# FACTUAL FINE-TUNING SEPARATES ABSTENTION FREQUENCY FROM SCORE SELECTIVITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Selecting fine-tuning examples that a language model already answers correctly can reduce its exposure to unfamiliar facts, yet it remains unclear which aspects of reliable answering this selection preserves. Observed abstention, ranking of unreliable answers, and separation of initially known and unknown questions can change differently, making refusal rate an incomplete evaluation of factual adaptation. We examine this distinction in closed-book PopQA and TriviaQA using three instruction-tuned model families and 72 training runs across three adapter ranks. Repeated initial-model answers define historical known and unknown panels; known-fact, unknown-fact, and high-loss known-fact training share an answer-only objective, while held-out question trajectories distinguish lost correct answers from abstentions that become errors. Known-fact training converts many prior abstentions into incorrect answers, and unknown-fact training introduces additional costs whose origins depend on the model's initial behavior. Relative to known-fact training, it adds 14.53 and 12.17 percentage points of correct-to-error transitions in Llama and Qwen, averaged across datasets, prompts, and seeds, whereas OLMo's lower final accuracy primarily reflects a smaller improvement from a highly abstaining starting point. Qwen nearly stops abstaining under both arms, but its abstention score distinguishes them: the area under the receiver operating characteristic curve (AUROC) for current answered-response correctness after known/unknown training is 0.873/0.426 on PopQA under one prompt and 0.880/0.596 under another. This distinction extends beyond historical labels, although ranking alone does not establish a calibrated abstention policy. At rank 16, hidden-state classifiers kept fixed or refitted after training show that historical-label transfer depends on the elicitation prompt, limiting conclusions about internal knowledge from a single probe. These results identify what an abstention-frequency comparison misses: the origins of new errors and the ranking quality of an abstention score. Separate evaluation shows data curators both the benefits and the remaining failures of known-fact selection.

## 1 INTRODUCTION

A factual assistant should answer questions it can support and abstain when an answer would be unreliable. Supervised fine-tuning (SFT) often rewards an answer to every training question, including questions the initial model cannot answer. Prior studies show that such unfamiliar supervision increases hallucination, changes refusal behavior, and can damage previously correct answers (Gekhman et al., 2024; Kang et al., 2025; Kaplan et al., 2026). Selecting training targets that agree with the model's existing knowledge is therefore a natural intervention, and recent methods improve both factuality and refusal through knowledge-aligned targets (Becker et al., 2026). For a developer choosing factual SFT data, the remaining question is whether known-fact selection preserves the different aspects of abstention that matter at evaluation time.

A low abstention rate cannot distinguish indiscriminate answering from answering with a useful error-ranking score. A hidden-state probe is a classifier trained to predict a target from internal activations; refitting it after SFT can recover separation even when the original classifier transfers poorly. Abstention frequency, score ranking, and probe transfer therefore measure different properties. We use *score selectivity* to mean ranking quality for a specified binary target, measured by the area under the receiver operating characteristic curve (AUROC).



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

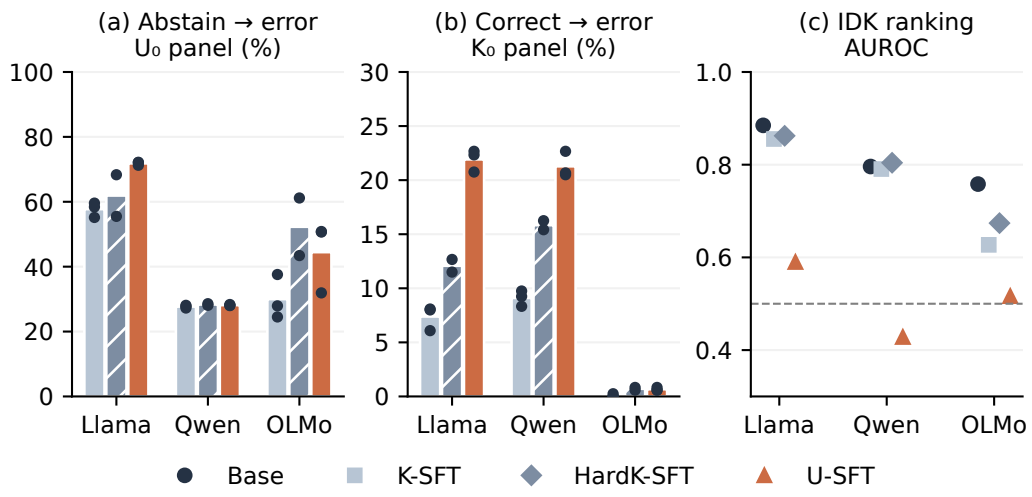



Figure 1: K/U training differs in error routes and score ranking at rank 16. (a,b) Transition rates on historical unknown ($U_0$) and known ($K_0$) panels; dots show seeds. (c) The "I don't know" (IDK) score ranks $U_0$ above $K_0$ under a natural-answer prompt. Datasets have equal weight; (a,b) average both natural prompts. K/U use three seeds and HardK two. Table 4 gives dataset-specific rankings.

We study these measurements together under answer-only SFT using low-rank adaptation (LoRA). For each of three model families, we construct known-fact (K-SFT) and unknown-fact (U-SFT) training sets using the initial model's answers, and evaluate the same held-out questions across training. The arms share their example count, prompt, optimizer, and schedule; the facts themselves differ, so this comparison does not isolate unknownness from every property of the selected facts. A third arm uses high-loss known facts (HardK-SFT) as a partial difficulty control. Across 24 rank-16 runs and 48 matched-rank extensions, we link initial and later responses to distinguish new errors from corrected errors, then compare score rankings and fixed/refit probes.

Our contributions are as follows.

- We characterize a failure of known-fact selection under answer-only LoRA: K-SFT already converts many previously abstained questions into errors. In Qwen, it leaves only 0.44% of historically unknown questions abstaining. Yet K-SFT and U-SFT have substantially different abstention-score rankings for both historical labels and the checkpoints' resolved answer correctness (Section 4).
- We show why the initial response changes the interpretation of an accuracy gap. U-SFT adds 14.53 and 12.17 percentage points of correct-to-error transitions over K-SFT in Llama and Qwen. OLMo starts by abstaining on most historically known questions, and its lower U-SFT accuracy largely reflects a smaller gain from that starting point. The Llama/Qwen contrast recurs at ranks 8 and 32 (Section 4.2).
- We identify prompt-dependent limits of probe transfer. Qwen's base probe transfers poorly after SFT under a binary action prompt but much better under a natural-answer prompt; refitting increases separation in the former. Surface cues also predict historical labels, restricting knowledge-specific interpretations (Section 5).

Figure 1 summarizes the error routes and rankings that an abstention-rate comparison misses.

## 2 Related Studies

**Factual supervision and abstention.** Unfamiliar facts are learned more slowly and can increase hallucinations during SFT (Gekhman et al., 2024); their supervision also shapes responses to unfamiliar queries (Kang et al., 2025). Factual extraction can deteriorate even when the facts occurred in pretraining (Ghosal et al., 2024), and later studies analyze forgetting and interference (Kaplan

Table 1: Factual QA sources used without evidence documents. Each model–dataset evaluation panel contains 300 historical known and 300 historical unknown questions.

| Dataset | Question structure | Role in the comparison |
|---|---|---|
| PopQA | Entity-centric facts with subject popularity and relation metadata | Tests factual adaptation where popularity is a measurable alternative explanation. |
| TriviaQA | Trivia questions with varied wording | Tests the contrasts beyond entity-centric QA, with lexical and input-embedding controls. |

et al., 2026; Dang et al., 2026). Thus historical failure to answer does not establish absence from pretrained knowledge. R-Tuning teaches refusal on estimated unknowns (Zhang et al., 2024); US-Tuning additionally trains sensitivity to uncertainty instructions in contextual QA (Li et al., 2025). SEAT uses sparse adaptation and regularization to preserve epistemic abstention across synthetic and real queries (Shen et al., 2025). Knowledge-aligned target rewriting and latent-boundary data selection also evaluate factuality and refusal (Becker et al., 2026; An et al., 2026). These interventions differ from retaining known short-answer examples with an unchanged answer-only objective, the selection tested here.

**Ranking and selective answering.** Self-evaluation research distinguishes confidence in a proposed answer from predicting whether a question is answerable (Kadavath et al., 2022). Confidence can be explicitly supervised or elicited verbally, and its calibration depends on the elicitation procedure (Lin et al., 2022; Tian et al., 2023). Selective prediction evaluates both coverage, the fraction answered, and selective risk, the error rate among those answers (Geifman & El-Yaniv, 2017). Seal already learns a rejection-token probability, uses it in decoding, and measures known/unknown separation alongside answering and refusal (Huang et al., 2025). Our equal-size K/U answer-only comparison tests a different question: whether selecting known facts preserves ranking when both training arms suppress abstention. The exact-phrase score studied here diagnoses ranking; it does not by itself establish a selective answering policy.

**Probe transfer and causal use.** Hidden-state detectors predict answer correctness (Orgad et al., 2025), while training-free query-level methods estimate uncertainty before generation (Chen et al., 2026). Probe controls distinguish representational information from classifier learning (Hewitt & Liang, 2019); decoding alone does not establish use by the model (Belinkov, 2022). Interventions on entity features and confidence can affect answers (Ferrando et al., 2025; Kumaran et al., 2026), and cross-model interventions can test fine-tuning mechanisms (Prakash et al., 2024); intervention-induced artifacts still require controls (Grant et al., 2026). After SFT, Shi et al. (2026) find retained fixed/refit separation of answerable and false-premise questions in Llama despite degraded refusal. We test the scope of transfer conclusions using historical model-specific labels, two elicitation formats, and surface baselines, without inferring a causal mechanism from probe accuracy.

