolated causal estimates of the KL contribution.

4.4 Mechanism Analysis

Two layer-wise measurements at the decision token are plotted for Phi-3 and the three Qwen scales (see Figure 2). Cosine similarity measures geometric similarity between the two decision-token states. The logit-lens measurement compares projected answer-token scores across views. On all three Qwen scales Vernier raises the cosine and lowers the disagreement, with the effect concentrated in the upper third of the network. On Phi-3 the final-layer summaries change little. The final-layer summaries across models are reported in Table A3. The logit-lens reduction factor is 5.5 on Qwen-7B, 5.0 on Qwen-14B, and 3.1 on Qwen-32B, with Qwen-32B also showing lower base disagreement. As a methodological control, we also computed the cosine pooled over all sequence positions rather than at the last non-pad token. The pooled signal moves by only 0.006 at the final layer on Qwen-7B against a 0.032 shift in the last-token cosine. The pooled and decision-token summaries therefore capture different aspects of the representations.

The probe analysis measures transfer of a lexical-content decoder (Figure 2 and table A3). A multinomial classifier fitted on original-view decision-token states predicts variable-name-family labels on the placeholder view. Reported final-layer transfer rises from 0.615 to 0.655 on Phi-3, 0.480 to 0.695 on Qwen-7B, 0.590 to 0.680 on Qwen-14B, and 0.585 to 0.705 on Qwen-32B. These increases concern a lexical proxy, rather than an annotation of all causal facts needed for the answer. Residual wording and template cues can also support proxy prediction. The results are therefore compatible with improved decoder transfer, without uniquely establishing recovery of task-relevant information. This diagnostic batch is distinct from the execution-verified 900-question comparison.

Paired supervision and answer-distribution agreement. The reference diagnostics in Table A4 compare paired supervision with and without KL. On Qwen-7B, probe transfer is 0.690 without KL and 0.695 with it; on Qwen-14B, the corresponding values are 0.730 and 0.680. Qwen-7B logit-lens disagreement is 0.074 without KL and 0.028 with it, against 0.155 at base. The observed

distinction is strongest for answer-distribution agreement, rather than for lexical-proxy transfer. These measurements characterize the saved checkpoints; they do not establish that KL alone caused the difference.

A causal test by activation patching. The cosine, logit-lens, and probe diagnostics are observational. Activation patching (Zhang & Nanda, 2024) instead tests the effect of an injected state. For each item, we copy the original-view decision-token state into the placeholder-view computation at a selected decoder layer. At relative depth 0.8, base Qwen-7B patched accuracy is 0.605 versus clean $P_1 = 0.535$ and $P_0 = 0.615$; agreement with the target item's clean original-view answer is 0.970. Qwen-14B gives 0.645 patched accuracy and 0.995 target agreement at the same relative depth. Llama-3.1-8B shows a smaller accuracy gain but high answer agreement (Table A5). These are effects at the intervention site, rather than a localization of the computation changed by training.

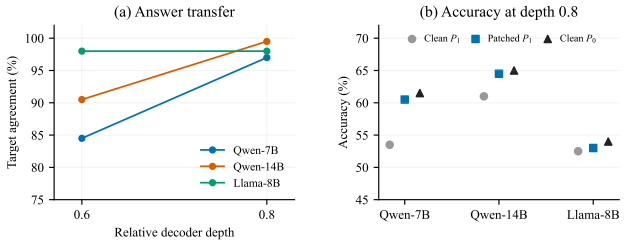

Figure 3: Activation patching on base Qwen-7B, Qwen-14B and Llama-8B (CLadder, $n = 200$). (a) Agreement with the target item's clean original-view prediction at intermediate decoder layers. (b) Clean and patched accuracies at relative depth 0.8. These are separate diagnostic batches. Matched-donor agreement measures answer transfer; donor specificity is evaluated in Figure A2.

The donor comparison is decisive for interpretation. An unrelated donor with the same clean prediction gives 0.620 patched accuracy and 0.965 target agreement on Qwen-7B, numerically close to the matched donor. A random donor gives 0.685 target agreement but 0.930 agreement with its own prediction. A different-prediction donor gives 0.175 target agreement and 0.825 donor agreement (Table A6). Thus the state strongly controls answer identity, while the matched outcome is not specific to an item-matched donor. Same/different groups are defined by model predictions, not gold correctness; numerical closeness is not a statistical equivalence claim.

What an injected state establishes. A matched donor supplies the recipient with an entire hidden state. Successful patching consequently shows that the injected state can control the answer, not that the recipient already retained all information carried by the donor. The unrelated answer-matched control further separates answer identity from item matching. Distinguishing a change in representational coding from a change in the decoder or decision boundary requires additional constraints on the intervention. We base the present interpretation on these donor comparisons, rather than on negative results from fitted global maps.

4.5 Sensitivity, robustness, and capability

The auxiliary checks support the same scoped conclusion. The recorded β and hyperparameter sweeps describe sensitivity across the listed configurations (see Figure A1 and table A9). Placeholder-scheme stress tests show that the learned invariance is not fully scheme-independent. Qwen-7B remains closed across held-out schemes, but Qwen-14B is brittle on `VAR_A, VAR_B` (see Table A10). Larger-n, paired, and stratified CLadder checks show that the in-distribution base gap is modest but the paired reduction remains positive on average (see Tables A12, A13, A15 and A16). Larger-OOD checks show a strong CRASS reduction on Qwen-7B and Qwen-14B but only a small e-CARE reduction (see Table A14). The sampled capability checks do not detect broad damage on the Qwen and Phi rows, but they are too small to certify capability preservation. Mistral is the clear failure case (see Figures A3 and A4). Joint training on CLadder/CRASS/e-CARE trades a little CLadder closure for a much smaller CRASS gap (see Table A11). A lexical-overlap audit finds no 8-gram or 13-gram overlap between the CLadder training prompts and the sampled MMLU, HellaSwag, or GSM8K capability items (see Table A8), so the sampled capability changes are unlikely to be explained by verbatim train-test contamination. These checks narrow the claim rather than broaden it. The recipe is useful in a working regime, not a universally safe fine-tune.

4.6 When the method works: capacity and base-model identity

To turn the question of when Vernier works into a measured one, we add three sub-4B models and run the full diagnostic suite on each (see Table A17 and fig. A1). Two are from the Qwen family (Qwen2.5-1.5B and Qwen2.5-3B), so comparing them against the larger Qwen models compares scale within one model family. The third is Gemma-2-2B, a second small family.

