EVIDENCE, ADAPTATION, AND TAIL RISK IN VERIFIABLE-REWARD TAXONOMY ALIGNMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

We audit verifiable-reward and supervised post-training for institutional taxonomy alignment: mapping national occupational records into ISCO-08 across 878 country/entity–system–source cells. On a development-exposed benchmark, five row-only post-SFT variants leave worst-cell reliability within three points of SFT, and five checkpoints share a 73-cell worst-decile core. Continued SFT on evidence-augmented prompts with gold targets only (three seeds) changes this pattern: worst-cell conditional value-at-risk increases by 0.072–0.081 relative to plain-prompt SFT, while overall strict accuracy falls by 4.7 points to 74.2–74.4%. The change is heterogeneous: cells defined as difficult from development records improve from 14.3% to 24.1–24.2%, while easy cells fall from 99.3% to 91.3–91.6%. At the coverage of the explicit `ABSTAIN`-target control, confidence-selected gold-only SFT reaches 96.1–96.2% selective accuracy versus 84.0–84.3%; at the strongest selective RL policy's 70.8% coverage, it reaches 92.6–92.8% versus 63.3%. On a prospectively frozen, single-source transfer of 976 queries, five labeled references raise gold-only accuracy from 9.1–9.9% to 28.9–29.6% at fixed weights; below a post-hoc nearest-code lookup (35.8%). These results identify recipe-specific tradeoffs between institutional-tail accuracy, aggregate error and deferral. They do not establish a general information-theoretic limit of verifiable rewards, or isolate training evidence from the additional optimization in continued SFT.

1 INTRODUCTION

Reinforcement learning with verifiable rewards (RLVR) trains models from automatically checked samples (Lambert et al., 2024; Shao et al., 2024; DeepSeek-AI, 2025). A verifier scores outputs; it does not supply missing source evidence or an abstention sample. We study whether such feedback improves reliability when errors concentrate in particular institutional groups.

The testbed is taxonomy alignment: national occupation systems, survey codebooks, and European job-matching bridges mapped into ISCO-08 (International Labour Office, 2012; le Vrang et al., 2014). This connects outcome verification, selective prediction and occupation coding: can a reward improve institutional tail reliability when the policy cannot recover the mapping from the isolated row? Figure 1 shows the persistent tail and context diagnosis; Figure 2 compares supervision and RL after exposing that context and a deferral action.

This is a diagnostic and mechanism study. Plain-prompt SFT with confidence thresholding remains the deployment reference. The augmented-input comparison asks what post-training learns from exposed evidence, whether that requires RL or abstention targets, and which institutional failures remain.

Contributions.

1. **A stable open-model tail not recovered by tested post-training.** Five forced-choice checkpoints share a 73-cell worst-decile core, tail-cell SFT does not remove it, and five post-SFT variants on Qwen2.5-3B change worst-cell CVaR@10% by -0.020 to $+0.010$ (Section 3).
2. **Source-evidence diagnostics of a persistent policy-specific tail.** Nearest-code and agreement-gated lookup diagnostics show that D_s, the other rows of the same source system, carries signal

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.

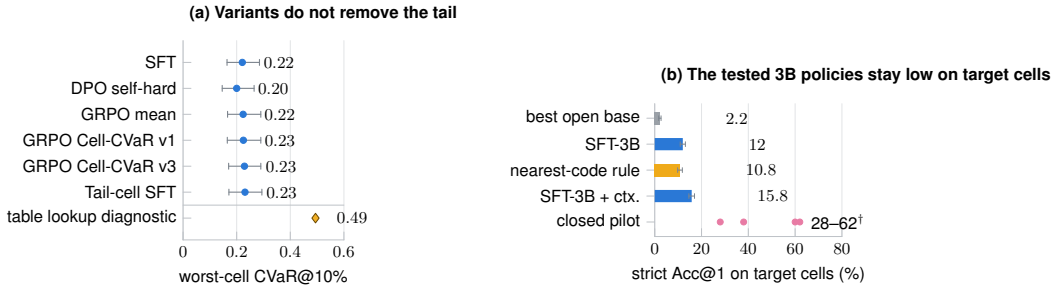


Figure 1: **The tail is stable across the forced-choice ladder and points to a missing-context gap. (a)** The SFT baseline plus five post-SFT variants leave a worst-cell tail whose 95% cell-bootstrap intervals (whiskers) overlap heavily; levels are in Table 2 and the pairwise deltas, all consistent with zero except DPO, in Section 3. The diamond *table lookup* marker is an upper-bound diagnostic that reuses the benchmark's own construction sources (Section 3.3), not a trainable policy, and is separated from the checkpoint ladder accordingly. **(b)** The best small open base model reaches 2.2%; SFT and a nearest-code rule stay below 16% (all bars use all 3,716 target rows; the rule scores 14.6% on the 74% of rows with a same-system neighbour); four closed-model pilots span 28–62% (†50-row pilots, one dot per model; intervals and caveats in Section 3 and Table 13).

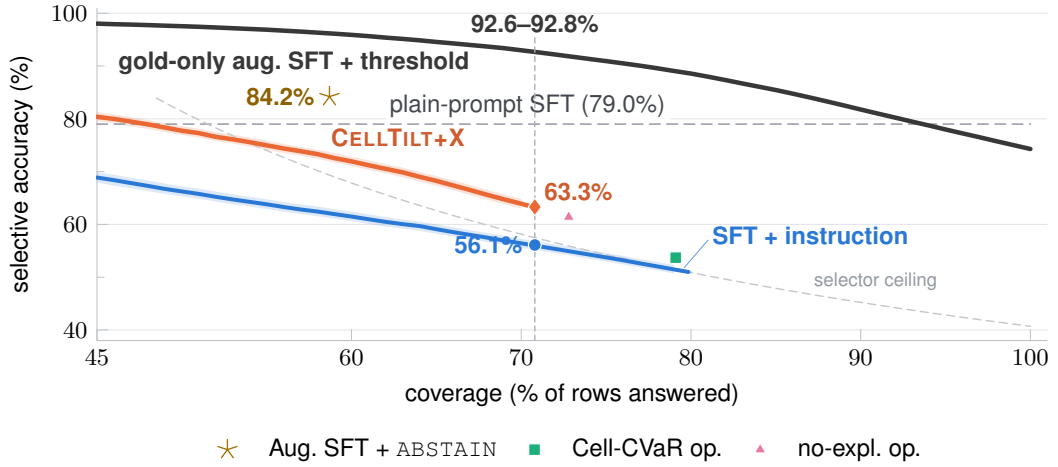


Figure 2: **With matched augmented inputs, gold-only supervised adaptation with confidence thresholding lies above every tested abstention-trained policy.** The dashed vertical guide reads off the three policies at CellTilt+X's own coverage. The gold-only curve is the mean of three seeds (band: seed range); it answers every row and is thresholded by sequence log-probability. Marked operating points use the original prediction pass; the curves use separate confidence passes and stop at their own answering rates (bands: 95% row-bootstrap). Plain-prompt SFT at full coverage is a separate deployment reference, and its own confidence curve lies higher still (Appendix Figure 7, which also carries the Aug. SFT + ABSTAIN curve); the selector ceiling bounds filtering the original SFT answers.

beyond the isolated row x for the tested policies. Two closed models answer about 60% of a 50-row pilot from x alone, so the observed failure is specific to the tested policies and protocol, not an impossibility implied by the row (Sections 3 and 3.3).

3. **Elementary estimator-level observations diagnose why direct abstention supervision is absent.** For centered group-relative estimators, zero-variance groups carry no gradient and full-weight within-group CVaR advantages are non-positive. Thus an unsampled ABSTAIN action receives no positive sample-specific learning term. An ablation isolates the estimator swap; one-sided forced ABSTAIN then adds 1.5–2.3 points of selective accuracy and no worst-cell gain in these single-run ablations (Sections 4 and 6.3).
4. **Supervised adaptation to exposed evidence moves the tail; abstention targets and the tested RL recipes are not needed for the selective gains.** Gold-only continued SFT on the augmented

prompt (three seeds) reaches 74.2–74.4% strict accuracy at full coverage and raises worst-cell CVaR over plain-prompt SFT by 0.072–0.081, trading 4.7 points of overall accuracy. With confidence thresholding it reaches 96.1–96.2% where explicit `ABSTAIN` targets reach 84.0–84.3%, and 92.6–92.8% where the strongest RL policy reaches 63.3%; rerun from the adapted model, that recipe adds a seed-sized 0.34 points and no abstention (one seed; Section 6, Appendix J).

2 Problem Setup

Institutional taxonomy alignment. An example is a source occupational record $x = (\text{country/entity}, \text{source system}, \text{national code}, \text{national title}, \text{source category}, \ldots)$ with an official set of target ISCO-08 unit-group codes $Y(x) \subseteq \mathcal{Y}$, where $\mathcal{Y}$ is the set of valid four-digit ISCO-08 unit groups, and a designated primary code $y^\star(x) \in Y(x)$ (the code the source table lists first; 33% of test rows have $|Y| > 1$, and Section 3 reports lenient accuracy $\hat{y} \in Y$ alongside strict). The model outputs a four-digit code, and we parse the first valid code in the response. ISCO-08 is a four-level hierarchy (major, sub-major, minor, unit group), which we use for reward shaping and diagnostics. A prediction can be strictly correct ($\hat{y} = y^\star$), lenient ($\hat{y} \in Y$), share a major group with a gold code, be a valid but wrong code, or be invalid.