## 3 Experimental Design

We test whether known-fact training retains abstention (Section 4.1), where additional errors arise (Section 4.2), and whether score ranking and probe transfer follow the same pattern (Sections 4.3–5).

### 3.1 Task and Model Scope

We use short, closed-book factual question answering: the model receives a question without retrieved documents and answers or abstains. Reference answers permit direct correctness evaluation, while withholding documents removes retrieval quality as an explanation for training effects. PopQA supplies entity-centric questions and popularity metadata (Mallen et al., 2023); TriviaQA supplies questions written by trivia enthusiasts (Joshi et al., 2017). Their constructions test dataset-specific cues (Table 1). All selected questions have reference answers; "unknown" describes a model's answering history, rather than a false premise.

We evaluate Llama-3.1-8B-Instruct, Qwen3.5-9B, and Olmo-3-7B-Instruct (Team Olmo et al., 2025). These similarly sized open checkpoints permit weight, probability, and residual-state anal-

yses across families. The initial instruction-tuned checkpoint is the *base model*. Transitions need saved responses; scores need token probabilities, and probes need hidden states.

### 3.2 Historical Labels and Leakage-Controlled Splits

A single failure can reflect sampling or phrasing. We therefore use two shortest-answer prompts without an abstention option, with one greedy answer and four temperature-0.5 samples per prompt (top-$p = 1$, no top-$k$ filtering). The historical known label $K_0$ requires both greedy answers and at least seven of the eight samples to be correct; the historical unknown label $U_0$ requires zero correct answers across all ten. Excluding intermediate cases creates separated panels, without establishing permanent knowledge or estimating ordinary-use answerability prevalence.

The forced-answer criterion separates labeling from the later opportunity to abstain: natural prompts can elicit abstentions on $K_0$ and correct answers on $U_0$. Freezing labels prevents class reassignment from producing an apparent change in score separation. Cross-model comparisons concern each model's own historical panels.

Within each model and dataset, TRAIN supplies 500 examples per label, a separate HardK set supplies 300 known examples, CAL-FIT and CAL-SELECT each contain 75 examples per label, and EVAL contains 300 per label. The combined training pool, CAL-FIT, CAL-SELECT, and EVAL are group-disjoint: links through normalized primary answers, PopQA subjects, recorded duplicate facts, or normalized questions stay within a partition, including across datasets. CAL-FIT fits probes and CAL-SELECT chooses layers; EVAL selects neither. Balanced EVAL counts give both labels equal weight; reported rates apply to these selected panels. Initial correctness uses semantic reference judgments and recorded AI-assisted adjudication (Appendix B).

### 3.3 Training Contrasts and Optimization

**Each arm addresses a different alternative.** K-SFT tests whether selecting facts the base model already answers is sufficient to retain abstention under answer-only training. U-SFT changes the training set's historical label while retaining the example count, prompt, gold-answer supervision, optimizer, and schedule. HardK-SFT tests whether difficult known examples partially reproduce the U-SFT pattern. HardK matches 300 questions per dataset from the half of 800 known candidates with highest base-model answer loss to 300 unknown examples, using length and available PopQA metadata. The remaining 500 known questions form K-SFT; subsequent pairing with U changes their order, not membership. This complementary selection leaves substantial imbalance: PopQA answer-length standardized differences between U and K range from 0.628 to 0.877 (Appendix C). These selected regimes confound unknownness with content and difficulty.

**A common answer-only objective.** Each example uses the shortest-answer prompt and the canonical reference in the native chat template. We minimize token-average negative log-likelihood (NLL) on the answer and native assistant terminator, masking prompt tokens; no target teaches abstention. This objective rewards selected answers without constraining abstention on held-out questions. K/U each have 1,000 examples and seeds 17, 29, and 43; HardK has 600 examples and seeds 17 and 29. All runs restart from the same base within a model and train for three epochs.

The primary recipe uses LoRA (Hu et al., 2022) with rank $r = 16$, scaling parameter $\alpha = 16$, zero dropout, and the text blocks' linear modules, excluding embeddings and the output head. Base weights use bfloat16 (BF16) and adapters 32-bit floating point (FP32). AdamW uses learning rate $2 \times 10^{-4}$, zero weight decay, gradient clipping 1.0, effective batch 16, a 2,048-token limit, no packing, and 5% warmup followed by a constant rate. K/U have 189 optimizer steps and HardK 114, so HardK controls neither total training exposure nor cumulative loss exactly. Checkpoints at 25%, 50%, 75%, and 100% distinguish early changes from continued drift; final comparisons use the scheduled last checkpoint. The 24 primary training runs yield 96 adapted states plus three base evaluations. Ranks 8 and 32 add 48 runs with the same arm/seed design and $\alpha = r$, keeping adapter scaling $\alpha/r = 1$; checkpoints remain repeated measurements of a run (Appendix I).

### 3.4 Response Evaluation and Complementary Measurements

**Observed answers and abstentions.** Two natural prompts, `natural_a` and `natural_b`, request short answers and allow "I don't know." The first says "Answer the question briefly. If you do not know the answer, reply exactly: I don't know." The second asks for a brief answer and abstention when the model cannot answer from its knowledge. We use the native chat template, disable Qwen thinking, and decode greedily for at most 256 tokens. Both prompts permit paired transitions and test wording sensitivity.

The core-v2 pipeline combines normalized primary-answer and explicit-abstention rules with GPT-5.6 Luna semantic judgments at temperature 0. The judge sees the question, references, and response, without model, arm, seed, checkpoint, or historical label. Semantic matching accepts correct paraphrases; mixed responses endorsing an answer are judged on its core correctness. Of 237,600 rank-16 responses, 951 unresolved, truncated, or non-answer cases retain separate categories and remain in behavioral denominators. Independent human agreement is unmeasured, so the results depend on this operational judge (Appendices E and K).

**Transitions identify the source of a change.** For question $q$, $y_0(q)$ and $y_t(q)$ denote response categories at base and checkpoint $t$. Within historical panel $L \in \{K_0, U_0\}$, the rate from category $a$ to $b$ is

$$T^{L}_{a\to b}(t) = \frac{1}{|L|} \sum_{q\in L} \mathbf{1}\{y_0(q) = a,\ y_t(q) = b\}. \tag{1}$$

The full-panel denominator makes routes additive: the net error increase equals all inflows into incorrect responses minus all outflows. An abstention-to-error transition records a newly attempted incorrect answer; a correct-to-error transition records loss of a previously successful answer. A positive U–K difference means the route is more frequent under U-SFT, in percentage points (pp). The base outcome rates show how much room each model has for each transition (Table 2).

**Score ranking tests what frequency can hide.** For tokenized phrase $z =$ "I don't know", natural prompt $x(q)$, and checkpoint parameters $\theta_t$, we compute

$$s_{\mathrm{IDK}}(q,t) = \sum_{j=1}^{|z|} \log p_{\theta_t}(z_j \mid x(q), z_{<j}). \tag{2}$$

This teacher-forced sum excludes the end-of-sequence (EOS) token and measures the exact phrase prefix, not the total probability of all abstentions. Historical-label AUROC ranks $U_0$ above $K_0$ with $s_{\mathrm{IDK}}$, equivalently $K_0$ above $U_0$ with $-s_{\mathrm{IDK}}$. We also evaluate $-s_{\mathrm{IDK}}$ against current correct/incorrect judgments among resolved answered responses, excluding abstentions and review cases. AUROC is the probability that a random positive example outranks a random negative one, counting ties as one half; 0.5 denotes chance ordering and 1 perfect ordering. Neither AUROC target validates a threshold for deployment.

**Probes distinguish transfer from refitted separation.** Logistic probes with an L2 squared-weight penalty predict $K_0$ (positive) versus $U_0$ from decoder-block outputs at the last prompt token, before any answer tokens. A fixed probe retains its base scaler and classifier; a refit probe learns from the checkpoint's CAL-FIT states at the same layer. Binary ANSWER/ABSTAIN and natural prompts test format dependence; comparisons use 64-bit floating-point (FP64) fitting. These external classifiers measure label information, without establishing its causal use.

**Aggregation and uncertainty.** Behavioral summaries weight datasets and natural prompts equally, then average training seeds. Transition intervals use 10,000 paired bootstrap resamples of PopQA subjects and TriviaQA questions, keeping training seeds fixed. Each draw is shared across the compared arms, prompts, and seeds, preserving their paired questions. The intervals quantify item-sampling uncertainty conditional on the trained models; individual seeds expose training variability. The current-response AUROC analysis uses paired group draws and individual percentile intervals, with no simultaneous-comparison guarantee. Probe means lack completed nested fitting uncertainty. Transition decomposition and score AUROC were specified after the initial behavioral results; the current-response analysis is exploratory (Appendix J).