A within-family scale contrast. Inside the Qwen family the 1.5B and 3B configurations differ under our joint descriptive criterion: the gap should close without P_0 regression, probe transfer should rise, and logit-lens disagreement should fall. The summaries report three-run means for these endpoints. Qwen-1.5B does not meet this criterion in the aggregate diagnostic summary: after training the gap is $+3.5 \pm 0.8$ pp, probe transfer drops by 0.09, and logit-lens disagreement barely moves. Qwen-3B meets it in the reported aggregate summary: P_0 and P_1 both rise to 0.88 with the gap at $+0.5 \pm 0.7$ pp, the probe transfer rises by 0.29, the largest gain of any model, and the logit-lens disagreement halves. The two endpoints separate cleanly on the mechanism diagnostics and the reported run-to-run spread (std under 1 pp on the gap) is far smaller than the gap between them. Since Qwen-1.5B and Qwen-3B share architecture and training family, this is a within-family contrast, although model size remains coupled to other model-design and training choices.

Scale alone does not determine the outcome. The threshold is not a parameter-count rule. Phi-3-mini has 3.8B parameters and fails, while Qwen-3B is smaller and succeeds. Model family, pretraining, and instruction tuning also matter. Mistral-7B makes the same point at a larger scale because it needs a lower learning rate than the other 7B-class models. Gemma-2-2B is useful for a different reason. Its in-distribution gap closes with both views rising, but its logit-lens disagreement widens and probe transfer moves only 0.07. Gemma therefore illustrates that accuracy gains need not coincide with the same representation measurements seen in Qwen. Across these configurations, parameter count alone does not predict the joint pattern.

Observed gains extend beyond Qwen. Qwen-7B, Qwen-14B, and Llama-3.1-8B agree across the behavioural diagnostics, and the recorded aggregate covers three runs. Qwen-32B is harder to compare because its base in-distribution gap is already near zero, but the mechanism still moves: last-token cosine rises by 0.032, logit-lens disagreement drops 3.1-fold, and variable-name probe transfer rises by 0.12. e-CARE is the main exception. Its gap changes little on Qwen-14B and reverses by 2.5 pp on Qwen-32B. We ran the same cosine diagnostic on e-CARE and found that Vernier still moves the two final-layer decision-token states closer together on all three Qwen scales (see Table A18). Thus greater final-layer cosine similarity can coexist with a remaining behavioural gap; the two measurements are not interchangeable.

4.7 Preliminary non-causal rename checks

We extend the rename checks to two tasks from BIG-Bench Hard (BBH): Logical Deduction, evaluated on 100 held-out items, and Reasoning About Colored Objects, evaluated on a 20-item diagnostic subset (Table A19). Paired augmentation reduces the absolute rename gap while improving both views. For example, on Qwen-7B colored objects, original/renamed accuracy increases from 0.500/0.300 to 0.800/0.850, changing the signed gap from $+20.0$ to -5.0 pp. Decision-token patching also transfers answers in the tested settings, although the most effective depth varies. These observations extend the evaluation beyond causal QA without establishing a shared mechanism across tasks. The colored-objects results are descriptive, and patching and adaptation may use different evaluation items.

5 Conclusion

Vernier combines paired supervision with answer-token consistency to address lexical sensitivity while distinguishing task improvement from lexical agreement. The recorded multi-model results show stronger transfer on CRASS than on e-CARE, with substantial variation across training recipes and model families. The execution-verified comparison confirms large accuracy gains on both lexical views without establishing a smaller absolute lexical gap. Representation diagnostics reveal a complementary distinction: transferable lexical decoding and control over answer identity do not uniquely identify the computation changed by adaptation. Together, the behavioural and donor-control results support evaluating lexical repair through correctness, agreement, and intervention specificity.

AI USE STATEMENT

AI tools assisted with language checking and polishing. No additional model-training or inference experiments were performed as part of these revisions. The authors take responsibility for the accuracy of the reported results, citations, and final manuscript.

ETHICS STATEMENT

This work studies lexical sensitivity using existing reasoning benchmarks. The reported improvements concern the evaluated perturbations and tasks; they do not establish reliability or fairness in deployment. In particular, entity anonymization is used as an experimental intervention and should not be interpreted as a privacy guarantee.

REPRODUCIBILITY STATEMENT

Code is available at https://anonymous.4open.science/r/Vernier. The repository includes a reference training configuration, saved prediction records, and scripts for recomputing statistical summaries. Experimental settings and evaluation protocols are documented in the main text and appendix.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024. URL https://arxiv.org/abs/2404.14219. arXiv:2404.14219.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens, 2023. URL https://arxiv.org/abs/2303.08112. arXiv:2303.08112.

Haoang Chi, He Li, Wenjing Yang, Feng Liu, Long Lan, Xiaoguang Ren, Tongliang Liu, and Bo Han. Unveiling causal reasoning in large language models: Reality or mirage? In *Advances in Neural Information Processing Systems*, volume 37, pp. 96640–96670, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/af2bb2b2280d36f8842e440b4e275152-Abstract-Conference.html.

Karl Cobbe, Vineet Kosaraju, Mohammad 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 preprint arXiv:2110.14168*, 2021. URL https://arxiv.org/abs/2110.14168.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1feb87871436031bdc0f2beaa62a049b-Abstract.html.

Li Du, Xiao Ding, Kai Xiong, Ting Liu, and Bing Qin. e-CARE: a new dataset for exploring explainable causal reasoning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 432–446, 2022. doi: 10.18653/v1/2022.acl-long.33. URL https://aclanthology.org/2022.acl-long.33/.

Jörg Frohberg and Frank Binder. CRASS: A novel data set and benchmark to test counterfactual reasoning of large language models. In *Proceedings of the 13th Language Resources and Evaluation Conference (LREC)*, 2022. URL https://aclanthology.org/2022.lrec-1.229/.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021. URL https://aclanthology.org/2021.emnlp-main.552/.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024. URL https://arxiv.org/abs/2408.00118. arXiv:2408.00118.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://aclanthology.org/2023.emnlp-main.751/.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783v3. arXiv:2407.21783v3.

Yuren Hao, Xiang Wan, and ChengXiang Zhai. An investigation of robustness of LLMs in mathematical reasoning: Benchmarking with mathematically-equivalent transformation of advanced mathematical problems, 2025. URL https://arxiv.org/abs/2508.08833.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2009.03300.