Cells. The reliability unit is the cell $c(x) =$ country/entity $\times$ source system $\times$ source category, where source categories are official national crosswalks, multilingual bridges from intergovernmental organizations (IGO) or EURES, survey codebooks, structural expansions of coarser codes, and composed bridges. Appendix Table 7 summarizes the splits (878 test cells, 162 country/entity labels; one source category per system, so cells are (country/entity, system) pairs).

Metrics. We report micro strict Acc@1, macro-cell strict Acc@1, **worst-cell CVaR@10%** (mean strict accuracy over the lowest-scoring 10% of cells), and invalid-code rate. Strict scoring uses the historical parser and normalization; it does not enforce a single-code response format (Appendix B). For selective prediction (Chow, 1970; Geifman & El-Yaniv, 2017) a response is an abstention if it contains `ABSTAIN` and no code; we report **coverage** (fraction answered), **selective accuracy** (strict accuracy among answered rows), and worst-cell CVaR@10% of per-cell selective accuracy over cells with at least one answered row. This last metric is fragile (floor at zero, policy-dependent cell set; Jones et al., 2021), so we also report cell-size sensitivities and a variant over the fixed set of all 878 cells that scores fully-abstained cells as zero. Thresholds on answered rows retain different cells for each policy; we therefore report the number retained at each threshold. Uncertainty is quantified with 2,000-sample bootstraps (Efron, 1979) (row-level and paired for accuracies and coverage, cell-level for cell metrics) and McNemar's test (McNemar, 1947) on rows two policies both answer.

Data and models. We use an anonymized CrosswalkBench-Global style collection of official occupational crosswalks (Anonymous, 2026). The metadata field called *country* is a source label, not a sovereign-state count: it includes countries, territories, survey programs, and multilateral entities. The collected pool (1.86M raw mapping rows before grouping and filtering; see Table 1) contains 1.26M rows spanning 220 such country/entity labels, 1,006 unique source-system names, 1,112 country/entity–system–source cells, and 68 primary languages. Experiments use cell-balanced 50,000/5,000/20,000 train/dev/test subsets from the five non-identity source families. Test rows are disjoint from training rows but mostly share their source systems, so the main split measures in-system record completion; the 1,369 national-crosswalk test rows are the exception, with no system seen in training. All open-model experiments start from Qwen2.5-3B-Instruct (Qwen Team, 2025) with low-rank adaptation (LoRA) (Hu et al., 2021). Supervised fine-tuning (SFT) trains one epoch on the primary code and is the reference policy for every later stage (Appendix B).

3 The Institutional Tail Persists under Tested Open-Model Post-Training

This section treats the forced-choice null as a protocol diagnostic, while keeping the optimizer-budget caveat visible. The claim would be weakened if lenient scoring removed the tail, if different checkpoints failed on different cells, or if same-system evidence did not move the target cells. The results below point the other way: the ladder is flat under strict and lenient scoring, the worst cells

Table 1: Dataset scope. The benchmark is a model-reliability stress test rather than a single-country occupation-coding task: the collected pool (grouped from the benchmark's 1.86M raw mapping rows, excluding near-identity ISCO-08 mappings) spans 220 country/entity labels (countries, territories, survey programs, and multilateral entities), 68 primary languages, and heterogeneous institutional source systems. "System names" counts unique source-system names; "Cells" counts country/entity–system–source units, the audit unit used in the paper. The experiment splits are cell-balanced subsets; "Multi" is the fraction of rows with more than one official target code.

Split / pool	Rows	Country/entity labels	System names	Cells	Languages	Multi
Collected pool	1,262,264	220	1,006	1,112	68	8.6%
Train, balanced	50,000	198	849	946	67	38.3%
Dev	5,000	157	656	720	57	33.6%
Test	20,000	162	807	878	59	33.1%

overlap heavily across checkpoints, and table/context diagnostics move the tail much more than the post-SFT variants.

3.1 SFT PLUS FIVE VARIANTS, ONE TAIL

Table 2 reports the forced-choice ladder on the test split: the SFT baseline, DPO, three GRPO variants, and a tail-cell SFT intervention. For the GRPO variants, the verifier reward is $+1$ for any code in $Y(x)$, $+0.2$ for a wrong code in the right major group, -0.2 for the wrong major group, and -1 for an invalid code. A direct preference optimization (DPO) (Rafailov et al., 2023) construction with *copy-code* negatives (the national code when it parses as a non-gold ISCO-08 code) collapses despite near-perfect pair separation (dev strict Acc@1 $0.78 \rightarrow 0.29$), because the policy never produces that negative. The ladder therefore uses *self-hard* negatives, the SFT checkpoint's own wrong predictions. Mean group-relative policy optimization (GRPO) and two Cell-CVaR variants (within-group CVaR advantages, the second with cell-level reweighting) preserve SFT and reduce invalid codes, and a second SFT pass on cleaned rows from the weak cells does not help either. The change in worst-cell CVaR relative to SFT is -0.020 $[-0.032, -0.007]$ for DPO and between $+0.003$ and $+0.010$ for the other four post-SFT variants, none of these four with an interval excluding zero. Levels fall with the minimum cell size (0.22, 0.14, 0.11 at 1, 10, 20 rows) but the spread stays within 0.03 (Appendix Table 8). The RL runs moved aggregate strict accuracy by at most 0.3 points and single slices by about one point at most (Appendix Table 11): a null for this training budget, not for every optimizer; the probes below provide complementary diagnostics on the development-exposed benchmark.

The null is not only a strict-primary-label artifact. Under lenient scoring ($\hat{y} \in Y(x)$), the same ladder remains flat: SFT is 0.800, GRPO-mean 0.804, the two Cell-CVaR variants 0.803 and 0.802, and tail-cell SFT 0.793. On the 73 target cells, SFT rises only from 12.0% strict to 14.8% lenient. This does not eliminate label ambiguity, but it makes it unlikely to explain the stable worst-cell core by itself.

Table 2: Forced-choice ladder on the 20k test split (Qwen2.5-3B, LoRA; 95% cell-bootstrap intervals). No post-SFT variant gives a clear tail improvement; the agreement-gated lookup is a benchmark-source upper-bound diagnostic (Section 3.3), not a trainable policy.

Method	Strict	Macro cell	Worst CVaR @10%	Invalid
Trainable post-training policies				
SFT	0.790	0.884	0.221 [0.16, 0.29]	2.6%
DPO (self-hard negatives)	0.780	0.879	0.200 [0.15, 0.26]	2.5%
GRPO, mean advantage	0.793	0.886	0.224 [0.17, 0.29]	2.2%
GRPO, Cell-CVaR (v1)	0.792	0.884	0.225 [0.17, 0.29]	2.3%
GRPO, Cell-CVaR (v3, cell-reweighted)	0.792	0.880	0.229 [0.17, 0.29]	2.1%
Tail-cell SFT	0.783	0.882	0.231 [0.17, 0.29]	2.7%
Diagnostic with external table evidence				
SFT + agreement-gated table lookup (diagnostic)	0.888	–	0.494	–

The failures are not spread over different cells. Taking the worst-decile cells (87 of 878) of each checkpoint, pairwise Jaccard overlap is 0.76–0.91 (Appendix Figure 6), and 73 cells are in the worst decile of all five checkpoints predating tail-cell SFT (70% of their union). We call these the *target cells*: 3,716 test rows (2,041 multilingual IGO bridges, 740 national crosswalks, 638 survey codebooks, mostly a West-African regional list, 195 structural expansions, 102 composed bridges). Defined on the test split, they are a diagnostic partition, not an evaluation set: later comparisons use the full split and the RL-excluded systems, and we flag design choices (oversampling, prompt design) informed by this diagnosis. This adaptive use of the test split is a limitation of the current artifact rather than a final validation protocol: the clean next step is to freeze the diagnosis and method on a diagnostic split, then re-run the final comparisons on a fresh held-out test split (Section 8).

3.2 Row-only open-model prompts do not elicit exact mappings; closed-model pilots partly do

On the 3,716 target-cell rows (Figure 1b, Appendix Table 12), zero-shot base models score 0.5–2.2% (Qwen2.5-3B to Qwen3.5-9B; Phi-3.5-mini 1.3%; Mistral-7B emits invalid output on 99% of rows). Invalid rates of 36–75% mean these numbers mix format failure with failed exact-code elicitation; major-group accuracy, the fairer signal, scales from 8% to 30% while exact accuracy stays below 2.2%. The SFT model reaches 12.0% $[11.0, 13.1]$. In a 50-row pilot on the same cells, two of four closed API models score about 60% (62% $[48, 74]$ and 60% $[46, 72]$), while the other two score 38% and 28%; the SFT model scores 18% on the same rows (Appendix Table 13).[1] The observations suggest a partly *elicitable* regime, where a large pretrained model often gives the exact mapping, and a regime that is *unrecovered for the 3B policy we post-train*, where verifier feedback alone does not elicit the exact unit-group mapping (Yue et al., 2025). This is an operational elicitation claim, not a statement about what no larger model or different training procedure could contain. Zero-shot probes up to 9B are consistent with this diagnosis, but they do not establish what a post-trainable 9B policy would learn. Two caveats bound the claim: the mapping may be present but not extracted by these models, prompts, and training budgets, and part of the residual error is label ambiguity (expert inter-agency agreement on ISCO major groups is only $\kappa \approx 0.7$, Massing et al., 2019).