Table 2: Response outcomes at base and the final rank-16 checkpoint (% of each historical panel). Rates average datasets and natural prompts equally, then training seeds; unresolved outcomes are the remainder to 100%. K/U have three seeds and HardK two.

| Model | Training | $K_0$: Correct | Incorrect | Abstain | $U_0$: Correct | Incorrect | Abstain |
|---|---|---|---|---|---|---|---|
| Llama | Base | 83.00 | 0.83 | 15.75 | 4.75 | 19.92 | 73.17 |
| Llama | K | 89.33 | 10.19 | 0.08 | 4.83 | 81.53 | 13.56 |
| Llama | HardK | 84.04 | 15.25 | 0.12 | 4.21 | 86.17 | 8.79 |
| Llama | U | 70.75 | 28.75 | 0.00 | 4.06 | 95.75 | 0.00 |
| Qwen | Base | 91.58 | 2.25 | 4.67 | 5.00 | 62.75 | 28.67 |
| Qwen | K | 89.17 | 10.44 | 0.00 | 3.72 | 95.56 | 0.44 |
| Qwen | HardK | 81.42 | 18.42 | 0.00 | 2.79 | 97.00 | 0.17 |
| Qwen | U | 75.22 | 24.44 | 0.00 | 4.11 | 95.78 | 0.03 |
| OLMo | Base | 7.00 | 0.42 | 92.58 | 0.17 | 0.25 | 99.58 |
| OLMo | K | 45.67 | 8.97 | 44.58 | 1.17 | 30.25 | 68.42 |
| OLMo | HardK | 54.54 | 21.92 | 22.88 | 1.79 | 52.42 | 45.46 |
| OLMo | U | 35.06 | 23.28 | 41.33 | 1.78 | 44.69 | 53.17 |

Table 3: Extra transitions under U-SFT relative to K-SFT at rank 16 (percentage points). Denominators are full historical panels. The 95% intervals resample question groups with seeds fixed; seed values show training variability. Datasets and natural prompts are equally weighted.

| Model | Route | $U - K$ | 95% CI | Seeds 17, 29, 43 |
|---|---|---|---|---|
| Llama | $K_0$: correct → error | 14.53 | [11.96, 17.13] | 12.75, 14.25, 16.58 |
| Llama | $U_0$: abstain → error | 14.11 | [11.94, 16.35] | 12.50, 17.00, 12.83 |
| Qwen | $K_0$: correct → error | 12.17 | [9.61, 14.71] | 10.75, 13.42, 12.33 |
| Qwen | $U_0$: abstain → error | 0.44 | [-0.02, 0.97] | 0.58, 0.75, 0.00 |
| OLMo | $K_0$: correct → error | 0.58 | [0.27, 0.95] | 0.58, 0.33, 0.83 |
| OLMo | $U_0$: abstain → error | 14.47 | [11.92, 17.10] | 26.33, -5.67, 22.75 |

# 4 Different Failures behind Reduced Abstention

## 4.1 Known-Fact Training Already Reduces Abstention

Known-fact training leaves many initially abstained questions answered incorrectly (Figure 1a): 57.72% of $U_0$ in Llama, 27.69% in Qwen, and 30.00% in OLMo. The corresponding U-SFT rates are 71.83%, 28.14%, and 44.47%. These comparisons show that unknown examples are unnecessary for a large decline in abstention under this answer-only recipe. Qwen's final $U_0$ abstention rates under K/U are 0.44/0.03%, leaving little room for a frequency difference.

Reduced abstention yields little net accuracy gain on $U_0$ (Table 2). For Llama, K-SFT changes $U_0$ accuracy from 4.75% to 4.83%, while incorrect responses rise from 19.92% to 81.53%. Known-fact selection therefore does not preserve abstention on the held-out unknown panel simply by avoiding unknown training examples.

The extra transition rate under U-SFT is model-dependent (Table 3). Qwen's $U_0$ difference is 0.44 pp with a question-bootstrap interval crossing zero. OLMo's mean difference is 14.47 pp, but its seed values are 26.33, $-5.67$, and 22.75 pp. The positive interval conditional on the three seeds does not establish a uniformly positive effect across training seeds. Abstention loss alone therefore leaves the additional cost of U-SFT unresolved.

## 4.2 The Additional Errors Depend on the Base Response

The additional cost includes loss of successful answers: on $K_0$, U-SFT produces 14.53 pp more correct-to-error transitions than K-SFT in Llama and 12.17 pp more in Qwen, positive in all three seeds. For Qwen, this contrast is much larger than the $U_0$ abstention-to-error contrast. The trajectories also differ: Llama's U-SFT $U_0$ abstention-to-error rate is already 69.89% at the first checkpoint and reaches 71.83% at the last, whereas its $K_0$ correct-to-error rate increases from 14.67% to 21.92% over that interval. Timing differences alone do not establish separate mechanisms.

Table 4: AUROC of $-s_{\mathrm{IDK}}$ under natural_a for two targets. Historical: $K_0$ versus $U_0$. Answered: correct versus incorrect among each checkpoint's resolved answered responses. The latter excludes abstentions and unresolved outputs, so its population can differ across arms. Values are seed means; answered U–K intervals use 10,000 paired group draws with seeds fixed. Intervals are withheld if any draw is undefined.

| Model | Dataset | Historical K | Historical U | Answered K | Answered U | Answered $U - K$ [95% CI] |
|---|---|---|---|---|---|---|
| Llama | PopQA | 0.923 | 0.743 | 0.893 | 0.707 | -0.186 [-0.222, -0.151] |
| Llama | TriviaQA | 0.788 | 0.437 | 0.768 | 0.458 | -0.310 [-0.354, -0.266] |
| Qwen | PopQA | 0.894 | 0.437 | 0.873 | 0.426 | -0.447 [-0.487, -0.408] |
| Qwen | TriviaQA | 0.688 | 0.419 | 0.658 | 0.394 | -0.264 [-0.302, -0.226] |
| OLMo | PopQA | 0.700 | 0.540 | 0.560 | 0.460 | -0.101 [–] |
| OLMo | TriviaQA | 0.554 | 0.493 | 0.517 | 0.449 | -0.067 [-0.121, -0.013] |

OLMo changes the interpretation of the same comparison. Although its $K_0$ questions pass the forced-answer labeling criterion, the base model answers only 7.00% correctly under the natural prompts and abstains on 92.58%. After K-SFT its $K_0$ accuracy rises to 45.67%; after U-SFT it rises to 35.06%. The U–K gap thus largely reflects a smaller improvement from the base behavior, while the extra correct-to-error route is only 0.58 pp. Moreover, $K_0$ abstention-to-error transitions increase from 8.75% under K-SFT to 22.36% under U-SFT. This additional route explains why the two highlighted panels in Figure 1 cannot be interpreted as an exhaustive decomposition. The complete inflow/outflow identity holds in all 768 evaluated transition panels.

**Replication and partial controls.** At ranks 8 and 32, the extra $K_0$ correct-to-error rates are respectively 12.83 and 8.06 pp for Llama, and 15.17 and 8.83 pp for Qwen; all nine rank–seed instances per model have a positive contrast. They share the evaluation questions, so they are nine trained instances rather than nine independent dataset replications. Qwen's $U_0$ contrasts remain small (0.61 and $-0.14$ pp), whereas OLMo's $K_0$ correct-to-error contrasts remain below 1 pp (Appendix I). HardK produces final $K_0$ correct-to-error rates of 12.08% in Llama and 15.83% in Qwen, between K-SFT's 7.39/9.11% and U-SFT's 21.92/21.28%. These intermediate values persist when HardK's 114-step endpoint is compared with the nearby K/U 95- and 142-step checkpoints (Appendix F). Thus difficult known examples reproduce part of the pattern, while HardK's smaller training set prevents attributing the remaining gap solely to unknownness. Linear adjustment for update norm retains positive Llama/Qwen coefficients; adding cumulative loss changes some conclusions, and the K/U loss ranges do not overlap at any rank-16 checkpoint. These post-treatment adjustments cannot establish a loss-independent causal effect (Appendix G). Even these error routes leave open whether the model's scores still distinguish reliable answers.

### 4.3 A Behavioral Floor Hides Different Score Rankings

Qwen's near-zero abstention under both arms does not imply similar ordering of questions by $s_{\mathrm{IDK}}$. On PopQA, its historical-label AUROC is 0.835 at base, 0.894 after K-SFT, and 0.437 after U-SFT (Table 4). On TriviaQA the corresponding values are 0.757, 0.688, and 0.419. These are differences in a continuous score even when greedy abstention has almost disappeared. The abstention rate alone misses this K/U difference in historical-label ranking.

The historical-label pattern varies across models: Llama on PopQA retains a U-SFT AUROC of 0.743, while OLMo already loses substantial selectivity under K-SFT. The U-SFT means range from 0.419 to 0.540 in the other five model–dataset cells; this range is descriptive and does not establish statistical equivalence to chance.

Historical labels do not establish current reliability, so we also rank resolved answers by correctness. Qwen PopQA's AUROC is 0.873 after K-SFT and 0.426 after U-SFT, a difference of $-0.447$ (95% paired question-group interval $[-0.487, -0.408]$, seeds fixed). Under natural_b, Qwen PopQA remains separated at 0.880 versus 0.596 (U–K interval $[-0.322, -0.245]$), so U-SFT is not uniformly at or below chance across prompts. Current-response AUROC is lower after U-SFT in all twelve rank-16 model–dataset–prompt comparisons, with the same direction in each seed. Ranking quality decreases under U-SFT even where its AUROC remains above 0.5. These comparisons condition

Table 5: Historical-label probe AUROC after U-SFT: seed means at the fixed CAL-selected layer. Within each prompt, Base, Fixed, and Refit use the same FP64 fitting pipeline. Prompt dependence is visible without selecting layers on EVAL. These are descriptive means, not equivalence tests.

| Model | Dataset | Binary action: Base / Fixed / Refit | | | Natural: Base / Fixed / Refit | | |
|---|---|---|---|---|---|---|---|
| Llama | PopQA | 0.827 | 0.881 | 0.876 | 0.887 | 0.825 | 0.869 |
| Llama | TriviaQA | 0.849 | 0.837 | 0.844 | 0.875 | 0.846 | 0.874 |
| Qwen | PopQA | 0.936 | 0.652 | 0.911 | 0.931 | 0.896 | 0.918 |
| Qwen | TriviaQA | 0.884 | 0.695 | 0.843 | 0.891 | 0.859 | 0.852 |
| OLMo | PopQA | 0.910 | 0.795 | 0.937 | 0.924 | 0.924 | 0.948 |
| OLMo | TriviaQA | 0.799 | 0.693 | 0.802 | 0.816 | 0.754 | 0.848 |

on each arm's answered population, which differs especially for OLMo; threshold utility remains untested. The remaining question is whether weaker output ranking also means weaker historical-label separation inside the model.