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 (ICLR)*, 2022. URL https://arxiv.org/abs/2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, et al. Mistral 7B, 2023. URL https://arxiv.org/abs/2310.06825. arXiv:2310.06825.

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*, volume 36, pp. 31038–31065, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/631bb9434d718ea309af82566347d607-Abstract-Conference.html.

Divyansh Kaushik, Eduard Hovy, and Zachary C. Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1909.12434.

Donggyu Lee, Hyeok Yun, Meeyoung Cha, Sungwon Park, Sangyoon Park, and Jihee Kim. EconCausal: A context-aware causal reasoning benchmark for large language models in social science, 2025. URL https://arxiv.org/abs/2510.07231v3. Version 3, revised 23 February 2026; first submitted in 2025.

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*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/5a66b9200f29ac3fa0ae244cc2a51b39-Abstract.html.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://arxiv.org/abs/1711.05101.

Tom McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019. URL https://aclanthology.org/P19-1334/.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/6f1d43d5a82a37e89b0665b33bf3a182-Abstract-Conference.html.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324.

Alexander Matt Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid. Activation addition: Steering language models without optimization, 2023. URL https://arxiv.org/abs/2308.10248v3. arXiv:2308.10248v3.

Yuzhe Wang, Yaochen Zhu, and Jundong Li. CausalFlip: A benchmark for LLM causal judgment beyond semantic matching. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 9963–9972. Association for Computing Machinery, 2026. doi: 10.1145/3770855.3817490. URL https://doi.org/10.1145/3770855.3817490.

Albert Webson and Ellie Pavlick. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2022. URL https://aclanthology.org/2022.naacl-main.167/.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 6256–6268, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi 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, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2024. URL https://arxiv.org/abs/2412.15115v2. arXiv:2412.15115v2.

Zhenyu Yu and Shuigeng Zhou. Caliper: Probing lexical anchors versus causal structure in LLMs, 2026. URL https://arxiv.org/abs/2606.04915.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 4791–4800, 2019. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, et al. Representation engineering: A top-down approach to AI transparency, 2023. URL https://arxiv.org/abs/2310.01405. arXiv:2310.01405.

A ADDITIONAL DIAGNOSTIC FIGURES

The consistency-weight sweep and the scale summary for the logit-lens reduction diagnostic are reported in Figure A1. The plotted summaries describe variation around $\beta \in [0.5, 1]$ and at $\beta = 5$; interpreting these differences as the isolated effect of regularization requires matched training configurations. The scale panel is kept in the appendix because the model-by-model interpretation is already reported in the main text.

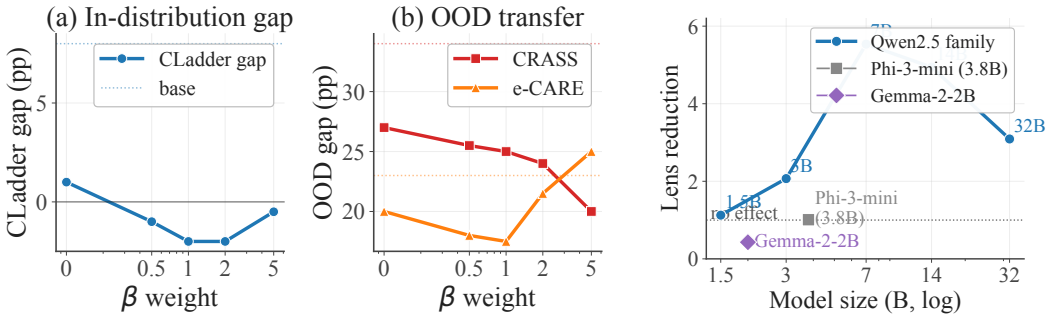

Figure A1: Additional diagnostics. The left panel gives the Qwen2.5-7B sweep of the consistency-loss weight β. The right panel gives logit-lens reduction by scale.

B ADDITIONAL RESULTS TABLES AND FIGURES

Table A1: Loss-component ablation on Qwen2.5-7B (held-out CLadder $n = 200$, CRASS $n = 200$, e-CARE $n = 200$), all at step 500 with seed 42. Best results in each column are in bold.

	In-distribution (CLadder)			Out-of-distribution		
Variant	P_0	P_1	**Gap**	**CRASS Gap**	**e-CARE Gap**	**Mean Δ_{OOD}**
Base (no fine-tuning)	0.625	0.545	+8.0	+34.0	+23.0	—
P_0-only SFT	0.850	0.755	+9.5	+28.0 (-18%)	+23.5 (-0%)	−9%
R-Drop (Liang et al., 2021) (no T)	0.965	0.780	+18.5	+25.5 (-25%)	+22.0 (-4%)	−15%
Twin task ($\beta = 0$, no KL)	0.885	0.875	+1.0 (-87.5%)	+27.0 (-21%)	+20.0 (-13%)	−17%
Full-vocab KL ($\beta = 1$)	0.930	0.925	+0.5 (-94%)	**+21.5 (-37%)**	+21.5 (-7%)	−22%
Vernier ($\beta = 1$, answer-subspace KL)	0.870	0.890	**−2.0 (-125%)**	+25.0 (-26%)	**+17.5 (-24%)**	**−25%**

Table A2: Counterfactual augmentation versus the full answer-subspace KL loss on CRASS. Both adapters are trained only on CLadder. The 274-item rows use all prepared CRASS items.

Model	n	**Setting**	P_0	P_1	**Gap**	**Δ gap**
Qwen2.5-7B	200	Base	0.845	0.505	+34.0	–
		Twin task	0.890	0.650	+24.0	−29%
		Vernier	0.880	0.635	+24.5	−28%
Qwen2.5-7B	274	Base	0.839	0.496	+34.3	–
		Twin task	0.891	0.708	**+18.2**	**−47%**
		Vernier	0.872	0.639	+23.4	−32%
Qwen2.5-14B	200	Base	0.880	0.610	+27.0	–
		Twin task	0.915	0.715	+20.0	−26%
		Vernier	0.920	0.720	+20.0	−26%
Qwen2.5-14B	274	Base	0.894	0.617	+27.7	–
		Twin task	0.920	0.745	+17.5	−37%
		Vernier	0.923	0.759	**+16.4**	**−41%**

Table A5: Causal activation patching on Qwen2.5-7B, Qwen2.5-14B, and Llama-3.1-8B (held-out CLadder, $n = 200$). The P_0 decision-token hidden state at the listed decoder depth is copied into the corresponding P_1 state. "P_1 patched" is post-patch P_1 accuracy, and "flip-to-P_0" is agreement with the clean P_0 prediction.