3.3 Source-system context provides diagnostic signal beyond the row

To separate a row-only optimizer failure from an input-evidence boundary, we ask whether same-system evidence outside the isolated row predicts the target cells. This is a diagnostic rather than a proof against every optimizer alternative, and it uses training data only (Appendix Table 10). Within a (country/entity, system), the national code space is a semantically ordered tree, and $(\text{country/entity}, \text{system}, \text{code}) \rightarrow$ ISCO-08 is a deterministic function on 99.9% of training rows. For the 74% of target-cell rows with a same-system training neighbour, the training row with the numerically nearest code shares the gold major group 55.9% of the time (random same-system row: 17.9%) and the exact unit group 14.6% (0.9%), against 13.6% for the fine-tuned SFT model on the same rows (10.8% vs. 12.0% unconditionally). In aggregate the rule beats term frequency-inverse document frequency (TF-IDF) title retrieval at every K (gold in the top-10: 24.5% vs. 16.1%). A dense multilingual title encoder closes most of the gap as K grows (9.9% vs. 14.6% at $K{=}1$, 23.7% vs. 24.5% at $K{=}10$), and title retrieval is stronger on structural expansions; part of the locality is inherited from national systems that extend ISCO-08. A row-only inference path does not directly condition on this neighbourhood unless it is supplied. Row-level training may internalize some system regularities, but the diagnostic signal is better described as evidence in D_s, the other rows of the same system, in addition to x (Appendix Table 20). This supports a narrower claim than "the code geometry fully explains the tail": source-system context provides useful diagnostic signal beyond the isolated row for the tested policies.

Benchmark-source correspondence tables give an upper-bound diagnostic for how much of the target tail is externally resolvable when the diagnostic is allowed to consult source-system evidence. Three lookups (the regional survey code list, the ISCO-88→ISCO-08 correspondence with prefix narrowing, exact match over 588 standardized crosswalk files), accepted only when they agree with

[1] The pilot only shows that exact mappings are not impossible for all models on these rows; a full closed-model evaluation is future work. In the sampled rows, closed models succeed on genuine multilingual titles (French *Pharmacien*, *Tanneur*) and fail on fine unit-group boundaries, a pattern consistent with semantic ambiguity but insufficient to distinguish reasoning from memorized crosswalk knowledge.

the *training* labels of the same system at least 90% of the time, cover 57 of the 73 target cells (2,310 rows) and lift worst-cell CVaR from 0.221 to 0.494 for SFT and by 25–28 points for every other checkpoint. The tables are among the benchmark's own construction sources, so this is a diagnostic, not a method: the table-access upper bound is much larger than the spread across the SFT-plus-variants ladder.

Injecting this context into the prompt without training helps on the target cells but hurts the easier full-split distribution: base Qwen2.5-3B rises from 0.5% to 13.0% on the target cells and SFT from 12.0% to 15.8%, but the un-adapted SFT model given the augmented prompt (forced-choice, no `ABSTAIN` option) on the full split falls from 0.790 to 0.450 strict Acc@1 (Appendix C). Context can add evidence that the row-only prompt did not expose to the tested policies, but it also changes the input distribution. A policy in this augmented regime must therefore learn when to use this context and, we argue next, when to decline.

4 Estimator Limits for Unsampled Abstention

GRPO-style methods score G sampled completions relative to their group (Shao et al., 2024). Two elementary observations explain the tested updates (formal statements and derivations in Appendix A). First, equal-reward groups have zero centered advantages; cell reweighting preserves zero and dynamic sampling drops these groups (Remark 1; Sagawa et al., 2020; Yu et al., 2025; Kool et al., 2019). Second, full-weight within-group CVaR advantages are non-positive: they suppress low-reward samples but never directly promote abstention, and they vanish whenever a group's two lowest rewards tie (Remark 2; Greenberg et al., 2022).

An absolute baseline can penalize all-wrong samples (Zhu et al., 2025), but a positive target for `ABSTAIN` requires that action in the training batch. Injecting it supplies the action; its mean advantage is positive only when the sampled mean reward is below the abstention reward (Remark 3). The target therefore still depends on sampled-group feedback, unlike direct supervision. A negative forced-completion term can be unbounded below as its probability vanishes (Remark 4; Welleck et al., 2020; Le Roux et al., 2025). The ablations test the estimator swap and a one-sided forced update separately.

5 Mechanism Test: Cell-Level Risk and Abstention Exploration

Selective reward. The policy may output a code or `ABSTAIN`:

$$r(x, o) = \begin{cases} +1 & \hat{y}(o) \in Y(x), \\ a & o = \texttt{ABSTAIN}, \\ -1 & \text{otherwise (wrong or invalid code).} \end{cases} \tag{1}$$

Under expected reward the optimal policy answers iff $\Pr[\text{correct} \mid x] > (1+a)/2$ (Chow, 1970), so a is the operating-point knob; we use $a = 0$, and the instruction states that `ABSTAIN` is allowed. The verifier rewards any code in $Y(x)$; strict accuracy scores $y^\star(x)$, the supervised target, so Section 6 also reports lenient scoring.

Estimator. We use normalized mean advantages, multiplied by clipped, mean-one exponential cell weights that amplify below-average cells (Li et al., 2021; Sagawa et al., 2020; Oren et al., 2019). Appendix A.1 gives the weights and objective. The no-exploration ablation retains these weights and changes only the within-group estimator relative to Cell-CVaR.

One-sided forced-`ABSTAIN` exploration. Before scoring, the last of the $G = 8$ rollouts of every group is replaced by the fixed completion `ABSTAIN`⟨eos⟩ (Appendix Figure 3). The forced rollout receives an unclipped REINFORCE update with importance ratio one; its negative advantages are clamped to zero. It is promoted when the sampled mean reward is negative, while sampled on-policy abstentions can still be penalized when they underperform their group. This separates action availability from target estimation: the action is injected, but the update depends on seven verifier-scored samples. The forced term is a weighted cross-entropy toward `ABSTAIN`, linking it to refusal tuning and off-policy guidance (Zhang et al., 2024; Xu et al., 2024; Yan et al., 2025; Le Roux et al., 2025). Appendix A.1 gives the objective (Equation 3) and its derivation; Algorithm 1 gives one step.

Inputs. Selective policies share the augmented prompt: the row, five same-system neighbours, correspondence candidates where available and an ABSTAIN instruction (Appendices C–D). We exclude 94 of 946 systems from selective training (5,506 training rows; 2,139 test rows). These RL-excluded systems are not unseen: SFT trained on them, and 1,705 test rows retain training neighbours. This measures transfer of deferral, not unseen task knowledge.

6 EXPERIMENTS

Protocol. RL runs start from the SFT model, which was never trained on the augmented prompt, and train 3,000 steps (Appendix Table 5); CELLTILT+X has 3 seeds, the other RL arms one. Table 3 compares Cell-CVaR ($\alpha = 0.25$, $\lambda = 1$), mean advantages without exploration, CELLTILT+X, supervised augmented SFT and SFT + instruction. Structural filters, token-confidence thresholds and an oracle selector provide selection controls. The supervised controls postdate the three predictions for CELLTILT+X (Appendix G) and use no RL rollouts or verifier sampling. Both continue SFT on the same augmented training rows, three seeds each. The gold-only control keeps the gold code as every target; the ABSTAIN-target control replaces it with ABSTAIN when candidate evidence is present and does not contain the gold code (per-seed results in Appendix J).

Table 3: Selective prediction on the 20k test split with matched augmented inputs (95% bootstrap intervals: row-level for selective accuracy, cell-level for worst-cell CVaR; "strict" counts abstentions as errors). Conditional CVaR excludes fully-abstained cells; Table 8 reports cell counts and the fixed-cell variant.

Policy	Coverage	Sel. Acc@1	RL-excluded cov. / acc.	Cond. CVaR @10%	Strict
SFT + instruction	79.6%	51.1% [50.3, 51.8]	76.3% / 52.5%	0.000 [0.000, 0.011]	0.407
+ strictest rule	56.4%	59.9%	–	0.000	0.338
Aug. SFT + gold-only targets (seed 17)	100.0%	74.2% [73.6, 74.8]	100.0% / 78.0%	0.301 [0.246, 0.361]	**0.742**
Aug. SFT + supervised ABSTAIN targets	58.7%	**84.2%** [83.5, 84.8]	57.3% / 90.9%	0.274 [0.174, 0.374]	0.494
Cell-CVaR selective	79.1%	53.7% [52.9, 54.4]	75.5% / 56.0%	0.002 [0.000, 0.019]	0.425
+ strictest rule	57.6%	61.9%	–	0.000	0.357
Mean + cell weights, no exploration	72.8%	61.4% [60.6, 62.2]	69.8% / 64.7%	0.019 [0.002, 0.059]	0.447
CELLTILT+X	70.8%	63.3% [62.5, 64.1]	67.9% / 67.4%	0.019 [0.001, 0.058]	0.448
+ invalid-code rule	67.2%	66.7%	–	0.058	0.448
+ strictest rule	54.9%	69.6%	–	0.000	0.382

6.1 MATCHED-INPUT MECHANISM RESULT

Plain-prompt SFT answers everything at 79.0% strict accuracy and has AURC 0.030 under sequence-log-probability thresholding (Appendix K). Table 3 is therefore not a deployment recommendation; it asks which estimator changes explain the matched-input selective behaviour and whether these matched-input gains require RL; the supervised controls show that they do not.

Augmented SFT with supervised ABSTAIN targets reaches 58.7% coverage and 84.2% selective accuracy [83.5, 84.8], with strict accuracy 0.494 and 90.9% selective accuracy on RL-excluded systems. Relative to this supervised control, CELLTILT+X answers 12.1 percentage points more rows but has 20.1–21.6 percentage points lower selective accuracy and 4.1–5.1 points lower strict accuracy. The control fully abstains in 283 cells (versus 58 for CELLTILT+X), so its high answered-cell CVaR (0.274 versus 0.019) conditions on different cell sets. Across all 878 cells, scoring fully-abstained cells as zero gives 0.000 for the control and 0.000 for CELLTILT+X (Appendix Table 8); neither policy resolves the all-cell tail.