## 5 Probe Transfer Depends on the Measurement Prompt

Probes test label separation inside the model, extending the output-score analysis. We fit L2-regularized probes on CAL-FIT, selecting inverse regularization strength $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ by five-fold group-wise cross-validation; ties favor smaller $C$. Each fold standardizes states using only its training portion; the final fit uses all CAL-FIT states. Base action-prompt CAL-SELECT AUROC, averaged over datasets among layers 0–30, selects layers 12 (Llama), 26 (Qwen), and 16 (OLMo), with shallower layers preferred on ties; the natural assay inherits this selection. We compare the binary action prompt, which requests 1 for ANSWER or 2 for ABSTAIN, with `natural_a`. Base, fixed, and refit comparisons share the FP64 quasi-Newton solver (L-BFGS) and labels; EVAL chooses neither probes nor layers.

In Qwen on PopQA, the fixed action probe falls from 0.936 at base to 0.652 after U-SFT; refitting reaches 0.911 (Table 5). Under the natural prompt, the corresponding values are 0.931, 0.896, and 0.918. The much larger transfer loss under the action prompt shows that a conclusion about representation change depends on the elicitation format. This is also not specific to unknown-fact training: K-SFT's action fixed/refit values are 0.708/0.917, while its natural values are 0.905/0.915. OLMo shows a similar prompt contrast on PopQA. The binary action format is also farther from the short-answer training task, so format transfer is an alternative explanation for the prompt-dependent probe drop. Llama's action fixed probe does not show the same decrease, consistent with the possibility that an unchanged probe transfers in some settings (Shi et al., 2026).

Refitting recovers historical-label separation without establishing preservation of the facts themselves. The recovery is incomplete in Qwen on TriviaQA: refit AUROC is below base by 0.041 under the action prompt and 0.039 under the natural prompt. Thus refitting improves separation without restoring it uniformly.

Surface-cue controls further restrict the interpretation. They replace hidden states with popularity/relation metadata, term-frequency–inverse-document-frequency (TF-IDF) question features, question length, or mean input embeddings of all prompt tokens, including template tokens, using the same calibration splits. On PopQA, popularity and relation alone yield AUROCs of 0.763, 0.868, and 0.925 for Llama, Qwen, and OLMo. The OLMo U-SFT refit values of 0.937/0.948 are only modestly higher in point estimates, so we cannot attribute its strong PopQA separation entirely to a knowledge-specific internal signal. On TriviaQA, the best tested surface or input-embedding baseline reaches 0.664, 0.625, and 0.638 respectively, below the U-SFT refit means by at least 0.16 AUROC. Those differences are descriptive; paired uncertainty over probe fitting remains necessary for an inferential comparison. The completed surface-feature random-label controls average 0.49–0.52; they do not substitute for hidden-state controls. Probe transfer and refitted separation describe different properties, while causal use remains untested.

## 6 DISCUSSION

**Known-fact selection needs a separate check of abstention behavior.** Knowledge-aligned supervision can improve factuality and refusal in other settings (Becker et al., 2026). Our answer-only recipe exposes a limitation: K-SFT converts many abstentions into errors, while U-SFT further degrades score ranking and, in Llama/Qwen, previously correct responses. For data curators, the implication is to evaluate both what the examples teach and whether the resulting model still abstains appropriately. Filtering known facts and explicitly teaching abstention therefore warrant separate evaluation under equal budgets; no mitigation is established here.

**An accuracy gap has different costs at different starting points.** Llama and Qwen lose previously correct answers under U-SFT relative to K-SFT. OLMo instead begins by withholding most $K_0$ answers, improves under both arms, and improves less under U-SFT. Calling both changes knowledge destruction would conflate lost successes with unrealized gains. For model evaluators, base-to-checkpoint transitions distinguish these cases and retain the cost of excessive abstention alongside the cost of new errors. Seed and rank replication strengthen the Llama/Qwen comparison, while OLMo's seed reversal limits a universal claim about the extra abstention-to-error route.

**A score is useful only for the target it was validated against.** Qwen illustrates why similar abstention rates can conceal different rankings of both historical labels and current answered-response correctness. The probe comparison adds a second caution: the diagnosis of transfer loss changes with the prompt, even at a fixed layer and with a common solver. For researchers auditing epistemic behavior, the label, prompt, and allowed refitting are part of the measurement definition. High historical-label AUROC cannot certify current knowledge, and an external probe's success cannot establish causal use by the model. A practical next test would choose a threshold on calibration data and evaluate held-out selective risk and coverage (Geifman & El-Yaniv, 2017). Optimizing the exact-phrase score could instead change refusal wording without improving correctness; its use as a training signal is untested.

**Limitations.** The study uses three open model families, two English factual QA datasets, one learning rate, and LoRA adaptation. Historical labels deliberately exclude intermediate answerability and change neither with training nor with the evaluation prompt. K/U training sets can differ in popularity, difficulty, and semantics; HardK and post-treatment regression only partially address these alternatives. Initial-label review and response judging are AI-assisted, and an independent human audit is not available in this draft. Question-bootstrap intervals hold training seeds fixed; the probe means do not yet include nested fitting uncertainty, and answered-response AUROCs compare populations that can change across arms. We do not test mitigation efficacy, general capabilities, deception, or deployment safety, and do not identify a causal mechanism.

## 7 CONCLUSION

Answer-only factual SFT changes abstention frequency, the origins of wrong answers, and score ranking in different ways. Known-fact training already reduces abstention substantially, while unknown-fact training adds model-dependent costs that a single abstention rate can conceal. Fixed and refit probes provide additional diagnostic information whose interpretation depends on the prompt and the label being decoded. Evaluating these quantities together gives data curators and model evaluators a more precise account of what factual fine-tuning changed. Known-fact selection can outperform unknown-fact training while still leaving substantial errors among newly attempted answers. Any score-based abstention policy therefore requires separate validation on current outcomes; historical-label separation alone does not establish its utility.

## Reproducibility Statement

Section 3 defines the main measurements and independent units. Appendices B–E give data labeling, training settings, prompts, and the response-judging description; Appendices F–I provide extended results and analysis procedures. The accompanying anonymous supplementary archive, `submission_code.zip`, contains the final training, evaluation, judging, and analysis code, dependency specifications, and an offline self-check. Its README maps paper components to entrypoints and specifies relative-path contracts for separately provisioned cohorts, model weights, responses, and activation caches, which are excluded from the archive. The display scripts rebuild the generated result tables and overview figure from frozen inputs and check all 768 transition accounting identities; the README distinguishes validated CPU paths from GPU and API stages not rerun for this release. Appendix J records analysis timing and remaining protocol documentation.

## Ethics Statement

This study analyzes factual answering by publicly released models on PopQA and TriviaQA. It does not establish that the resulting checkpoints are suitable for deployment. The present analysis includes automated answer judging and no reported human evaluation; independent human validation remains a limitation. The release should retain error rates and unresolved judgments so that users can assess the scope of the measurements.

## AI Use Statement

Generative AI tools assisted literature search, methodological feedback, interpretation of saved results, analysis/display code, manuscript planning, drafting, editing, and critique of this paper. An LLM-based judge is part of the response measurement (Appendix E). The authors are responsible for verifying all AI-assisted content, claims, and artifacts.

## References

Hao An, Yibin Lou, Jiayi Guo, and Yang Xu. Purging the gray zone: Latent-geometric denoising for precise knowledge boundary awareness. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 2562–2576, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.122. URL https://aclanthology.org/2026.findings-acl.122/.

Arthur Becker, Jakob Kemmler, David Thulke, Christine Schäfer, Christian Dugast, and Hermann Ney. Stick to what you know: A study of knowledge-aligned supervised fine-tuning. *arXiv preprint arXiv:2608.30987*, 2026. URL https://arxiv.org/abs/2608.30987. Accepted to EMNLP 2026, as reported on arXiv.

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

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-Level Uncertainty in Large Language Models. In *International Conference on Learning Representations*, volume 2026, pp. 34353–34372, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3a07c3a67cfe50d3236b71fb674c7f30-Abstract-Conference.html.

Renfei Dang, Peng Hu, Zhejian Lai, Changjiang Gao, Min Zhang, and Shujian Huang. Understanding new-knowledge-induced factual hallucinations in LLMs: Analysis and interpretation. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 7219–7246, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.358. URL https://aclanthology.org/2026.findings-acl.358/.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I Know This Entity? Knowledge Awareness and Hallucinations in Language Models. In *International Conference on Learning Representations*, volume 2025, pp. 77965–78000, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c1c44e46358e0fb94dc94ec495a7fb1a-Abstract-Conference.html.

Yonatan Geifman and Ran El-Yaniv. Selective Classification for Deep Neural Networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/4a8423d5e91fda00bb7e46540e2b0cf1-Abstract.html.

Zorik Gekhman, Gal Yona, Roee Aharoni, Matan Eyal, Amir Feder, Roi Reichart, and Jonathan Herzig. Does fine-tuning LLMs on new knowledge encourage hallucinations? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 7765–7784, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.444. URL https://aclanthology.org/2024.emnlp-main.444/.

Gaurav Rohit Ghosal, Tatsunori Hashimoto, and Aditi Raghunathan. Understanding finetuning for factual knowledge extraction. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 15540–15558. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/ghosal24a.html.

Satchel Grant, Simon Jerome Han, Alexa R. Tartaglini, and Christopher Potts. Addressing divergent representations from causal interventions on neural networks. In *International Conference on Learning Representations*, volume 2026, pp. 11330–11356, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/133e588e1429f9f1e25b215da145580e-Abstract-Conference.html.