Model	Setting	Depth	P_0	P_1 clean	P_1 patched	flip-to-P_0
Qwen2.5-7B	Base	0.6	0.615	0.535	0.540	0.845
		0.8	0.615	0.535	**0.605**	**0.970**
		1.0	0.615	0.535	0.590	0.635
	Vernier	0.6	0.870	0.890	0.890	0.920
		0.8	0.870	0.890	0.870	1.000
		1.0	0.870	0.890	0.865	0.985
Qwen2.5-14B	Base	0.6	0.650	0.610	0.625	0.905
		0.8	0.650	0.610	**0.645**	**0.995**
		1.0	0.650	0.610	0.645	0.905
	Vernier	0.6	0.910	0.935	0.935	0.955
		0.8	0.910	0.935	0.915	0.995
		1.0	0.910	0.935	0.885	0.925
Llama-3.1-8B	Base	0.6	0.540	0.525	0.530	0.980
		0.8	0.540	0.525	0.530	0.980
		1.0	0.540	0.525	0.540	1.000
	Vernier	0.6	0.945	0.960	0.945	1.000
		0.8	0.945	0.960	0.945	1.000
		1.0	0.945	0.960	0.945	0.980

Table A3: Final-layer mechanism diagnostics across models, base versus Vernier. Cosine is last-token hidden-state cosine $\cos(h_L(P_0), h_L(P_1))$ at final layer L. Logit-lens disagreement is $\langle |p_\theta(y \mid P_0) - p_\theta(y \mid P_1)| \rangle$. Probe transfer is linear-probe accuracy on final-layer P_1 hidden states.

	Hidden cosine			Logit-lens gap			Probe transfer		
Model	Base	Vernier	Δ	Base	Vernier	Ratio	Base	Vernier	Δ
Phi-3-mini (3.8B)	0.990	0.990	0.000	0.044	0.044	1.0×	0.615	0.655	+0.040
Qwen2.5-7B	0.963	**0.995**	**+0.032**	0.155	**0.028**	**5.5×**	0.480	**0.695**	**+0.215**
Qwen2.5-14B	0.979	0.995	+0.016	0.173	0.035	5.0×	0.590	0.680	+0.090
Qwen2.5-32B	0.960	0.992	+0.032	0.136	0.044	3.1×	0.585	0.705	+0.120

Table A4: Mechanism diagnostics for augmentation without KL and full Vernier (held-out CLadder, $n = 200$). Augment is the $\beta = 0$ twin-task adapter. Probe transfer is final-layer variable-name probe accuracy on P_1 hidden states, and logit-lens gap is final-layer answer-token disagreement.

Model	Diagnostic	Base	Augment (β=0)	Vernier (β=1)
Qwen2.5-7B	Probe transfer (↑)	0.480	**0.690**	0.695
	Logit-lens gap (↓)	0.155	0.074	**0.028**
Qwen2.5-14B	Probe transfer (↑)	0.590	**0.730**	0.680

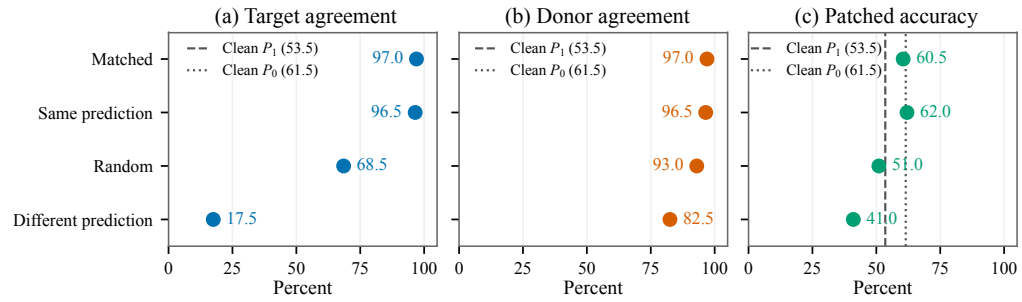

Figure A2: Donor controls on base Qwen-7B (CLadder, $n = 200$, relative depth 0.8). Matched donors use the same item's original-view state; the other groups use unrelated donors. Panels report agreement with the target prediction, agreement with the donor prediction, and patched accuracy. Same/different prediction groups use clean model predictions rather than gold labels. Dashed and dotted reference lines give clean accuracies for this batch. Values are listed in Table A6.

Table A6: Recorded Qwen-7B CLadder donor controls ($n = 200$, depth 0.8). Target agreement refers to the target item's clean P_0 prediction; donor agreement refers to the source item's prediction. Both are unconditional agreement rates, not conditional flip rates. Base accuracies belong to this patching batch.

Setting	Donor	P_0	P_1 clean	P_1 patched	Target agr.	Donor agr.
Base	matched	0.615	0.535	0.605	0.970	0.970
Base	random	0.615	0.535	0.510	0.685	0.930
Base	same pred.	0.615	0.535	0.620	0.965	0.965
Base	different pred.	0.615	0.535	0.410	0.175	0.825

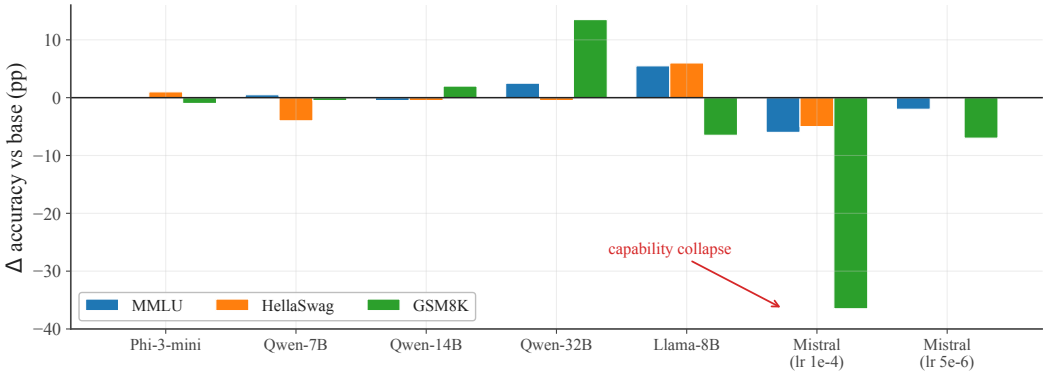

Figure A3: General-capability change after Vernier on MMLU (Hendrycks et al., 2021), HellaSwag, and GSM8K ($n = 200$ items each under a fixed seed), reported as Δ accuracy in percentage points versus base. Qwen2.5-32B is evaluated with QLoRA. The Qwen and Phi rows show no broad degradation, whereas Mistral-7B under the canonical learning rate collapses on GSM8K; the rescued low-rate recipe (step 150) recovers most of the loss. The per-benchmark numbers are in Table A7.