Given the same prompt, SFT already abstains on 20.4% of rows at 51.1% selective accuracy. The Cell-CVaR policy adds 2.6 points $[2.3, 2.9]$ with an abstention set that overlaps SFT's at Jaccard 0.94: abstention came from the instruction, not training, as Remark 2 predicts. The no-exploration ablation reaches 61.4%, isolating the effect of replacing within-group CVaR with mean advantages (cell weights present in both). CELLTILT+X reaches 63.3% at 70.8% coverage, 12.2 points $[11.7, 12.8]$ over SFT and 9.6 $[9.1, 10.1]$ over the CVaR policy. This is not just lower coverage: at 70.8%, a perfect selector over SFT + instruction reaches at most 57.4% (60.0% for the CVaR policy), and strict Acc@1 rises from 0.407 to 0.448 $[+3.7, +4.5]$ (Appendix Table 16).

Lenient scoring. Scored as the verifier rewards ($\hat{y} \in Y(x)$), CELLTILT+X reaches 68.8% and the ABSTAIN-target control 85.9%: their gap narrows from 20.9 to 17.1 points, the gap in correct answers over all rows from 4.6 to 1.7, and the forced rollout's gain over the no-exploration ablation grows from 0.1 to 1.0 points.

Confidence-threshold baselines for the RL policies are in Appendix K; AURC is 0.030 for plain-prompt SFT and 0.058–0.059 for the gold-only control.

6.2 GAINS MIX PROMPT ADAPTATION AND SELECTION

Gold-only control. Continued SFT with the gold code as every target answers all rows at 74.2–74.4% strict accuracy (40.7% for SFT + instruction), recovering most of the matched-input deficit through continued training. Ranked by sequence log-probability and cut at the paired ABSTAIN-target model's coverage, it reaches 96.1–96.2% selective accuracy against 84.0–84.3% (90.2–90.7% with first-token probability), and 92.6–92.8% at CELLTILT+X's coverage. It answers 60.8–61.5% of the rows the ABSTAIN-target model declines correctly: explicit abstention targets are not the source of the supervised gain, and the label rule discards answerable rows.

Where the trade-off occurs. On 39 development-defined difficult cells, test accuracy rises from 14.3% to 24.1–24.2%, while easy-cell accuracy falls from 99.3% to 91.3–91.6%. At 70% automated coverage, plain SFT leaves 10.6 wrong answers per 1,000 inputs versus 48.5–49.6 for gold-only, but fully defers 59 cells versus 5–8. Thus lower aggregate error and broader institutional coverage favor different policies. These are exploratory comparisons on development-exposed splits (Appendix L).

For CELLTILT+X, 7.9 of the 12.2 points of selective-accuracy gain come from rows both policies answer (63.2% vs. 55.4%, McNemar $p < 10^{-100}$) and the rest from selection; its abstention set overlaps SFT's at Jaccard 0.69 (Appendix Figure 5, Table 14).

6.3 ABLATION: ESTIMATOR VERSUS EXPLORATION

Training dynamics (Appendix E) match the remarks: zero-variance groups fall to 0.30 for CELLTILT+X versus 0.78 for the CVaR policy (a lower bound on its zero-gradient groups, Remark 2), and the forced ABSTAIN signal decays as on-policy abstention rises. Appendix Table 6 separates the two changes. Replacing only the within-group estimator already reaches 61.4% selective accuracy at 72.8% coverage and worst-cell CVaR@10% of 0.019 $[0.002, 0.059]$. Adding one-sided forced ABSTAIN raises selective accuracy to 63.3% and RL-excluded accuracy by 2.7 points $[1.5, 4.0]$ (McNemar $p < 10^{-4}$), with no distinguishable worst-cell difference (delta $[-0.016, 0.013]$). The two-sided variant with the CVaR estimator shows the failure mode of Remark 4; it is a diagnostic, not a one-factor test of two-sided forcing under mean advantages.

6.4 REMAINING FAILURE MODES

Over all 878 test cells, with fully abstained cells scored as zero, worst-cell CVaR@10% is 0.000 for every abstaining policy. The ABSTAIN-target control fully abstains in 283 cells and CELLTILT+X in 58, so their conditional values summarize different cell sets (Appendix Table 8).

Within the RL comparisons, conditional CVaR@10% rises from 0.000/0.002 to 0.019 $[0.001, 0.058]$, below our pre-specified threshold of 0.05, and CELLTILT+X does not improve on the no-exploration ablation; target-cell strict Acc@1 is unchanged (0.129 vs. 0.135). The national-crosswalk slice exposes failed deferral under jointly evidence-free and out-of-system inputs: all 1,369 rows come from systems absent from the training split, their prompts carry no reference or candidate block, and every policy answers nearly all of them (Appendix Table 14). Abstention is associated with input regime: on the 1,699 rows without an evidence block, coverage is 99.8% for CELLTILT+X and 99.9% for the ABSTAIN-target control, against 68.1% and 54.9% where a block is present.

The tail moves under supervised adaptation. At full coverage, the gold-only control raises worst-cell CVaR@10% from 0.221 to 0.292–0.301 (paired cell-bootstrap differences 0.072–0.081, intervals within $[+0.033, +0.119]$); the forced-choice interventions moved it by at most $+0.010$. This is a redistribution: overall accuracy falls by 4.7 points, nearly all in the 18,301 evidence-block

rows (-4.9 against -3.2 points). The recipe helps where the plain model is weak (IGO multilingual, 41.1% to 47.0%) and hurts where it knows the mapping (composed bridges, 97.2% to 84.7%; Appendix Table 21). On the development split, which played no role in selecting the target cells, the difference is +0.038 $[+0.001, +0.077]$ over cells with at least five rows; inconclusive over all 720 (Appendix J).

Frozen-checkpoint source transfer. A prospectively sealed NOL protocol uses 243 support codes and 976 queries (Appendix M). Five labeled references raise gold-only accuracy from 9.1–9.9% to 28.9–29.6% across three seeds. Adapted RL adds +0.10 points $[-0.82, +1.13]$ over its SFT parent with references. A post-hoc nearest-code support lookup reaches 35.8%. This single English source/version shares classification lineage and historical titles; it does not establish cross-institution tail generalization.

7 Related Work

RLVR and sampled feedback. GRPO and subsequent work use verifiable scores to train reasoning policies (Shao et al., 2024; Yu et al., 2025; Liu et al., 2025); dynamic sampling removes groups whose rewards do not vary. Our estimator remarks specialize these known properties to a deferral action. The contribution is the diagnostic comparison between evidence, action availability and supervision, rather than a new general estimator. The protocol emphasis also relates to AI control (Greenblatt et al., 2024), although our setting is not adversarial.

Selective prediction and supervised refusal. Selective classification trades coverage for accuracy (Chow, 1970; Geifman & El-Yaniv, 2017); refusal tuning supplies targets for when to decline (Zhang et al., 2024; Xu et al., 2024). Our `ABSTAIN`-target control is such a refusal-tuning baseline built from candidate evidence; the gold-only comparison on the original benchmark favors thresholding the adapted model's confidence. Our cell tilt is a Group-DRO-style reweighting (Sagawa et al., 2020), which cannot restore a zero gradient (Remark 1). Selective prediction can magnify group accuracy disparities (Jones et al., 2021); we therefore also report cell coverage.

Institutional mapping. Occupation-coding systems use retrieval and supervised classifiers at a target production rate (Gweon et al., 2017; Schierholz & Schonlau, 2021). We study institution-specific failures across heterogeneous source systems and compare row-only post-training with explicit source evidence and deferral. Official table lookup remains a benchmark-source diagnostic, not an independent deployment baseline. Appendix F provides the wider context.

8 Limitations and Conclusion

One 3B model and one task family; the supervised controls and CellTilt+X have three seeds each, other RL arms one. Bootstraps quantify evaluation, not training, variance, so single-run RL deltas need replication. The 73 target cells are test-derived and informed prompt and sampling design; full-test reporting does not remove that exposure. The frozen NOL transfer covers one source, so broader independent evaluation remains necessary. The 50-row closed-model pilot is a plausibility check, not a benchmark. The selective results live in the augmented-input regime, and the RL policies started from a model not adapted to it; one post-hoc rerun from that model never abstains (Appendix J). Continued training on augmented prompts recovers most of that deficit, and the tail improvement is a same-system result costing overall accuracy. Missing controls: matched-budget continued SFT without evidence, more seeds of that rerun, plain-prompt CellTilt+X, cell routing, mean advantages without cell weights, and one-factor two-sided forcing.

Forced-choice variants, lenient scoring, checkpoint overlap and benchmark-source diagnostics support the negative row-only result; more seeds would tighten the selective-RL effect sizes. The supervised controls change how the matched-input mechanism study reads on this model and task: the gold-only continued-SFT recipe moves accuracy and the tail without RL or abstention targets. Separating extra optimization from training evidence needs the matched-budget control. The stable null remains the main message: verifiable rewards did not make the tested open policies elicit exact source-system mappings from row-only prompts; the tail moves once the prompt exposes that evidence.

ETHICS STATEMENT

Occupational crosswalks affect how labor-market statistics describe workers across countries and institutions. The experiments use official crosswalk records and no personal data. We report cell-level and source-category failures rather than only aggregate accuracy, and treat abstention as deferral: abstentions fall most on survey-codebook and multilingual IGO rows and on low-income systems (coverage 60% to 52%, Appendix Table 9), so deferred rows would return to the least-resourced coding offices. Because ISCO-08 supports international labor-statistics harmonization, unaudited LLM classification could concentrate errors on lower-resourced systems and corrupt the comparisons that crosswalks are meant to support.