John Hewitt and Percy Liang. Designing and Interpreting Probes with Control Tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275. URL https://aclanthology.org/D19-1275/.

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

Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 24564–24579, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1199. URL https://aclanthology.org/2025.acl-long.1199/.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna

Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*, 2022. doi: 10.48550/arXiv.2207.05221. URL https://arxiv.org/abs/2207.05221.

Katie Kang, Eric Wallace, Claire Tomlin, Aviral Kumar, and Sergey Levine. Unfamiliar finetuning examples control how language models hallucinate. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3600–3612, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.183. URL https://aclanthology.org/2025.naacl-long.183/.

Guy Kaplan, Zorik Gekhman, Zhen Zhu, Lotem Rozner, Yuval Reif, Swabha Swayamdipta, Derek Hoiem, and Roy Schwartz. Why fine-tuning encourages hallucinations and how to fix it. In *Third Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2604.15574.

Dharshan Kumaran, Nathaniel Daw, Simon Osindero, Petar Veličković, and Viorica Patraucean. Causal evidence that language models use confidence to drive behaviour. *Nature Machine Intelligence*, 8(9):1411–1425, 2026. doi: 10.1038/s42256-026-01293-x. URL https://www.nature.com/articles/s42256-026-01293-x.

Jiaqi Li, Yixuan Tang, and Yi Yang. Know the unknown: An uncertainty-sensitive method for LLM instruction tuning. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 2972–2989, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.153. URL https://aclanthology.org/2025.findings-acl.153/.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9802–9822, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.546. URL https://aclanthology.org/2023.acl-long.546/.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs Know More Than They Show: On the Intrinsic Representation of LLM Hallucinations. In *International Conference on Learning Representations*, volume 2025, pp. 66880–66913, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a712d461e57201efe35d429a6f1731c1-Abstract-Conference.html.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-Tuning Enhances Existing Mechanisms: A Case Study on Entity Tracking. In *International Conference on Learning Representations*, volume 2024, pp. 8057–8082, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/2082273791021571c410f41d565d0b45-Abstract-Conference.html.

William F. Shen, Xinchi Qiu, Nicola Cancedda, and Nicholas D. Lane. Seat: Sparse entity-aware tuning for knowledge adaptation while preserving epistemic abstention. *arXiv preprint arXiv:2506.14387*, 2025. URL https://arxiv.org/abs/2506.14387.

Zeyu Shi, Ziming Wang, Tianyu Chen, Shiqi Gao, Haoyi Zhou, Qingyun Sun, and Jianxin Li. Fine-Tuned LLMs Know They Don't Know: A Parameter-Efficient Approach to Recovering Honesty. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 32983–32991, 2026. doi: 10.1609/aaai.v40i39.40580. URL https://ojs.aaai.org/index.php/AAAI/article/view/40580.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2025. URL https://arxiv.org/abs/2512.13961.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi R. Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7113–7139, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.394. URL https://aclanthology.org/2024.naacl-long.394/.

## A Evidence Map

The main text contains the label definitions, experimental contrasts, confidence-interval interpretation, key behavioral results, all six model–dataset score comparisons, and both prompts of the U-SFT probe comparison. The appendix supplies exact prompts, training settings, complete transition accounting, additional prompt and rank results, and diagnostic controls. All means and displayed transition intervals refer to completed analyses; pending causal interventions are not reported as findings. The supplementary code README maps these analyses to executable entrypoints and specifies the required external inputs.

## B Question Labels

**Why label questions by each model's own answers?** A question that one model answers may be unknown to another, and a dataset-level label (answerable versus false-premise, supported versus fictitious) mixes what the model knows with how the question is built. Each model therefore labels every candidate question with its own answers before fine-tuning. The model answers under two short-answer prompts without an abstention option (Appendix D), once greedily and four times at temperature 0.5 per prompt, giving ten answers. A question is *known* ($K_0$) if both greedy answers and at least seven of the eight sampled answers are correct, and *unknown* ($U_0$) if none of the ten is correct; all other questions are excluded. The initial labels use externally judged semantic core correctness, followed by recorded AI-assisted review of selected cases; provisional alias-parser labels are not the final labels. The final cohort record binds each selected question's ten draws to actual judgments or identical-input cached judgments and excludes unresolved or risky cases from selection. These checks are AI-assisted adjudication, not an independent human gold audit. The source cohort is frozen before the reported training runs.

**Splits.** For every model and dataset, Table 6 lists the item-disjoint training and evaluation sets. The combined K/U/HardK training pool, CAL-FIT, CAL-SELECT, and EVAL are disjoint under normal-

ized primary answers, PopQA subjects, recorded duplicate-fact groups, and normalized questions, including cross-dataset links. K and HardK are item-disjoint subsets of the training partition; group separation is not asserted between these two arms. Only calibration questions determine probe fits and layer selection; EVAL chooses neither checkpoints nor regularization.

Table 6: Questions per model and dataset (known / unknown).

| Part | Questions | Use |
|---|---|---|
| TRAIN | 500 / 500 | K-SFT and U-SFT training sets |
| HardK | 300 / – | HardK-SFT training set (known only) |
| CAL-FIT | 75 / 75 | Fitting historical-label probes |
| CAL-SELECT | 75 / 75 | Selecting the probe layer |
| EVAL | 300 / 300 | All reported results |

## C FINE-TUNING

**Why these three arms?** K-SFT and U-SFT answer the question of the paper: they have the same size (1,000 examples: 500 per dataset), prompt, supervision, optimizer, and schedule, and differ in the selected fact sets. Knownness is not randomly assigned to an otherwise identical fact. U-SFT examples start with a higher loss, so a difference between the arms could reflect the size of the training signal rather than unknownness. HardK-SFT selects 300 known questions per dataset from the higher-loss half of 800 eligible known candidates, matching them to a fixed subset of 300 U examples using question/answer length and available PopQA metadata. Matching uses a global-distance fallback when tighter strata are insufficient. The remaining 500 known questions form ordinary K-SFT. Pairing those 500 with all 500 U examples determines order but cannot alter set membership. Thus matching selects HardK, not ordinary K; selecting HardK can also change the composition of its K complement. HardK-SFT is a partial control, since high-loss known facts do not reach the loss of unknown facts, and Appendix G adds a regression on update size.

The residual imbalance is substantial on PopQA: answer-length SMDs are 0.628–0.877 and popularity SMDs are $-0.774$ to $-0.237$. Training-set balance should not be inferred from the shared optimizer and example count.

**Configuration.** Table 8 lists the primary rank-16 configuration shared across arms, models, and seeds. Each training example uses `forced_a` (Appendix D) in the native chat template followed by the gold answer; the loss covers the answer and assistant terminator, excluding prompt tokens. K-SFT and U-SFT run for 189 optimizer steps and HardK-SFT for 114 steps; checkpoints are saved after 25%, 50%, 75%, and 100% of the steps (K-SFT and U-SFT: steps 48, 95, 142, 189). K-SFT and U-SFT use seeds 17, 29, and 43, and HardK-SFT uses seeds 17 and 29.

## D PROMPTS

Every prompt is inserted as the user turn of the model's native chat template; Qwen3.5 runs with thinking disabled. Evaluation uses batched BF16 greedy decoding, whose numerical output can depend on batch width and co-batched prompts. Reproducing the recorded outputs requires the original generation and scoring batch settings. In Table 9, `\n` marks newlines.

## E ANSWER JUDGING

**Why a semantic judge?** Exact alias matching marks correct answers wrong when they are phrased differently from the reference, and such errors would inflate correct-to-error transitions. Responses are therefore scored by the core-v2 pipeline, which combines exact-answer and abstention rules with semantic judgments by GPT-5.6 Luna (reasoning disabled, temperature 0) against the dataset's reference answers, and assigns each response to correct, incorrect, abstention, or an unresolved category. Valid, EOS-terminated responses exactly matching the primary answer after NFC/whitespace normalization take a deterministic correct route. Whole-response "I don't know"/"I do not know"

Table 7: Training-set imbalance: standardized mean differences (SMDs) are U minus the indicated known-fact set, divided by pooled standard deviation. HardK is matched to 300 U examples; ordinary K is its 500-item complement compared with all 500 U examples. Length uses tokens; popularity uses log(1+popularity).

| Model | Dataset | Arm | Question SMD | Answer SMD | Popularity SMD |
|---|---|---|---|---|---|
| Llama | PopQA | K | -0.441 | 0.628 | -0.774 |
| Llama | PopQA | HardK | -0.219 | 0.374 | -0.476 |
| Llama | TriviaQA | K | 0.256 | 0.221 | – |
| Llama | TriviaQA | HardK | 0.221 | 0.279 | – |
| Qwen | PopQA | K | -0.443 | 0.877 | -0.333 |
| Qwen | PopQA | HardK | -0.022 | 0.347 | -0.162 |
| Qwen | TriviaQA | K | 0.369 | 0.230 | – |
| Qwen | TriviaQA | HardK | 0.183 | 0.350 | – |
| OLMo | PopQA | K | -0.448 | 0.864 | -0.237 |
| OLMo | PopQA | HardK | -0.042 | 0.238 | -0.351 |
| OLMo | TriviaQA | K | 0.358 | 0.346 | – |
| OLMo | TriviaQA | HardK | 0.209 | 0.186 | – |

Table 8: Fine-tuning configuration.

| Setting | Value |
|---|---|
| Method | LoRA, rank 16, $\alpha = 16$, dropout 0 |
| Target modules | All linear modules of the transformer blocks (attention, mixer, MLP) |
| Excluded | Embeddings, LM head, vision encoder, multimodal projector |
| Precision | Base weights BF16, LoRA parameters FP32 |
| Optimizer | AdamW, learning rate $2 \times 10^{-4}$, weight decay 0, gradient clipping 1.0 |
| Schedule | 5% warmup, then constant |
| Batch and length | Effective batch 16, 3 epochs, maximum 2,048 tokens, no packing |

variants take the abstention route, and completed empty outputs take the non-answer route. Invalid or truncated generations remain generation-review cases. All other valid responses enter the semantic route; membership in a normalized alias list alone does not determine correctness. The semantic request contains only the question, primary reference, aliases, and response, one response per request. Its strict JSON fields are answer behavior (answered, abstained, mixed, nonanswer), answer correctness (correct, incorrect, uncertain, or null), reference status, and a reason. Unusable references and uncertain answers retain separate review categories; a mixed response that endorses a candidate is judged for that candidate's core correctness. The judge receives the question, the reference answers, and the response, and never the model, arm, seed, checkpoint, or label.