Table A7: General-capability evaluation on MMLU (Hendrycks et al., 2021), HellaSwag, and GSM8K ($n = 200$ items each under a fixed seed). Qwen2.5-32B is evaluated with QLoRA. The two Mistral-7B rows compare the default and rescued training recipes.

	MMLU			HellaSwag			GSM8K		
Model	Base	Vernier	Δ	Base	Vernier	Δ	Base	Vernier	Δ
Phi-3-mini (3.8B)	0.655	0.655	0.0	0.795	0.805	+1.0	0.815	0.805	−1.0
Qwen2.5-7B	0.685	0.690	+0.5	0.830	0.790	−4.0	0.705	0.700	−0.5
Qwen2.5-14B	0.780	0.775	−0.5	0.855	0.850	−0.5	0.820	0.840	+2.0
Qwen2.5-32B	0.825	0.850	+2.5	0.900	0.895	−0.5	0.740	0.875	+13.5
Llama-3.1-8B	0.565	0.620	+5.5	0.645	0.705	+6.0	0.830	0.765	−6.5
Mistral-7B-v0.3, lr=1e-4[†]	0.525	0.465	−6.0	0.650	0.600	−5.0	0.525	0.160	−36.5
Mistral-7B-v0.3, lr=5e-6, step 150	0.525	0.505	−2.0	0.650	0.650	0.0	0.525	0.455	−7.0

Table A8: Lexical-overlap contamination check for the capability benchmarks. The 1500 CLadder training prompts are compared with the 200-item MMLU, HellaSwag, and GSM8K samples using long n-gram overlap.

Benchmark	**Items**	**8-gram overlap**	**13-gram overlap**
MMLU	200	0.000	0.000
HellaSwag	200	0.000	0.000
GSM8K	200	0.000	0.000

Table A9: Hyperparameter sweeps. Top: KL vs MSE consistency form on Qwen-7B. Middle: LoRA rank on Qwen-7B. Bottom: β sweep on Qwen-14B. Best per column in each block in bold.

Configuration	**CLadder Gap**	**CRASS Gap**	**e-CARE Gap**	**Avg. OOD reduction**
Consistency form on Qwen-7B ($\beta = 1$, $r = 16$):				
Symmetric KL (default)	**−2.0**	**+25.0**	**+17.5**	**−25.2%**
MSE on answer-subspace logits	−1.5	+24.5	+19.0	−22.8%
LoRA rank on Qwen-7B ($\beta = 1$, symmetric KL, $\alpha_{\text{LoRA}} = 2r$):				
$r = 4$	−2.0	+26.0	+20.0	−18.3%
$r = 8$	**−4.5**	+23.5	+22.0	−17.6%
$r = 16$ (default)	−2.0	+25.0	**+17.5**	**−25.2%**
$r = 32$	−3.0	**+21.5**	+22.5	−19.5%
β sweep on Qwen-14B ($r = 16$, symmetric KL):				
$\beta = 0.5$	**−1.5**	**+18.5** (-31%)	**+18.5** (-14%)	**−22%**
$\beta = 1.0$ (default)	−1.5	+20.5 (-24%)	+21.0 (-2%)	−13%
$\beta = 2.0$	+2.0	+18.5 (-31%)	+22.0 (+2%)	−14%

Table A10: Adversarial-placeholder robustness on held-out CLadder ($n = 200$). Each Vernier adapter is trained only on the canonical $X_1, X_2, \ldots$ scheme (top block) and evaluated under three held-out schemes the adapter has never seen.

		Phi-3-mini			Qwen-7B			Qwen-14B		
Scheme	**Setting**	P_0	P_1	Gap	P_0	P_1	Gap	P_0	P_1	Gap
$X_1, X_2, \ldots$	Base	0.585	0.540	+4.5	0.625	0.545	+8.0	0.615	0.580	+3.5
	Vernier	0.550	0.550	0.0	0.870	0.890	−2.0	0.920	0.935	−1.5
`foo, bar, baz, ...`	Base	0.585	0.575	+1.0	0.625	0.555	+7.0	0.615	0.605	+1.0
	Vernier	0.550	0.575	−2.5	0.870	0.855	+1.5	0.920	0.915	+0.5
`alpha_a, alpha_b, ...`	Base	0.585	0.565	+2.0	0.625	0.540	+8.5	0.615	0.590	+2.5
	Vernier	0.550	0.570	−2.0	0.870	0.845	+2.5	0.920	0.870	+5.0
`VAR_A, VAR_B, ...`	Base	0.585	0.550	+3.5	0.625	0.515	+11.0	0.615	0.530	+8.5
	Vernier	0.550	0.545	+0.5	0.870	0.870	0.0	0.920	0.775	+14.5

Table A11: Setting A (CLadder-only training) versus Setting B (joint training on CLadder, CRASS, and e-CARE) on Qwen2.5-7B and Qwen2.5-14B. Setting B uses 5,154 pooled training items, single seed. Best CRASS and best e-CARE per model in bold.

		Gap (pp)			
Model	**Setting**	**CLadder**	**CRASS**	**e-CARE**	Δ_{OOD}
Qwen2.5-7B	Base	$+8.0$	$+34.0$	$+23.0$	—
	Setting A	$\mathbf{-2.5_{\pm 0.9}}$	$+24.8_{\pm 0.3}$	$+18.8_{\pm 1.5}$	-22%
	Setting B	$+0.5$	$\mathbf{+5.5}$	$\mathbf{+16.5}$	$\mathbf{-56}\%$
Qwen2.5-14B	Base	$+3.5$	$+27.0$	$+21.5$	—
	Setting A	$\mathbf{-1.5}$	$+20.5$	$+21.0$	-13%
	Setting B	$+1.0$	$\mathbf{+5.5}$	$\mathbf{+14.0}$	$\mathbf{-56}\%$

Table A12: Paired statistical tests for Qwen2.5-7B. Gaps are $P_0 - P_1$ in percentage points. CIs are paired bootstrap 95% intervals over items. McNemar tests whether the paired P_0/P_1 correctness asymmetry differs from zero. The final column is the paired reduction in gap relative to the base row for the same benchmark.