REPRODUCIBILITY STATEMENT

Analysis scripts regenerate tables and plot data from saved predictions, training logs and gold labels, with fixed bootstrap seeds and recorded inputs. The retrieval probes of Section 3.3 and the closed-model pilot come from the audit scripts named in Appendix B. Training configurations, prompts, reward definitions, and the advantage patch are specified in Appendices B–D; compute and cost are reported in Appendix H. Code, configurations, and prediction files will be released with the benchmark artifact.

AI USE STATEMENT

Generative AI assisted with manuscript editing. Saved predictions, input and checkpoint fingerprints, analysis scripts, and citation audit records support verification of this work. The authors take responsibility for the final text, claims, and artifacts.

REFERENCES

Anonymous. CrosswalkBench-Global: Official occupational crosswalks for institutional reliability evaluation, 2026. Anonymized benchmark artifact used for double-blind review.

Sapana Chaudhary, Ujwal Dinesha, Dileep Kalathil, and Srinivas Shakkottai. Risk-averse fine-tuning of large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. URL http://arxiv.org/abs/2501.06911v1. arXiv:2501.06911.

Xujun Che, Yuchen Yuan, Weida Zhao, and Chenyang Yu. Abstention as an action can kill both the reward gradient and the KL anchor: Collapse law and repair for error-penalized reinforcement learning, 2026. URL https://arxiv.org/abs/2608.00301.

Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. Can AI assistants know what they don't know? In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL https://proceedings.mlr.press/v235/cheng24i.html.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/tit.1970.1054406.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a CVaR optimization approach. In *Advances in Neural Information Processing Systems 28 (NeurIPS)*, pp. 1522–1530, 2015. doi: 10.48550/arxiv.1506.02188.

Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, and Yi Ma. SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL https://arxiv.org/abs/2501.17161.

Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. Learning with rejection. In *Algorithmic Learning Theory (ALT)*, pp. 67–82, 2016. doi: 10.1007/978-3-319-46379-7_5.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty, 2025. URL https://arxiv.org/abs/2507.16806.

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2501.12948.

John C. Duchi and Hongseok Namkoong. Learning models with uniform performance via distributionally robust optimization. *The Annals of Statistics*, 49(3):1378–1406, 2021. doi: 10.1214/20-aos2004.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Ran El-Yaniv and Yair Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11:1605–1641, 2010. URL https://jmlr.org/papers/v11/el-yaniv10a.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017. URL https://arxiv.org/abs/1705.08500.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Ido Greenberg, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Efficient risk-averse reinforcement learning. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022. doi: 10.52202/068431-2365.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 16295–16336. PMLR, 2024. URL https://proceedings.mlr.press/v235/greenblatt24a.html.

Hyukjun Gweon, Matthias Schonlau, Lars Kaczmirek, Michael Blohm, and Stefan Steiner. Three methods for occupation coding based on statistical learning. *Journal of Official Statistics*, 33(1): 101–122, 2017. doi: 10.1515/jos-2017-0006.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*, 2017. URL https://arxiv.org/abs/1610.02136.

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, 2021. URL https://arxiv.org/abs/2106.09685.

International Labour Office. *International Standard Classification of Occupations: ISCO-08. Volume I: Structure, Group Definitions and Correspondence Tables*. International Labour Organization, Geneva, 2012. URL https://researchrepository.ilo.org/esploro/outputs/book/International-Standard-Classification-of-Occupations-ISCO-08/995673746502676.

Heinrich Jiang, Been Kim, Melody Y. Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018. doi: 10.48550/arxiv.1805.11783.

Erik Jones, Shiori Sagawa, Pang Wei Koh, Ananya Kumar, and Percy Liang. Selective classification can magnify disparities across groups. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2010.14134.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems, 2025. URL https://arxiv.org/abs/2411.06037.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Why language models hallucinate, 2025. URL https://arxiv.org/abs/2509.04664.

Amita Kamath, Robin Jia, and Percy Liang. Selective question answering under domain shift. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5684–5696, 2020. doi: 10.18653/v1/2020.acl-main.503.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. AbstentionBench: Reasoning LLMs fail on unanswerable questions, 2025. URL https://arxiv.org/abs/2506.09038.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR 2019 Workshop on Deep RL Meets Structured Prediction*, 2019. URL https://wouterkool.github.io/publication/buy-4-samples-free-baseline/.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2302.09664.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tülu 3: Pushing frontiers in open language model post-training, 2024. URL https://arxiv.org/abs/2411.15124.

Nicolas Le Roux, Marc G. Bellemare, Jonathan Lebensold, Arnaud Bergeron, Joshua Greaves, Alex Fréchette, Carolyne Pelletier, Eric Thibodeau-Laufer, Sándor Toth, and Sam Work. Tapered off-policy REINFORCE: Stable and efficient reinforcement learning for LLMs, 2025. URL https://arxiv.org/abs/2503.14286.

Martin le Vrang, Agis Papantoniou, Erika Pauwels, Pieter Fannes, Dominique Vandensteen, and Johan De Smedt. ESCO: Boosting job matching in Europe with semantic interoperability. *Computer*, 47(10):57–64, 2014. doi: 10.1109/mc.2014.283.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://arxiv.org/abs/2007.01162.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. URL https://arxiv.org/abs/2205.14334.

Evan Z. Liu, Behzad Haghgoo, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 6781–6792, 2021.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 100–114, 2022. doi: 10.18653/v1/2022.deelio-1.10.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-Zero-Like training: A critical perspective, 2025. URL https://arxiv.org/abs/2503.20783.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7052–7063, 2021. doi: 10.18653/v1/2021.emnlp-main.565.

Bolei Ma, Cynthia A. Huang, and Anna-Carolina Haensch. Can large language models advance crosswalks? The case of Danish occupation codes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 392–399, Albuquerque, USA, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-srw.38. URL https://aclanthology.org/2025.naacl-srw.38/.

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 *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9802–9822, 2023. doi: 10.18653/v1/2023.acl-long.546.

Natascha Massing, Martina Wasmer, Christof Wolf, and Cornelia Züll. How standardized is occupational coding? A comparison of results from different coding agencies in Germany. *Journal of Official Statistics*, 35(1):167–187, 2019.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 2791–2809, 2022. doi: 10.18653/v1/2022.naacl-main.201.

Christopher Mohri and Tatsunori Hashimoto. Language models with conformal factuality guarantees. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 36029–36047, 2024. doi: 10.48550/arxiv.2402.10978.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 7076–7087, 2020. doi: 10.48550/arxiv.2006.01862.

Yonatan Oren, Shiori Sagawa, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust language modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pp. 4227–4237, 2019. doi: 10.18653/v1/d19-1432.

Muyu Pan, Shu Zhao, Nan Zhang, Philip Shin, Varun Parekh, Vijaykrishnan Narayanan, and Rui Zhang. TIAR: Trajectory-informed advantage reweighting for LLM abstention learning, 2026. URL https://arxiv.org/abs/2605.25850.

Qwen Team. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023. URL https://arxiv.org/abs/2305.18290.

Shyam Sundhar Ramesh, Yifan Hu, Iason Chaimalas, Viraj Mehta, Pier Giuseppe Sessa, Haitham Bou Ammar, and Ilija Bogunovic. Group robust preference optimization in reward-free RLHF. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. doi: 10.52202/079017-1171.

Noam Razin, Sadhika Malladi, Adithya Bhaskar, Danqi Chen, Sanjeev Arora, and Boris Hanin. Unintentional unalignment: Likelihood displacement in direct preference optimization. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2410.08847.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models, 2023. URL https://arxiv.org/abs/2312.09300.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000. doi: 10.21314/jor.2000.038.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 2655–2671, 2022. doi: 10.18653/v1/2022.naacl-main.191.

Daniel E. Russ, Kwan-Yuet Ho, Joanne S. Colt, Karla R. Armenti, Dalsu Baris, Wong-Ho Chow, Faith Davis, Alison Johnson, Mark P. Purdue, Margaret R. Karagas, et al. Computer-based coding of free-text job descriptions to efficiently identify occupations in epidemiological studies. *Occupational and Environmental Medicine*, 73(6):417–424, 2016. doi: 10.1136/oemed-2015-103152.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1911.08731.

Malte Schierholz and Matthias Schonlau. Machine learning for occupation coding—A comparison study. *Journal of Survey Statistics and Methodology*, 9(5):1013–1034, 2021. doi: 10.1093/jssam/smaa023.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms, 2017. URL https://arxiv.org/abs/1707.06347.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Carlos N. Silla and Alex A. Freitas. A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, 22(1–2):31–72, 2011. doi: 10.1007/s10618-010-0175-9.

Tasuku Soma and Yuichi Yoshida. Statistical learning with conditional value at risk, 2020. URL https://arxiv.org/abs/2002.05826.

Linxin Song, Taiwei Shi, and Jieyu Zhao. The hallucination tax of reinforcement finetuning, 2025. URL https://arxiv.org/abs/2505.13988.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. LACIE: Listener-aware finetuning for calibration in large language models. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024. doi: 10.52202/079017-1364.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the CVaR via sampling. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2993–2999, 2015. doi: 10.1609/aaai.v29i1.9561.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. TruthRL: Incentivizing truthful LLMs via reinforcement learning, 2025. URL https://arxiv.org/abs/2509.25760.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations (ICLR)*, 2020. URL https://arxiv.org/abs/1908.04319.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision (ECCV)*, pp. 148–166, 2022. doi: 10.1007/978-3-031-20059-5_9.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992. doi: 10.1007/bf00992696.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2305.13300.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback. In *First Conference on Language Modeling (COLM)*, 2024. doi: 10.48550/arxiv.2403.18349.