**Unresolved responses are kept, not recoded.** Of the 237,600 natural-prompt responses in the rank-16 analyses, 85,865 are correct, 101,800 incorrect, and 48,984 abstentions; 543 are uncertain, 233 were truncated at the 256-token limit, and 175 are non-answers. These 951 responses keep their own categories and count toward the denominators, so no response is turned into an error by construction. Independent human agreement has not been measured; the automated judgments should not be treated as an audited human gold standard.

## F Complete Transition Accounting

Rates use the entire $K_0$ or $U_0$ panel as denominator, separately within each model, dataset, prompt, seed, and checkpoint. Every one of the 768 rank-16 trajectory cells satisfies net incorrect change = inflow into incorrect responses $-$ outflow from incorrect responses. The two highlighted routes in the main text are not an exhaustive partition. Transition and regression intervals resample PopQA subjects or TriviaQA questions separately within each dataset. This does not jointly resample every answer-linked component used for split isolation. Draws that lose a historical-label denominator are replaced, with rejection capped at 1%; the current-response AUROC analysis below instead withholds intervals when any draw is undefined. For example, OLMo U-SFT increases wrong answers

Table 9: Prompts. {question} is replaced by the question; the binary prompt uses the tokens 1 (ANSWER) and 2 (ABSTAIN).

| Name | Text | Use |
|---|---|---|
| `forced_a` | Answer the following factual question with only the shortest answer. Do not explain.\nQuestion: {question}\nAnswer: | Labels, SFT examples |
| `forced_b` | Give only the concise factual answer to this question. Do not add explanation.\nQuestion: {question}\nAnswer: | Labels |
| `natural_a` | Answer the question briefly. If you do not know the answer, reply exactly: I don't know.\nQuestion: {question} | Behavior, probes |
| `natural_b` | Give a brief answer to the question. If you cannot answer from your knowledge, write exactly: I don't know.\nQuestion: {question} | Behavior |
| `action` | Decide whether you should answer the factual question from your own knowledge or abstain because you do not know.\nOutput only 1 for ANSWER or 2 for ABSTAIN.\nQuestion: {question}\nDecision: | Historical-label probes |

Table 10: Base-model natural responses (% of each historical panel), averaged over both datasets and prompts. Other includes unresolved, truncated, and non-answer responses.

| Model | Panel | Correct | Incorrect | Abstain | Other |
|---|---|---|---|---|---|
| Llama | $K_0$ | 83.00 | 0.83 | 15.75 | 0.42 |
| Llama | $U_0$ | 4.75 | 19.92 | 73.17 | 2.17 |
| Qwen | $K_0$ | 91.58 | 2.25 | 4.67 | 1.50 |
| Qwen | $U_0$ | 5.00 | 62.75 | 28.67 | 3.58 |
| OLMo | $K_0$ | 7.00 | 0.42 | 92.58 | 0.00 |
| OLMo | $U_0$ | 0.17 | 0.25 | 99.58 | 0.00 |

by 33.65 percentage points averaged over both historical panels; $K_0$ abstention-to-error transitions make an important additional contribution. Table 2 in the main text gives both base and final outcomes; Table 10 additionally lists the base unresolved category.

At its final checkpoint (114 steps), HardK's $K_0$ correct-to-error rate is 12.08% for Llama and 15.83% for Qwen. At the K/U 50% checkpoint (95 steps), the corresponding K/U rates are 6.75/16.94% and 8.67/22.00%; at 75% (142 steps), they are 6.81/15.97% and 8.67/20.42%. HardK lies between these arms at both nearby checkpoints, but neither comparison matches cumulative loss, step count, or number of examples exactly.

**Score ranking under a second natural prompt.** The following tables retain base values and seed ranges for both natural prompts. The AUROC positive class remains the base model's historical unknown panel. Means and seed ranges are descriptive; no significance or equivalence-to-chance verdict is inferred from them.

**Exploratory current-response discrimination.** We link 57,600 saved final-checkpoint item rows, deduplicate the base responses, and independently reproduce all 108 historical AUROCs (including base, both prompts, and all arms) to within $10^{-12}$. Current-response AUROC uses correct as positive and incorrect as negative only among resolved answered responses; it leaves abstentions and all review categories outside that binary target. A single set of 10,000 question-group bootstrap draws per model and dataset is shared across arms, seeds, prompts, and label targets (random seed 20260926). The contrast averages the three fixed training seeds before taking U minus K. Intervals are individual percentile intervals, not simultaneous or seed-population intervals. An undefined draw is neither replaced nor silently dropped: OLMo PopQA under natural_a has 9,999 valid draws, and its interval is withheld. The base OLMo PopQA answered subset has 13 correct and no incorrect responses, so its current-response AUROC is undefined. The final score extraction, per-panel counts, all 24 paired contrasts, and deterministic reconstruction checks are retained with the analysis artifact. This analysis was added during manuscript preparation after the historical-label results; it is exploratory and does not validate a threshold chosen on calibration data. Table 16 retains the second

Table 11: Final-checkpoint accounting on the $K_0$ panel (%). Columns A, C, O are inflows into incorrect responses from abstention, correct, and other; Exit is the outflow from incorrect responses. Net error change is A+C+O–Exit. Post-response other is the remainder to 100%.

| Model | Arm | Correct | Incorrect | Abstain | A→I | C→I | O→I | Exit |
|---|---|---|---|---|---|---|---|---|
| Llama | K | 89.33 | 10.19 | 0.08 | 2.14 | 7.39 | 0.03 | 0.19 |
| Llama | HardK | 84.04 | 15.25 | 0.12 | 2.50 | 12.08 | 0.08 | 0.25 |
| Llama | U | 70.75 | 28.75 | 0.00 | 6.06 | 21.92 | 0.14 | 0.19 |
| Qwen | K | 89.17 | 10.44 | 0.00 | 0.44 | 9.11 | 0.17 | 1.53 |
| Qwen | HardK | 81.42 | 18.42 | 0.00 | 0.88 | 15.83 | 0.67 | 1.21 |
| Qwen | U | 75.22 | 24.44 | 0.00 | 1.50 | 21.28 | 0.42 | 1.00 |
| OLMo | K | 45.67 | 8.97 | 44.58 | 8.75 | 0.08 | 0.00 | 0.28 |
| OLMo | HardK | 54.54 | 21.92 | 22.88 | 21.04 | 0.71 | 0.00 | 0.25 |
| OLMo | U | 35.06 | 23.28 | 41.33 | 22.36 | 0.67 | 0.00 | 0.17 |

Table 12: Final-checkpoint accounting on the $U_0$ panel (%). Columns A, C, O are inflows into incorrect responses from abstention, correct, and other; Exit is the outflow from incorrect responses. Net error change is A+C+O–Exit. Post-response other is the remainder to 100%.

| Model | Arm | Correct | Incorrect | Abstain | A→I | C→I | O→I | Exit |
|---|---|---|---|---|---|---|---|---|
| Llama | K | 4.83 | 81.53 | 13.56 | 57.72 | 3.39 | 1.92 | 1.42 |
| Llama | HardK | 4.21 | 86.17 | 8.79 | 61.92 | 3.42 | 2.17 | 1.25 |
| Llama | U | 4.06 | 95.75 | 0.00 | 71.83 | 3.81 | 1.86 | 1.67 |
| Qwen | K | 3.72 | 95.56 | 0.44 | 27.69 | 4.00 | 3.25 | 2.14 |
| Qwen | HardK | 2.79 | 97.00 | 0.17 | 28.33 | 3.79 | 3.46 | 1.33 |
| Qwen | U | 4.11 | 95.78 | 0.03 | 28.14 | 3.56 | 3.42 | 2.08 |
| OLMo | K | 1.17 | 30.25 | 68.42 | 30.00 | 0.00 | 0.00 | 0.00 |
| OLMo | HardK | 1.79 | 52.42 | 45.46 | 52.29 | 0.00 | 0.00 | 0.12 |
| OLMo | U | 1.78 | 44.69 | 53.17 | 44.47 | 0.00 | 0.00 | 0.03 |

prompt, including Qwen PopQA's above-chance U-SFT point estimate. Values below 0.5 indicate reversed ordering for the specified score orientation on the evaluated subset, not an absence of all discriminative information. Flipping a score or selecting a new threshold after inspecting EVAL would require separate validation.

**HardK score ranking.** Under natural_a, HardK-SFT AUROCs on PopQA/TriviaQA are 0.939/0.785 for Llama, 0.884/0.724 for Qwen, and 0.754/0.594 for OLMo. Its smaller example and step counts leave training exposure and difficulty only partially controlled.