Benchmark	**Setting**	n	P_0	P_1	**Gap**	**Gap CI**	**McNemar** p	**Δ gap vs base**
CLadder	Base	200	62.5	54.5	+8.0	[+3.0, +13.0]	0.004	–
CLadder	Vernier	200	87.0	89.0	-2.0	[-6.0, +2.0]	0.454	+10.0 [+3.5, +17.0]
CRASS	Base	200	84.5	50.5	+34.0	[+26.5, +41.5]	$< 10^{-4}$	–
CRASS	Vernier	200	88.5	63.5	+25.0	[+18.0, +32.0]	$< 10^{-4}$	+9.0 [+3.0, +15.5]
e-CARE	Base	200	81.5	58.5	+23.0	[+15.5, +30.5]	$< 10^{-4}$	–
e-CARE	Vernier	200	81.0	63.5	+17.5	[+10.0, +25.0]	$< 10^{-4}$	+5.5 [+0.5, +11.0]

Table A13: Larger CLadder held-out evaluation ($n = 1000$). Gaps are $P_0 - P_1$ in percentage points with paired bootstrap 95% intervals. The final column reports paired gap reduction relative to the corresponding base model.

Model	**Setting**	n	P_0	P_1	**Gap**	**Gap CI**	**Δ gap vs base**
Qwen2.5-7B	Base	1000	55.8	53.4	+2.4	[+0.0, +4.8]	–
Qwen2.5-7B	Vernier	1000	84.2	85.5	-1.3	[-3.0, +0.4]	+3.7 [+0.7, +6.7]
Qwen2.5-14B	Base	1000	56.8	54.0	+2.8	[+0.5, +5.2]	–
Qwen2.5-14B	Vernier	1000	86.6	87.9	-1.3	[-2.7, +0.1]	+4.1 [+1.3, +6.8]

Table A14: Larger out-of-distribution evaluation for Qwen2.5-7B and Qwen2.5-14B. CRASS uses all 274 prepared items, and e-CARE uses a 1000-item sample. The adapters are trained only on CLadder. Gap reductions are relative to the corresponding base row.

Model	**Benchmark / Setting**	n	P_0	P_1	**Gap**	**Δ Gap**
Qwen2.5-7B	CRASS Base	274	0.839	0.489	$+35.0$	–
	CRASS Vernier	274	0.876	0.639	$\mathbf{+23.7}$	$\mathbf{-32}\%$
	e-CARE Base	1000	0.802	0.595	$+20.7$	–
	e-CARE Vernier	1000	0.808	0.621	$\mathbf{+18.7}$	$\mathbf{-10}\%$
Qwen2.5-14B	CRASS Base	274	0.894	0.620	$+27.4$	–
	CRASS Vernier	274	0.923	0.755	$\mathbf{+16.8}$	$\mathbf{-39}\%$
	e-CARE Base	1000	0.824	0.592	$+23.2$	–
	e-CARE Vernier	1000	0.830	0.606	$+22.4$	-3%

Table A15: Base-model CLadder lexical gaps on a larger held-out slice ($n = 1000$). Gaps are $P_0 - P_1$ in percentage points with paired bootstrap 95% intervals, and p is the exact two-sided McNemar test on paired correctness.

Model	P_0	P_1	Gap	95% CI	McNemar p
Qwen2.5-7B	0.553	0.534	$+1.9$	$[-0.6, +4.4]$	0.151
Qwen2.5-14B	0.564	0.549	$+1.5$	$[-0.9, +3.9]$	0.255
Qwen2.5-32B	0.568	0.558	$+1.0$	$[-1.3, +3.3]$	0.440
Llama-3.1-8B	0.530	0.508	$+2.2$	$[-1.0, +5.3]$	0.194

Table A16: Random stratified CLadder held-out splits on Qwen2.5-7B. Each run trains a fresh adapter on 1500 training items and evaluates on a disjoint 200-item held-out set stratified by rung. Values are mean $\pm$ standard deviation across split seeds $\{7, 42, 123\}$.

Aggregate	Base Gap	Vernier Gap	Δ Gap
Mean $\pm$ std	+6.2 $\pm$ 5.2	**+0.2 $\pm$ 1.4**	**+6.0 $\pm$ 5.0**

Table A17: Capacity and mechanism sweep on held-out CLadder ($n = 200$). The gap is $P_0 - P_1$. "Probe Δ" is the change in final-layer P_1 probe-transfer accuracy, and "lens ratio" is the base-to-Vernier logit-lens disagreement ratio. Small Qwen rows (†) are three-seed means $\pm$ std, and other rows are single-seed.

Model	Params	Base Gap	Vernier Gap	Probe Δ	Lens Ratio	Verdict
Qwen2.5-1.5B†	1.5B	-9.0	+3.5	-0.09	1.1×	fail
Qwen2.5-3B†	3B	+0.5	+0.5	**+0.29**	2.1×	**success**
Phi-3-mini	3.8B	+4.5	+0.0	+0.04	1.0×	fail
Gemma-2-2B	2B	+9.5	-2.0	+0.07	0.4×	mixed
Qwen2.5-7B	7B	+8.0	-2.5	+0.22	5.5×	success
Qwen2.5-14B	14B	+3.5	+0.5	+0.09	5.0×	success
Qwen2.5-32B	32B	-2.0	+0.5	+0.12	3.1×	success

Table A18: e-CARE final-layer decision-token alignment. Cosine is $\cos(h_L(P_0), h_L(P_1))$ at the last non-pad token. The remaining e-CARE behavioural gap is therefore not explained by a failure to move the two final-layer representations closer together.

Model	Base Cosine	Vernier Cosine	Δ
Qwen2.5-7B	0.904	0.923	+0.019
Qwen2.5-14B	0.917	0.936	+0.020
Qwen2.5-32B	0.902	0.924	+0.021

Table A19: Preliminary non-causal BBH checks with object names renamed to typed placeholders. Logical_deduction uses $n = 100$ held-out items; colored_objects uses a smaller $n = 20$ diagnostic slice. Augmentation closes or reduces the gap with both views rising, and activation patching shows the decision-token representation is a control point in the tested cells.