Jianhao Yan, Yafu Li, Zican Hu, Zhi Wang, Ganqu Cui, Xiaoye Qu, Yu Cheng, and Yue Zhang. Learning to reason under off-policy guidance, 2025. URL https://arxiv.org/abs/2504.14945.

Hiyori Yoshikawa and Naoaki Okazaki. Selective-LAMA: Selective prediction for confidence-aware evaluation of language models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 2017–2028, 2023. doi: 10.18653/v1/2023.findings-eacl.150.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, et al. DAPO: An open-source LLM reinforcement learning system at scale, 2025. URL https://arxiv.org/abs/2503.14476.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Yang Yue, Shiji Song, and Gao Huang. Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model?, 2025. URL https://arxiv.org/abs/2504.13837.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi 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 *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, 2024. doi: 10.18653/v1/2024.naacl-long.394.

Xinyu Zhu, Mengzhou Xia, Zhepei Wei, Wei-Lin Chen, Danqi Chen, and Yu Meng. The surprising effectiveness of negative reinforcement in LLM reasoning. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2025. doi: 10.52202/085713-4220. URL https://arxiv.org/abs/2506.01347.

A PROOFS

GRPO-style estimators use a group of G rewards with mean $\bar{r}$ and population standard deviation σ. The standard advantage is $a_i = (r_i - \bar{r})/(\sigma + \epsilon)$.

Remark 1 (Zero-variance groups carry no signal)**.** *If $r_1 = \cdots = r_G$, then the mean advantage, the within-group CVaR advantage of Section 4, and any estimator considered here that is a function of $(r_i - \bar{r})_i$ or of the empirical group quantiles and vanishes at zero centred rewards is identically zero, so the group contributes no policy gradient. Cell-level reweighting (Sagawa et al., 2020) rescales this gradient and therefore leaves it zero, and group-filtering schemes that drop zero-variance groups (Yu et al., 2025) drop the same prompts from the batch. This is the property behind dynamic sampling and the leave-one-out baseline (Kool et al., 2019).*

Our Cell-conditional-value-at-risk (Cell-CVaR) estimator follows the classical construction (Rockafellar & Uryasev, 2000; Tamar et al., 2015; Soma & Yoshida, 2020): with q_α the empirical α-quantile of the group rewards, the tail advantage is $a_i^{\text{tail}} = (r_i - q_\alpha)/\alpha$ for $r_i \le q_\alpha$ and 0 otherwise,

mixed as $a_i = (1-\lambda)(r_i - \bar{r}) + \lambda\, a_i^{\text{tail}}$ with $\lambda \in [0,1]$ controlling the mean–tail mixture, and normalized by σ.

Remark 2 (Full-weight CVaR advantages are non-positive)**.** *For $\lambda = 1$, $a_i \le 0$ for every i, with equality for every completion at or above q_α. A completion above the group's α-quantile receives no gradient, and one below it is only ever pushed down: the "blindness to success" of CVaR policy gradients (Greenberg et al., 2022), here in its sharpest form because the quantile is taken within the group. With discrete rewards the advantage also vanishes on many groups that have non-zero variance: taking q_α as the order statistic $r_{(\lfloor (G-1)\alpha \rfloor + 1)}$ (our implementation; $r_{(2)}$ for $G = 8$, $\alpha = 0.25$), every a_i is zero whenever the two lowest rewards are tied, for example whenever two or more completions are wrong.*

Derivation of Remark 1. If $r_1 = \cdots = r_G = r$, then $\bar{r} = r$, $\sigma = 0$, every empirical quantile equals r, and $r_i - \bar{r} = 0$ for all i. The mean advantage is $0/(0+\epsilon) = 0$; the CVaR tail advantage is $(r-r)/\alpha = 0$ and its mixture with the mean advantage is 0; any estimator that is a function of $(r_i - \bar{r})_i$ or of the empirical quantiles takes the value it takes at the all-zero argument, which is 0 for every estimator considered here (mean, CVaR, entropic, and clipped variants all vanish at zero centered rewards). The policy-gradient contribution of the group is $\sum_i a_i \nabla \log \pi(o_i \mid x) = 0$. Multiplying by a cell weight $w(c)$ preserves zero. Dynamic sampling (Yu et al., 2025) removes groups with $\sigma = 0$ from the batch, which removes exactly these prompts. □

Derivation of Remark 2. With $\lambda = 1$, $a_i = a_i^{\text{tail}}/(\sigma+\epsilon)$. If $r_i \le q_\alpha$ then $a_i^{\text{tail}} = (r_i - q_\alpha)/\alpha \le 0$; if $r_i > q_\alpha$ then $a_i^{\text{tail}} = 0$. Equality $a_i = 0$ holds iff $r_i \ge q_\alpha$. Hence no completion receives a positive advantage, and a completion strictly below the quantile receives a strictly negative one. For $G = 8$ and $\alpha = 0.25$, $q_\alpha = r_{(2)}$. If $r_{(1)} = r_{(2)}$, every $r_i \le q_\alpha$ equals q_α, so $a_i^{\text{tail}} = 0$ for all i even when $\sigma > 0$; a non-zero advantage requires a unique lowest reward. The share of zero-variance groups is therefore a lower bound on the share of groups that contribute no gradient under this estimator. □

Remark 3 (Sign of a forced rollout's advantage)**.** *Replace one of the G rollouts by a fixed completion with reward a (here `ABSTAIN`, $a = 0$) and use mean advantages. The forced rollout's advantage is positive if and only if the mean reward of the $G-1$ sampled rollouts is below a. For $a = 0$, this holds when wrong sampled answers outnumber correct ones; sampled abstentions contribute zero. If all sampled rollouts are code answers, this is the empirical counterpart of the Chow threshold (Chow, 1970), and in an all-wrong group its normalized advantage equals $\sqrt{G-1}$, the largest value a group of size G can assign.*

Derivation of Remark 3. Let $r_1, \ldots, r_{G-1}$ be the sampled rewards with mean m, and $r_G = a$ the forced reward. The group mean is $\bar{r} = \big((G-1)m + a\big)/G$, so $r_G - \bar{r} = a - \bar{r} = \frac{G-1}{G}(a-m)$, which is positive iff $m < a$; the normalization by $\sigma + \epsilon > 0$ preserves the sign. Under the selective reward with $a = 0$, code answers score ± 1 and abstentions score zero, so $m < 0$ iff wrong answers outnumber correct ones. If all $G-1$ sampled answers score -1 and the forced reward is 0, the rewards are $G-1$ copies of -1 and one 0, whose population standard deviation is $\sigma = \sqrt{(G-1)}/G$ and whose forced-rollout deviation is $(G-1)/G$; the normalized advantage is $\sqrt{G-1}$ (up to ϵ). Advantages sum to zero, so each sampled wrong answer receives $-1/\sqrt{G-1}$ (about -0.38 for $G = 8$) before the cell weight: the forced rollout breaks the zero-variance condition of Remark 1 and the sampled wrong codes are pushed down at the same time. For any group of size G with population-normalized advantages, $\max_i a_i \le \sqrt{G-1}$, attained exactly when one reward exceeds $G-1$ equal others. □

Remark 4 (The negative forced term is unbounded below)**.** *For a fixed negative advantage $a_G < 0$, an unclipped forced-completion term $-a_G \log \pi_\theta(\texttt{ABSTAIN} \mid x)$ is unbounded below as the completion's probability tends to zero. Repeatedly injecting the completion prevents its contribution from becoming rare through on-policy sampling. This identifies an unstable direction of the term, not a guarantee that the full optimizer will follow it. In particular, it does not imply divergence of the exact forward KL to a full-support reference. The one-sided clamp removes this negative term, as motivated by work on unlikelihood, likelihood displacement and tapered off-policy guidance (Williams, 1992; Welleck et al., 2020; Razin et al., 2025; Le Roux et al., 2025).*

Derivation of Remark 4. Write the next-token distribution at the first completion position as a softmax over logits z, with $\pi_A = \pi(\texttt{ABSTAIN} \mid x)$ the probability of the abstain token. Then $\partial \log \pi_A / \partial z_A = 1 - \pi_A$. The REINFORCE term contributes $-a_G(1-\pi_A) = |a_G|(1-\pi_A) > 0$ to the loss gradient on z_A, while its value is $|a_G| \log \pi_A \to -\infty$. Other token positions contribute analogously. This component therefore keeps pushing the forced token down as its probability vanishes; opposing gradients from the rest of the loss may change the full update. For a fixed finite vocabulary and reference q with $q_{\min} > 0$, the exact forward KL satisfies $D_{\mathrm{KL}}(p\|q) \le \log(1/q_{\min})$, so $p_A \to 0$ does not itself make it infinite. The logged KL statistic exceeding 10^{12} in the failed run is an implementation-level sampled-penalty observation, not a measurement of that exact distributional divergence. □

A.1 The forced rollout as weighted cross-entropy

Cell-weighted mean advantages. Within a group we use the mean advantage, so that above-group completions can be promoted, and apply risk sensitivity across cells by scaling each group:

$$A_i = w(c(x))\,\frac{r_i - \bar{r}}{\sigma + \epsilon}, \qquad w(c) = \frac{1}{\bar{w}}\,\mathrm{clip}\Big(e^{-(\bar{R}_c - \bar{R})/\tau},\, 0.25,\, 4\Big), \tag{2}$$

where $\bar{R}_c$ is an exponential moving average (decay 0.98) of the mean reward of cell c, $\bar{R}$ the population average, $\tau = 0.5$, and $\bar{w}$ the batch mean of the raw weights. This is a cell-level exponential tilt of the objective (Li et al., 2021): unlike Group-DRO's exponentiated-gradient weights (Sagawa et al., 2020; Oren et al., 2019) it interpolates smoothly between average and worst-cell risk, and the clipping and mean-one normalization keep the effective learning rate fixed while amplifying cells currently below average.