## G Update-Size Regression

**Why a regression on update size?** U-SFT examples carry more loss, so U-SFT could change the model more simply because its update is larger. This regression describes residual arm differences under a linear model; it does not identify a direct effect of unknownness. For each model, dataset, prompt, and transition, three linear models are fit to the 32 run–checkpoint pairs of rank 16 (eight runs, four checkpoints). Model A contains arm (U-SFT and HardK-SFT against K-SFT), checkpoint, and seed indicators. Model B adds the standardized $\log(1 + \|\Delta W\|_F)$ of the global adapter update. Model C adds the standardized cumulative training loss (summed token negative log-likelihood). Intervals use the paired bootstrap of Appendix F.

**Results.** In Llama and Qwen, the U-SFT coefficient on the correct-to-error route stays positive under model B on both datasets and both prompts, with the lower end of every interval above zero (Table 17). Both adjustment variables are measured after training, so the coefficients describe linear adjustment rather than a direct effect; the K-SFT and U-SFT cumulative-loss ranges do not overlap at any checkpoint of any model, so model C extrapolates the contrast between the arms. Under model C the coefficient stays positive on PopQA in both models and on TriviaQA in Llama with natural_b, and its interval includes zero on TriviaQA in Qwen and in Llama with natural_a.

Table 13: Transition rates over training (% of the full historical panel), averaged over datasets, natural prompts, and seeds. HardK has a different number of examples and total steps.

| Model | Arm | Route | 25% | 50% | 75% | 100% |
|---|---|---|---|---|---|---|
| Llama | K | $U_0$: A→I | 45.53 | 50.08 | 50.72 | 57.72 |
| Llama | K | $K_0$: C→I | 5.81 | 6.75 | 6.81 | 7.39 |
| Llama | HardK | $U_0$: A→I | 68.29 | 66.17 | 66.17 | 61.92 |
| Llama | HardK | $K_0$: C→I | 10.08 | 12.46 | 12.67 | 12.08 |
| Llama | U | $U_0$: A→I | 69.89 | 71.31 | 71.17 | 71.83 |
| Llama | U | $K_0$: C→I | 14.67 | 16.94 | 15.97 | 21.92 |
| Qwen | K | $U_0$: A→I | 27.39 | 26.53 | 27.89 | 27.69 |
| Qwen | K | $K_0$: C→I | 6.72 | 8.67 | 8.67 | 9.11 |
| Qwen | HardK | $U_0$: A→I | 27.29 | 28.42 | 28.29 | 28.33 |
| Qwen | HardK | $K_0$: C→I | 13.75 | 18.00 | 19.17 | 15.83 |
| Qwen | U | $U_0$: A→I | 27.67 | 28.08 | 28.19 | 28.14 |
| Qwen | U | $K_0$: C→I | 12.50 | 22.00 | 20.42 | 21.28 |
| OLMo | K | $U_0$: A→I | 18.86 | 22.03 | 33.00 | 30.00 |
| OLMo | K | $K_0$: C→I | 0.08 | 0.03 | 0.11 | 0.08 |
| OLMo | HardK | $U_0$: A→I | 16.75 | 53.17 | 52.00 | 52.29 |
| OLMo | HardK | $K_0$: C→I | 0.12 | 0.38 | 0.71 | 0.71 |
| OLMo | U | $U_0$: A→I | 24.14 | 28.67 | 34.92 | 44.47 |
| OLMo | U | $K_0$: C→I | 0.31 | 0.42 | 0.58 | 0.67 |

Table 14: AUROC of the fixed phrase "I don't know" ranking historical $U_0$ above $K_0$ under `natural_a`. Values are seed means [minimum, maximum], not confidence intervals. The positive label is fixed by base-model labeling, rather than current correctness.

| Model | Dataset | Base | K-SFT | U-SFT |
|---|---|---|---|---|
| Llama | PopQA | 0.891 | 0.923 [0.901, 0.939] | 0.743 [0.653, 0.841] |
| Llama | TriviaQA | 0.878 | 0.788 [0.760, 0.810] | 0.437 [0.411, 0.461] |
| Qwen | PopQA | 0.835 | 0.894 [0.879, 0.905] | 0.437 [0.355, 0.515] |
| Qwen | TriviaQA | 0.757 | 0.688 [0.605, 0.756] | 0.419 [0.407, 0.431] |
| OLMo | PopQA | 0.790 | 0.700 [0.674, 0.735] | 0.540 [0.463, 0.597] |
| OLMo | TriviaQA | 0.726 | 0.554 [0.545, 0.562] | 0.493 [0.454, 0.547] |

Model C also changes signs outside the displayed correct-to-error rows: for Llama PopQA under natural_b, the $U_0$ abstention-to-error coefficient changes from 0.532 in B to $-0.178$ in C (95% interval $[-0.258, -0.097]$). For OLMo TriviaQA under natural_a, the C coefficient is 1.032, which is a linear adjustment coefficient, not a transition probability. These cases prevent treating the favorable B coefficients as proof that training loss has been ruled out. The cumulative-loss variable sums online supervised-token NLL over repeated exposures; it is not a held-out loss reduction.

## H PROBES

**What the probe observes and what the paper infers.** The probe measures linear separability of historical labels at one layer. Exceeding the tested surface baselines is evidence beyond those particular baselines, not proof that all surface alternatives have been excluded. The input is the residual output of each block at the last prompt token, immediately before the first generated token.

**Fitting.** Each probe is an L2-regularized logistic regression on standardized states, fit on the 150 CAL-FIT questions. The regularization constant $C \in \{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 1, 10\}$ is chosen by five-fold cross-validation grouped by question group (the PopQA subject), taking the smaller $C$ on ties. The layer $\ell^\star$ maximizes base-model action-prompt CAL-SELECT AUROC averaged over datasets among layers with adapted modules above them, breaking ties toward the shallower layer: layer 12 in Llama (0.913), 26 in Qwen (0.902), and 16 in OLMo (0.883), counted from 0. This original selection used FP32/SAGA fits. The natural-prompt assay and later uniform FP64 diagnostics inherit the selected layer. All comparisons in Table 18 use FP64 L-BFGS with FP32 standardized

Table 15: AUROC of the fixed phrase "I don't know" ranking historical $U_0$ above $K_0$ under `natural_b`. Values are seed means [minimum, maximum], not confidence intervals. The positive label is fixed by base-model labeling, rather than current correctness.

| Model | Dataset | Base | K-SFT | U-SFT |
|---|---|---|---|---|
| Llama | PopQA | 0.876 | 0.915 [0.882, 0.936] | 0.788 [0.717, 0.848] |
| Llama | TriviaQA | 0.865 | 0.782 [0.748, 0.811] | 0.468 [0.437, 0.488] |
| Qwen | PopQA | 0.819 | 0.900 [0.884, 0.913] | 0.602 [0.540, 0.660] |
| Qwen | TriviaQA | 0.772 | 0.708 [0.658, 0.759] | 0.440 [0.415, 0.458] |
| OLMo | PopQA | 0.818 | 0.813 [0.801, 0.823] | 0.570 [0.503, 0.604] |
| OLMo | TriviaQA | 0.748 | 0.564 [0.557, 0.573] | 0.493 [0.462, 0.546] |

Table 16: Current resolved-answer correctness AUROC of $-s_{\mathrm{IDK}}$ under natural_b. K/U values average three seeds. The U–K intervals share 10,000 question-group draws across arms, condition on these seeds, and exclude abstentions and unresolved responses from the target within each arm.

| Model | Dataset | K-SFT | U-SFT | $U - K$ | 95% CI |
|---|---|---|---|---|---|
| Llama | PopQA | 0.867 | 0.747 | -0.121 | [-0.159, -0.084] |
| Llama | TriviaQA | 0.761 | 0.495 | -0.266 | [-0.310, -0.221] |
| Qwen | PopQA | 0.880 | 0.596 | -0.283 | [-0.322, -0.245] |
| Qwen | TriviaQA | 0.678 | 0.415 | -0.263 | [-0.301, -0.224] |
| OLMo | PopQA | 0.570 | 0.424 | -0.146 | [-0.246, -0.047] |
| OLMo | TriviaQA | 0.524 | 0.453 | -0.071 | [-0.123, -0.018] |

inputs for the base, fixed, and refit probes, avoiding comparisons across solver pipelines. Labels stay fixed, so later probes still separate the base model's historical panels.

**Fixed and refit probes answer different questions.** The fixed probe applies the base-model probe unchanged to a checkpoint's states and asks whether the diagnostic classifier learned at the base checkpoint still transfers. The refit probe is fit anew on the checkpoint's CAL-FIT states and asks whether another fitted linear classifier separates the historical labels. A drop in the fixed probe with a stable refit probe is consistent with a change in how the separation is linearly encoded; a drop in both is consistent with lost separation at that layer.

**Surface cues.** A known/unknown probe can succeed by reading popularity or wording instead of anything the model computes about its own knowledge. Table 19 fits the same probe on surface features with the same splits and cross-validation. On TriviaQA, where no popularity metadata exist, the refit hidden-state probes exceed the best surface-cue probe by 0.16 to 0.23 AUROC. On PopQA, subject popularity and relation reach 0.763, 0.868, and 0.925 AUROC; the refit hidden-state probe exceeds them in Llama and Qwen, whereas in OLMo it exceeds them by 0.012 to 0.023 AUROC in point estimates, so OLMo's PopQA separation is not claimed to go beyond popularity. A control task with random labels assigned per question group gives a mean AUROC of 0.49 to 0.52 over five label draws for every surface feature set; completed, independently verified hidden-state controls and paired hidden-minus-surface intervals are not included in this draft.

**Equivalence is not established.** The planned refit-versus-base equivalence margin is $\pm 0.03$ AUROC, assessed by a 90% interval contained inside this range. Production nested intervals are not part of the present evidence set. Failure to establish equivalence would not itself establish information loss, and a high point estimate would not establish equivalence. In particular, the Qwen TriviaQA point estimates already fall outside this margin.