		Base			After Augmentation		
Task	**Model**	P_0	P_1	**Gap**	P_0	P_1	**Gap**
Logical deduction	Qwen-7B	0.740	0.650	+9.0	0.950	0.950	**+0.0**
	Qwen-14B	0.900	0.890	+1.0	0.990	0.980	+1.0
Colored objects	Qwen-7B	0.500	0.300	+20.0	0.800	0.850	**-5.0**
	Qwen-14B	0.750	0.300	+45.0	0.850	0.650	+20.0
Activation patching (P_0 decision-token state copied into P_1)							
logical deduction, Qwen-7B, depth 0.8: P_1 0.63 → **0.77**, flip-to-P_0 **0.89**							
colored objects, Qwen-14B, depth 0.8: P_1 0.25 → **0.75**, flip-to-P_0 **1.00**							

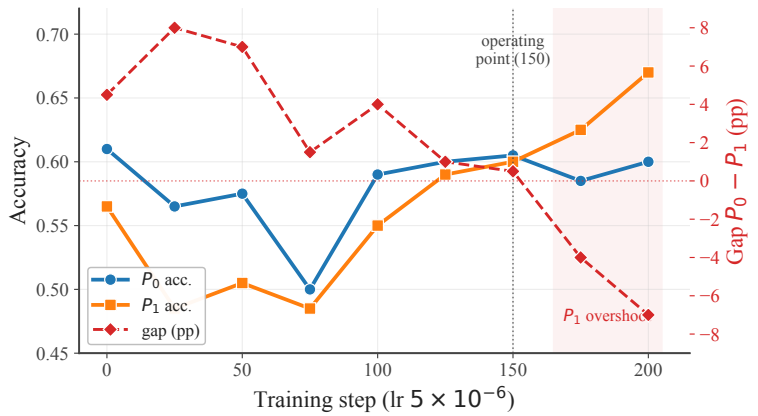

Figure A4: Mistral-7B-Instruct-v0.3 checkpoint trajectory under the rescued recipe (lr 5×10^{-6}) on the in-distribution CLadder held-out slice. P_0 and P_1 accuracy (left axis) and the signed gap $P_0 - P_1$ (right axis) are plotted against the training step. Step 150 is the operating point reported in Table 1 and fig. A3; later checkpoints overshoot into a negative gap as P_1 overtakes P_0. The per-step numbers are in Table A20.

Table A20: Mistral-7B-Instruct-v0.3 checkpoint trajectory under the rescued recipe (lr=5×10^{-6}) on the in-distribution CLadder held-out slice. Step 150 is the operating point reported in Tables 1 and A7.

Step	P_0	P_1	**CLadder gap**	**Comment**
Base	0.610	0.565	+4.5	—
25	0.565	0.485	+8.0	no learning yet
50	0.575	0.505	+7.0	
75	0.500	0.485	+1.5	
100	0.590	0.550	+4.0	partial closure
125	0.600	0.590	+1.0	
150	0.605	0.600	**+0.5**	sweet spot, reported in Tables 1 and A7
175	0.585	0.625	-4.0	P_1 overshoot
200	0.600	0.670	-7.0	P_1 overshoot

C PRE-SPECIFICATION, RECIPE AUDIT, AND FALSIFICATION SUMMARY

The reference recipe specifies attention-projection LoRA with rank 16, $\alpha = \beta = 1$, learning rate 10^{-4}, and a nominal 500-step checkpoint. The historical summaries include model-specific learning-rate, quantization, and checkpoint variants. Step numbers identify reported checkpoints rather than establishing a common optimizer schedule; repeated-run standard deviations summarize the available runs, without assigning unverified seed identities. For Phi-3, prediction-file adapter paths identify attention checkpoint 70; both attention and MLP logs span 100 optimizer steps with matching learning-rate trajectories. The available records do not establish equality of every remaining configuration field or training member, so this is a trajectory comparison rather than an isolated projection-target effect. The Mistral lower-rate table reports step 150 as an operating point; the saved trajectory supports this descriptive comparison but does not establish a prospective selection rule.

The complete Qwen-7B execution record provides a separate, traceable reference: seed 42, 1,500 examples, 500 total steps, a saved step-500 adapter, exclusion of training question IDs, and paired evaluation on 900 questions. The training and evaluation commands, question IDs, and prediction records are provided in the evidence supplement. Base and adapter use the same evaluation script and scoring configuration. This provenance applies to that comparison and is not transferred to the historical multi-model or ablation summaries.

Table A21 organizes the observations by the question each measurement can answer. The controls constrain interpretation without assuming statistically independent tests or an exhaustive choice between information loss and misalignment.

Table A21: Diagnostic questions and interpretation controls. Each row specifies the observable property and a comparison that limits its interpretation.

Question	**Interpretation control**	**Evidence**
Do both views improve?	Component accuracies alongside the gap	Table 1
Does a lexical decoder transfer?	Proxy labels are not causal-fact annotations	Table A3
Do projected answer scores agree?	Projection-dependent logit-lens measurement	Table A3
What differs across recipes?	Training membership and schedule matching	Table A1 and section C
Does patching require the matched item?	Unrelated answer-matched donors	Figure A2

D PHI-3 WITH MLP-TARGETED LORA

The attention-targeted Phi-3 checkpoint exhibits a small lexical gap with original-view regression. We examine whether the same behavioural pattern appears along an MLP-targeted trajectory, using `gate_up_proj` and `down_proj`. Table A22 compares attention checkpoint 70 with MLP checkpoints at steps $\{25, 50, 75, 100\}$. Both recorded runs use a 100-step learning-rate schedule with a peak of 5×10^{-6}, 50 warmup steps, and cosine decay. The table describes their measured trajectories without assuming all other training conditions are identical.

Table A22: Phi-3-mini with the LoRA target switched from attention to MLP projections (`gate_up_proj`, `down_proj`), keeping all other hyperparameters identical. The default attention-LoRA row is included for reference.