The sampled rollouts use the clipped surrogate; the injected completion uses an unclipped REINFORCE term with importance ratio one and its negative advantage clamped to zero. One step minimizes

$$\begin{aligned}\mathcal{L}(\theta) = -\frac{1}{G}\Big[\sum_{i<G} \min\big(\rho_i A_i,\ \mathrm{clip}(\rho_i, 1\pm\varepsilon)A_i\big) + \max(A_G, 0)\,\log \pi_\theta(\texttt{ABSTAIN} \mid x)\Big] \\ + \beta\,\widehat{K}\big(\pi_\theta, \pi_{\mathrm{SFT}}\big),\end{aligned} \tag{3}$$

where ρ_i abbreviates the per-token importance ratios of completion i (the surrogate is evaluated per token and aggregated by the trainer's token-level normalization) and A_i is defined in Equation 2. Here $\widehat{K}$ denotes the implementation's sampled KL penalty; it is distinct from evaluating the exact forward KL over all possible completions, especially when the batch contains forced completions.

Remark 5 (Forced term)**.** *With ratio one and the clamp, the forced rollout's contribution to the gradient of Equation 3 is* $-\frac{1}{G}\max(A_G, 0)\,\nabla_\theta \log \pi_\theta(\texttt{ABSTAIN} \mid x)$*, i.e., the gradient of a cross-entropy loss on the target* `ABSTAIN` *with weight* $\frac{w(c)}{G}\max\big(0, \frac{G-1}{G}(a-m)/(\sigma+\epsilon)\big)$*, where* m *is the mean reward of the* $G-1$ *sampled rollouts.*

Derivation. Substitute $r_G - \bar{r} = \frac{G-1}{G}(a-m)$ from the derivation of Remark 3 into Equation 2; the REINFORCE term with ratio one is $A_G \log \pi_\theta(\texttt{ABSTAIN} \mid x)$ and the clamp keeps only $A_G > 0$. □

The weight is therefore an online refusal label whose strength grows with how wrong the sampled group is. Its sign is decided by $G - 1 = 7$ samples at the sampling temperature. In the diagnostic case with no sampled abstentions and true code-answer correctness probability p, it is positive with probability $\Pr[\mathrm{Bin}(7, p) \le 3]$ (Table 4). The one-sided clamp turns non-positive estimates off; it does not push the forced action down. With sampled abstentions, the exact condition remains that wrong answers outnumber correct ones.

B Experimental Details

Forced-choice ladder (Table 5). The shaped reward is $+1$ for any code in $Y(x)$, $+0.2$ same major group, -0.2 wrong major group, -1 invalid. GRPO runs use KL coefficient $\beta = 0.04$ to the SFT reference and a maximum completion length of 32 tokens. Cell reweighting uses EMA decay 0.98, temperature 0.5, and clip range $[0.25, 4]$.

Table 4: Probability that the forced ABSTAIN rollout receives a positive advantage ($a = 0$, $G = 8$) in the diagnostic case with no sampled abstentions, as a function of code-answer correctness probability p.

p	0.2	0.3	0.4	0.5	0.6	0.7	0.8
$\Pr[A_G > 0]$	0.97	0.87	0.71	0.50	0.29	0.13	0.03

Algorithm 1 CELLTILT+X: one optimizer step

for each prompt x in the batch, with cell $c = c(x)$ **do**
  sample $G - 1$ completions from $\pi_\theta(\cdot \mid x)$; append the forced completion ABSTAIN⟨eos⟩
  score all G completions with the selective verifier: $+1$ if $\hat{y} \in Y(x)$, -1 wrong or invalid, a for ABSTAIN
  $a_i \leftarrow (r_i - \bar{r})/(\sigma + \epsilon)$ ▷ mean advantage; zero if $\sigma = 0$
  $A_i \leftarrow w(c)\, a_i$ with $w(c) = \mathrm{clip}\big(e^{-(\bar{R}_c - \bar{R})/\tau}, 0.25, 4\big)/\bar{w}$ ▷ clipped, mean-one cell tilt weight
  $A_G \leftarrow \max(A_G, 0)$ ▷ one-sided: forced rollout only ever raises $\pi(\texttt{ABSTAIN} \mid x)$
  update $\bar{R}_c$ with the group's mean reward
end for
take a clipped-surrogate step (Schulman et al., 2017) on the sampled rollouts and a REINFORCE step on the forced rollouts, with KL penalty β to the SFT reference

Selective RL runs. Maximum completion length 16 tokens; KL $\beta = 0.04$; linear learning-rate decay; checkpoints every 300 steps. Training rows from the 94 RL-excluded systems are dropped; national-crosswalk and IGO rows are oversampled $3\times$. The Cell-CVaR run used steps 0–600 on an RTX 4090 and 600–3,000 on an H100 with an identical effective batch (only the log of its final restart, from step 1,200, was retained); CELLTILT+X ran on a single RTX 4090. Predictions are greedy with 16 new tokens. The two-sided forcing variant is identical to CELLTILT+X except that the forced rollout's negative advantages are kept and the within-group estimator is the CVaR estimator of Section 4; it was stopped at step 340.

Output parsing. Strict accuracy compares the historical parser's normalized prediction to the primary gold code; it is not exact-format accuracy. Noncanonical responses receive correct credit on 45 test rows for CELLTILT+X (0.225 percentage points) and one row for augmented SFT (0.005 points). These counts identify a scoring sensitivity; they do not constitute a new exact-format evaluation.

C CONTEXT HELPS THE TARGET TAIL BUT SHIFTS THE FULL DISTRIBUTION

Injecting the same-system neighbour codes and correspondence candidates into the prompt without training helps on the target cells and hurts the easier full-split distribution. On the target cells the base 3B model rises from 0.5% to 13.0% and SFT from 12.0% to 15.8% $[14.6, 17.0]$. But the unadapted SFT model applied to the augmented prompt (forced-choice, no ABSTAIN option) on the full test split falls from 0.790 to 0.450 strict Acc@1 (worst-cell CVaR 0.221 to 0.016), because on the ISCO-88-labelled rows it maps at 94% from memory and a candidate list turns that into 35%. Closed model A likewise drops from 62% to 56% on the pilot rows. Label-free gates recover about one point. This is the familiar tension between parametric and contextual knowledge (Longpre et al., 2021; Mallen et al., 2023; Xie et al., 2024). The boundary it exposes is the one the SFT-plus-variants ladder could not cross; in this augmented regime, the policy must learn *when* to use context and, we argue next, when to decline.

D PROMPT TEMPLATE

The augmented prompt lists up to five same-system training rows with the numerically nearest national codes (prefix-3 siblings preferred), each with its ISCO-08 code; for rows whose national code is a coarse ISCO-88 group, the candidate ISCO-08 codes from the official correspondence with their titles; the source row (country/entity, system, national code, national title, source category, primary language); and the instruction "Answer with the four-digit ISCO-08 code only, no explanation. If you are not confident that one code is correct, output exactly ABSTAIN instead of a code." The plain prompt omits the neighbour and candidate blocks and the abstention sentence. Rows without same-system training neighbours receive the plain prompt plus the abstention sentence, so the ac-

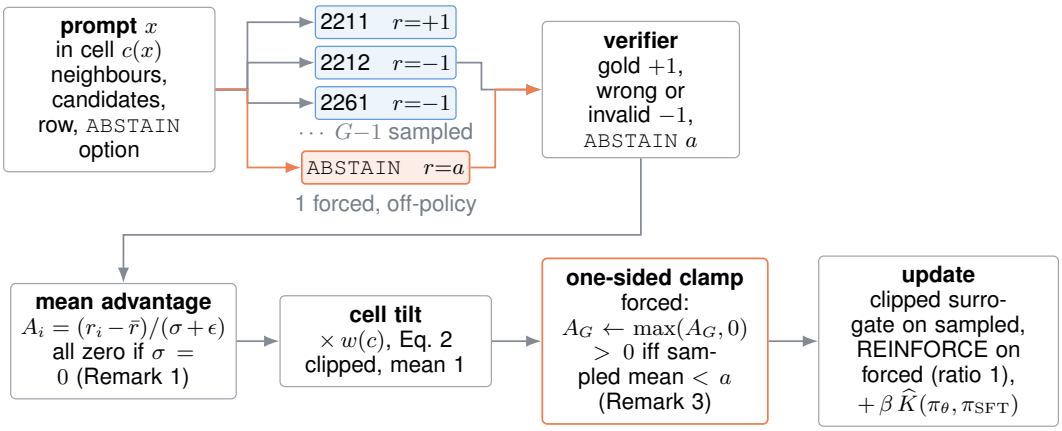


Figure 3: One optimizer step of CELLTILT+X for one prompt group. With all sampled answers wrong, injection gives the forced ABSTAIN advantage $+\sqrt{G-1}$ before cell weighting and ignoring ϵ; sampled wrong answers then have negative advantages. In a mostly-right group, a sampled ABSTAIN is pushed down on-policy while the forced update is clamped to zero.

Table 5: Training configurations. All runs use LoRA rank 16 (scale 32, dropout 0.05) on all attention and MLP projections of Qwen2.5-3B-Instruct (29.9M trainable parameters, 0.97% of the model), and bf16. Each configuration has one historical run; initialization-seed provenance is incomplete. Retrieval probes: `code_locality_audit.py`, `code_locality_vs_semantic.py`; lookup diagnostic: `retrieval_override.py`; pilot: `closed_model_pilot.py`.