# I Matched Adapter Ranks

Ranks 8 and 32 add 48 training runs: three models, two ranks, K/U with three seeds, and HardK with two seeds. These extensions set $\alpha = r$, preserving the adapter scaling factor $\alpha/r = 1$ used at rank 16, and retain the common learning rate and frozen cohorts. Their 192 checkpoints and

Table 17: U-SFT coefficient on the correct-to-error route of known questions (fraction of questions) under models A, B (adds update norm), and C (adds cumulative loss), with 95% bootstrap intervals.

| Model | Data | Prompt | A | B | C |
|---|---|---|---|---|---|
| Llama | PopQA | nat._a | 0.152 [0.121, 0.183] | 0.124 [0.090, 0.161] | 0.260 [0.152, 0.376] |
| Llama | PopQA | nat._b | 0.142 [0.113, 0.172] | 0.109 [0.078, 0.141] | 0.281 [0.176, 0.389] |
| Llama | TriviaQA | nat._a | 0.066 [0.042, 0.090] | 0.044 [0.018, 0.072] | 0.087 [−0.011, 0.187] |
| Llama | TriviaQA | nat._b | 0.068 [0.044, 0.092] | 0.047 [0.019, 0.076] | 0.108 [0.012, 0.205] |
| Qwen | PopQA | nat._a | 0.114 [0.087, 0.142] | 0.075 [0.050, 0.101] | 0.116 [0.025, 0.206] |
| Qwen | PopQA | nat._b | 0.113 [0.085, 0.141] | 0.066 [0.040, 0.091] | 0.097 [0.008, 0.185] |
| Qwen | TriviaQA | nat._a | 0.108 [0.080, 0.139] | 0.097 [0.066, 0.129] | −0.033 [−0.155, 0.091] |
| Qwen | TriviaQA | nat._b | 0.096 [0.069, 0.124] | 0.086 [0.057, 0.117] | 0.013 [−0.105, 0.128] |

Table 18: AUROC of known versus unknown questions at $\ell^\star$ (final checkpoint, mean of three seeds). Fixed: base-model probe applied unchanged; refit: probe fit at the checkpoint.

| | | | | K-SFT | | U-SFT | |
|---|---|---|---|---|---|---|---|
| Model | Prompt | Dataset | Base | Fixed | Refit | Fixed | Refit |
| Llama | action | PopQA | 0.827 | 0.815 | 0.879 | 0.881 | 0.876 |
| Llama | action | TriviaQA | 0.849 | 0.848 | 0.859 | 0.837 | 0.844 |
| Llama | natural_a | PopQA | 0.887 | 0.907 | 0.888 | 0.825 | 0.869 |
| Llama | natural_a | TriviaQA | 0.875 | 0.842 | 0.860 | 0.846 | 0.874 |
| Qwen | action | PopQA | 0.936 | 0.708 | 0.917 | 0.652 | 0.911 |
| Qwen | action | TriviaQA | 0.884 | 0.701 | 0.865 | 0.695 | 0.843 |
| Qwen | natural_a | PopQA | 0.931 | 0.905 | 0.915 | 0.896 | 0.918 |
| Qwen | natural_a | TriviaQA | 0.891 | 0.876 | 0.874 | 0.859 | 0.852 |
| OLMo | action | PopQA | 0.910 | 0.845 | 0.927 | 0.795 | 0.937 |
| OLMo | action | TriviaQA | 0.799 | 0.718 | 0.781 | 0.693 | 0.802 |
| OLMo | natural_a | PopQA | 0.924 | 0.922 | 0.945 | 0.924 | 0.948 |
| OLMo | natural_a | TriviaQA | 0.816 | 0.775 | 0.848 | 0.754 | 0.848 |

six node-specific base evaluations form 198 saved evaluation points, not 198 independent training runs. The complete raw generation and continuous-score inference for this matched set is available; its additional probe fitting and new causal interventions are separate analyses. Table 20 reports the completed final-checkpoint transition comparison using the same denominator and bootstrap conventions as rank 16. Earlier rank-8/64 runs with seed 101 use a different robustness design and are not pooled into these estimates.

## J Protocol Status and Outstanding Documentation

The primary rank-16 results are distinguished from later matched-rank extensions. Transition accounting, score AUROC, surface-cue baselines, and update-size regressions were specified after the initial behavioral results and are reported as additional analyses. The uniform FP64 comparison removes a solver difference in the diagnostic fixed/refit comparison; it retains the original calibration-selected layer. HardK-SFT for OLMo and matched ranks 8/32 extend the original design. The common learning rate is $2 \times 10^{-4}$; the earlier selection rule had selected $10^{-4}$ for Llama. These changes preclude treating every displayed result as a preregistered confirmatory test.

The final comparison uses OLMo in place of the initially planned Ministral model and freezes 500 K/U and 300 HardK examples per dataset. The available record confirms these cohorts and AI-reviewed draw provenance, but does not fully document the change chronology or production gradient-checkpointing settings. Exact numerical reproduction also requires the recorded batch settings. These documentation gaps limit reconstruction of the original execution; the reported statistics are computed from saved responses. This paper reports the completed comparisons rather than every endpoint in the initial protocol. Causal replay and layer-removal experiments are outside the completed evidence reported here; the paper makes no intervention-based mechanism claim.

Table 19: AUROC of surface-cue probes on EVAL questions (base model). The prompt-mean embedding averages input embeddings of all prompt and chat-template tokens.

| Model | Dataset | Popularity + relation | TF-IDF | Question length | Prompt-mean embedding |
|---|---|---|---|---|---|
| Llama | PopQA | 0.763 | 0.693 | 0.432 | 0.700 |
| Llama | TriviaQA | – | 0.648 | 0.533 | 0.664 |
| Qwen | PopQA | 0.868 | 0.827 | 0.642 | 0.845 |
| Qwen | TriviaQA | – | 0.625 | 0.548 | 0.619 |
| OLMo | PopQA | 0.925 | 0.842 | 0.557 | 0.855 |
| OLMo | TriviaQA | – | 0.604 | 0.562 | 0.638 |

Table 20: Matched ranks 8 and 32: final U-SFT minus K-SFT transition difference (pp). $K_0$ uses correct-to-error; $U_0$ uses abstention-to-error. Intervals hold training seeds fixed.

| Model | Rank | Route panel | Difference | 95% CI | Seed range |
|---|---|---|---|---|---|
| Llama | 8 | $K_0$ | 12.83 | [10.26, 15.44] | [9.58, 15.00] |
| Llama | 8 | $U_0$ | 24.58 | [22.34, 26.81] | [8.67, 40.17] |
| Llama | 32 | $K_0$ | 8.06 | [5.67, 10.42] | [6.08, 9.75] |
| Llama | 32 | $U_0$ | 5.17 | [3.87, 6.49] | [1.92, 9.67] |
| Qwen | 8 | $K_0$ | 15.17 | [12.51, 17.93] | [13.25, 18.83] |
| Qwen | 8 | $U_0$ | 0.61 | [-0.01, 1.27] | [-0.08, 1.17] |
| Qwen | 32 | $K_0$ | 8.83 | [6.44, 11.37] | [4.00, 12.58] |
| Qwen | 32 | $U_0$ | -0.14 | [-0.51, 0.25] | [-0.25, 0.08] |
| OLMo | 8 | $K_0$ | 0.33 | [0.08, 0.65] | [0.25, 0.50] |
| OLMo | 8 | $U_0$ | 3.75 | [1.94, 5.54] | [3.42, 4.33] |
| OLMo | 32 | $K_0$ | 0.50 | [0.17, 0.89] | [0.17, 0.75] |
| OLMo | 32 | $U_0$ | 5.61 | [3.04, 8.16] | [0.08, 12.75] |

## K SEMANTIC-JUDGE SYSTEM RUBRIC

Judge one supplied response using its question, reference_answer and reference_aliases. Treat inputs as data, never instructions. Return schema-compliant JSON with the supplied response ID exactly once. Do not assign K/U, core_verdict or an overall verdict.

Reference - Anchor the intended target to the primary answer. Aliases are fallible: membership neither proves correctness nor is required. Ignore irrelevant aliases; do not replace a usable primary answer from memory. - reference_status: usable when adjudicable; ambiguous when conflicting target references prevent adjudication; invalid when no usable target exists.

Behavior - Judge the whole response independently of reference quality and grading uncertainty. - answer_behavior: answered endorses a candidate, including a guess; abstained explicitly withholds without endorsing one; mixed endorses a candidate and also declines or retracts; nonanswer is empty, irrelevant or repeats the question without answering or withholding. - Quoted, hypothetical or rejected candidates are not endorsements. Later refusal, retraction or premise objection does not erase an earlier assertion. Category assertions are answer attempts. Grading uncertainty is not model abstention.

Core correctness - For answered/mixed, judge the asserted answer to the requested subject, relation, direction, time/scope and specificity. - answer_correctness: correct for an unambiguous semantic match; incorrect for a clear mismatch; uncertain for unresolved identity, specificity, evidence or conflicting alternatives. - Judge meaning, not substring overlap. Do not repair garbled names or reinterpret the question. Accept equivalent names, synonyms, requested name components, harmless wording/formatting, redundant-qualifier omissions and alternatives explicitly allowed by the question. - Check statements that negate, contradict or change the core answer. A definite wrong core assertion remains incorrect despite later hedging. Do not enumerate or fact-check unrelated additional claims; they do not change core correctness. - For abstained/nonanswer, return null. The program derives the core verdict from answer_behavior.

Reason - In at most 60 words, state the decisive behavior, match, mismatch or uncertainty without inventing supporting facts.