LoRA target	**Step**	P_0	P_1	**CLadder gap**	**OOD reduction**
Base (no adapter)	—	0.585	0.540	+4.5	—
Attention projections (`q, k, v, o_proj`), default Vernier:					
`attn`	500	0.550	0.550	0.0	CRASS -3%, e-CARE -10%
MLP projections (`gate_up_proj, down_proj`):					
`mlp`	25	0.580	0.545	+3.5	—
`mlp`	50	0.570	0.535	+3.5	CRASS -3%, e-CARE -7%
`mlp`	75	0.510	0.490	+2.0	CRASS +2%, e-CARE -5%
`mlp`	100	0.490	0.490	0.0	CRASS -5%, e-CARE -2%

Table A24: A worked CLadder example. Gold answer $y^\star =$ no, Pearl rung observational. Underlined surface forms in q are mapped by T to placeholders. Numeric quantities, query, and answer are unchanged.

Original prompt q	**Perturbed prompt** $T(q)$
The overall probability of alarm set by husband is 77%. For husbands that don't set the alarm, the probability of ringing alarm is 26%. For husbands that set the alarm, the probability of ringing alarm is 76%. Is ringing alarm less likely than silent alarm overall? Answer (yes or no):	The overall probability of X_3 set by X_2 is 77%. For X_2 that don't set the X_3, the probability of ringing X_3 is 26%. For X_2 that set the X_3, the probability of ringing X_3 is 76%. Is ringing X_3 less likely than silent X_3 overall? Answer (yes or no):

Along the MLP trajectory, the CLadder gap narrows to zero while accuracy falls to 49.0% on both views. The reported CRASS and e-CARE gaps remain within 2 pp of their base values. These checkpoints reproduce the behavioural pattern seen in the attention reference: a smaller lexical gap need not correspond to improved task accuracy. The observation does not identify a unique optimization mechanism or imply that target-module choice has no effect.

E PER-RUNG CLADDER BREAKDOWN

The held-out CLadder results broken down by Pearl-ladder rung are reported in Table A23.

Table A23: Per-rung CLadder breakdown ($n = 200$ held-out), grouped by model so the three rungs can be compared within each model. Rung n is the count of held-out items at that rung. "flip" marks a base gap that reverses sign after Vernier (relative reduction undefined). Δ_{gap} is the relative gap reduction versus base. Within each model, **bold** marks the best and underline the second best across the three rungs per column (P_0/P_1 higher is better; Gap closest to 0 is better). Δ_{gap} is shown but not ranked, because the per-rung base gaps differ and flips and overcorrection make it non-monotonic.

		Base			Vernier			
Model	**Rung** (n)	P_0	P_1	**Gap**	P_0	P_1	**Gap**	**Rel. gap**
Phi-3-mini	Observational (75)	0.547	0.520	**+2.7**	0.520	0.547	-2.7	flip
	Interventional (57)	**0.667**	**0.614**	+5.3	**0.596**	**0.614**	**-1.8**	flip
	Counterfactual (68)	0.559	0.500	+5.9	0.544	0.500	+4.4	-25%
Qwen2.5-7B	Observational (75)	0.587	0.547	**+4.0**	0.880	0.853	**+2.7**	-32%
	Interventional (57)	**0.684**	**0.579**	+10.5	**0.947**	**1.000**	-5.3	-150%
	Counterfactual (68)	0.618	0.515	+10.3	0.794	0.838	-4.4	-143%
Qwen2.5-14B	Observational (75)	0.587	0.573	**+1.3**	0.893	0.907	**-1.3**	flip
	Interventional (57)	**0.632**	**0.596**	+3.5	**0.947**	**1.000**	-5.3	-251%
	Counterfactual (68)	**0.632**	0.574	+5.9	0.926	0.912	+1.5	-75%

On both Qwen models the largest baseline gap is on the interventional and counterfactual rungs and the smallest is on the observational rung. After Vernier the interventional P_1 accuracy reaches 1.000 on both Qwen models, the largest single-cell accuracy improvement reported in the paper, describing a larger observed gain on these cells without identifying its internal cause. Phi-3 shows the opposite pattern. The interventional rung gap changes sign through P_0 regression, with P_0 falling from 0.667 to 0.596 while P_1 stays at 0.614. This shows why rung-level accuracies should accompany aggregate gap measurements.

F WORKED LEXICAL PERTURBATION

Table A24 illustrates the lexical transformation used to construct paired views.

G TRAINING PROCEDURE

Algorithm 1 summarizes the paired training update. Task cross-entropy is normalized over the full vocabulary, whereas consistency is computed over the candidate answer-token IDs.

Algorithm 1 Vernier training: one update step

Input: Frozen base model f_θ; LoRA parameters ϕ; answer-token ID set $\mathcal{A}$; batch $\mathcal{B} = \{(q_i, T(q_i), t_i^\star)\}_{i=1}^{B}$; loss weights α, β; AdamW optimizer state s.
Output: Updated LoRA parameters ϕ' and optimizer state s'.

$P_0 \leftarrow \textsc{TokenizeAndPad}(\{q_i\}_{i=1}^{B})$
$P_1 \leftarrow \textsc{TokenizeAndPad}(\{T(q_i)\}_{i=1}^{B})$
for $v \in \{0, 1\}$ **do**
    $Z^{(v)} \leftarrow f_{\theta,\phi}(P_v)$
    $\mathbf{z}^{(v)} \leftarrow \textsc{GatherLastNonPaddingLogits}(Z^{(v)}, P_v)$
    $\mathbf{p}^{(v)} \leftarrow \mathrm{softmax}\big(\mathbf{z}^{(v)}_{\cdot,\mathcal{A}}\big)$
end for

$$\mathcal{L}_{\text{task}} \leftarrow \frac{1}{B}\sum_{i=1}^{B}\Big[\mathrm{CE}(\mathbf{z}_i^{(0)}, t_i^\star) + \mathrm{CE}(\mathbf{z}_i^{(1)}, t_i^\star)\Big]$$

▷ Cross-entropy over the full vocabulary $\mathcal{V}$

$$\mathcal{L}_{\text{cons}} \leftarrow \frac{1}{2B}\sum_{i=1}^{B}\Big[\mathrm{KL}(\mathbf{p}_i^{(0)} \| \mathbf{p}_i^{(1)}) + \mathrm{KL}(\mathbf{p}_i^{(1)} \| \mathbf{p}_i^{(0)})\Big]$$

$\mathcal{L} \leftarrow \alpha\mathcal{L}_{\text{task}} + \beta\mathcal{L}_{\text{cons}}$
$g \leftarrow \nabla_\phi \mathcal{L}$
$(\phi', s') \leftarrow \textsc{AdamWStep}(\phi, g, s)$
return (ϕ', s')