Stage	Init	Data	Steps / epochs	Batch	LR	G / temp.	Reward, risk
SFT	base	50k rows	1 epoch	32	2×10^{-5}	–	primary code, CE
DPO (self-hard)	SFT	SFT's wrong preds	1 epoch	32	5×10^{-6}	–	$\beta = 0.1$
GRPO mean	SFT	50k rows	1 epoch	64	10^{-6}	8 / 0.9	shaped, mean
GRPO Cell-CVaR v1	SFT	50k rows	1 epoch	64	10^{-6}	8 / 0.9	shaped, CVaR α=0.25
GRPO Cell-CVaR v3	SFT	50k, 3× weak cats.	1 epoch	80	10^{-6}	10 / 1.15	shaped, CVaR + cell weights
Tail-cell SFT	SFT	clean rows of 73 cells	2 epochs	16	5×10^{-6}	–	primary code, CE
Cell-CVaR selective	SFT	44.5k aug. rows	3,000 steps	64	10^{-6}	8 / 1.15	selective, CVaR + cell weights
CELLTILT+X	SFT	44.5k aug. rows	3,000 steps	64	10^{-6}	8 / 1.15	selective, mean + cell weights + forced ABSTAIN
No-exploration ablation	SFT	44.5k aug. rows	3,000 steps	64	10^{-6}	8 / 1.15	selective, mean + cell weights
Aug. SFT + ABSTAIN targets	SFT	44.5k aug. rows	1 epoch	32	5×10^{-6}	–	candidate-miss targets, CE; seeds 17/29/43
Aug. SFT + gold-only targets	SFT	44.5k aug. rows	1 epoch	32	5×10^{-6}	–	gold-code targets, CE; seeds 17/29/43

tion space is identical for every row. One augmented prompt verbatim (gold 2211; CELLTILT+X answers 2211, the SFT + instruction model 2212):

```
Reference records from the same source system (national code | national title
-> ISCO-08):
- 102 | Spécialistes des sciences de la vie non classés ailleurs -> 2133
- 104 | Médecin spécialiste (chirurgie, gynécologie, pédiatre, ophtalmologue,
orthopédiste, etc.)  -> 2212
- 105 | Dentiste -> 2261
- 109 | Infirmier diplômé -> 2221
- 116 | Professeur d'enseignement secondaire -> 2330
Candidate ISCO-08 unit groups in the same region as the references (code:
title):
- 2131:  Biologists, Botanists, Zoologists and Related Professionals
- 2132:  Farming, Forestry and Fisheries Advisers
- 2133:  Environmental Protection Professionals
- 2211:  Generalist Medical Practitioners
- 2212:  Specialist Medical Practitioners
- 2221:  Nursing Professionals  [...10 more candidates ...]
- 2330:  Secondary Education Teachers
Map the occupational record to one valid four-digit ISCO-08 unit group.
Return only the four-digit ISCO-08 code.
Country or entity:  Togo
Source system:  EHCVM 2018-2019 AFRISTAT profession/metier via UEMOA ISCO-08
National code:  103
National title:  Médecin généraliste
Source category:  igo_multilingual
Primary source language:  French
Answer with the four-digit ISCO-08 code only, no explanation.  If you are not
confident that one code is correct, output exactly ABSTAIN instead of a code.
```

E TRAINING DYNAMICS

Figure 4 tracks the quantities the remarks single out. CELLTILT+X's zero-variance-group fraction averages 0.30 over the run (0.09–0.46; 0.30 on the steps after 1,200 where both logs exist), against 0.78 (0.64–0.89) for the CVaR policy. Early in training the forced `ABSTAIN` is preferred in 40–49% of groups. As the policy learns to abstain on its own (on-policy abstention rate at sampling temperature 16%→42%), that fraction falls to 10–18% and the mean reward rises from -0.02 to $+0.32$, against $+0.19$ for the CVaR policy at the same step. The logged KL statistic reaches 0.19–0.29, an order of magnitude above the CVaR run.

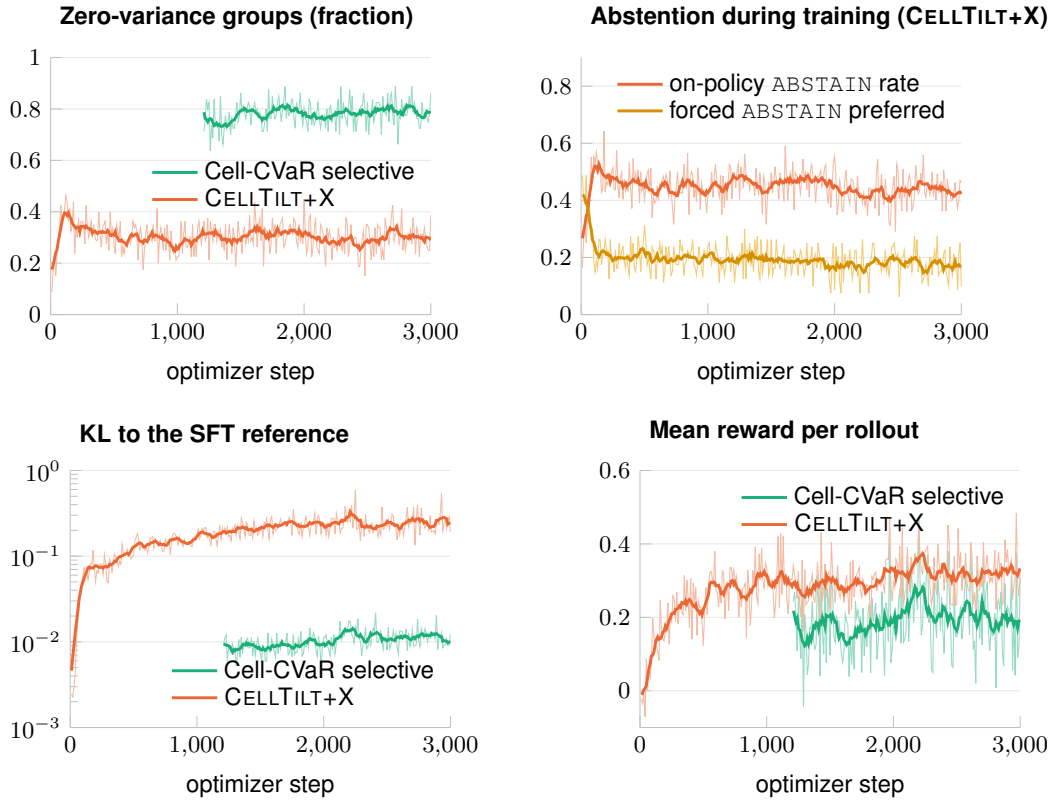


Figure 4: Training dynamics. Faint lines are the per-10-step logs, solid lines 9-point moving averages. CELLTILT+X keeps the zero-variance-group fraction near 0.3 (the within-group CVaR run stays at 0.64–0.89), its on-policy abstention rate rises from 16% to over 40% as the fraction of groups in which the forced `ABSTAIN` is preferred falls, and it moves an order of magnitude further from the reference while its reward ends above the CVaR run's. The Cell-CVaR curves begin at step 1,200 because only the log of that run's final restart was retained.

F EXTENDED RELATED WORK

Context sufficiency and abstention. Joren et al. (2025) distinguish insufficient retrieved context from failures to use sufficient context, and study selective generation informed by that distinction. Our evidence-presence categories do not certify sufficiency: code-neighbour references can omit or contradict the desired mapping. The question here is how prompt evidence and post-training redistribute errors across institutional cells, and whether selective gains retain those cells. Likewise, Kirichenko et al. (2025) evaluate abstention across diverse unanswerability conditions; missing reference records in our task do not by themselves establish that a query is unanswerable.

Occupation coding and crosswalks. Occupation coding uses nearest-neighbour and supervised classifiers on titles, deployed at a target production rate (Gweon et al., 2017; Russ et al., 2016;

Schierholz & Schonlau, 2021), and LLM crosswalk construction has been explored with prompting (Ma et al., 2025); ISCO-08 is a classic hierarchical target (Silla & Freitas, 2011). We add the cell audit, the code-space locality prior (a structural analogue of demonstration retrieval (Liu et al., 2022; Rubin et al., 2022; Min et al., 2022)), and the two-regime diagnosis.

Verifiable-reward RL and its blind spots. GRPO (Shao et al., 2024; DeepSeek-AI, 2025) and its variants (Yu et al., 2025; Liu et al., 2025) optimize group-relative advantages.[2] DAPO's dynamic sampling drops zero-variance groups, which are common in our tail, and Yue et al. (2025) find that tested RLVR setups improve low-budget sampling without expanding large-k reasoning coverage. In different arithmetic and navigation tasks, Chu et al. (2025) find stronger out-of-distribution generalization from RL than SFT. These findings concern different notions and settings of generalization. Our claim is confined to exact mappings under the tested prompts and training budget; row-only inputs omit other source-system records. Song et al. (2025) report degraded refusal on unanswerable math after reinforcement finetuning, while Kirichenko et al. (2025) find reduced abstention in tested reasoning models across broader conditions. Kalai et al. (2025) analyze how evaluation incentives can favor guessing over acknowledging uncertainty. Recent work adds abstention or calibration terms to the reward (Xu et al., 2024; Damani et al., 2025; Wei et al., 2025; Stengel-Eskin et al., 2024) or reweights the abstention advantage inside GRPO (Pan et al., 2026). Our point is narrower: when the policy samples neither abstention nor a correct answer, and when within-group CVaR removes positive advantages, the estimator decides whether abstention receives a positive learning signal. The forced rollout is an online, group-conditional form of refusal tuning (Zhang et al., 2024; Xu et al., 2024; Cheng et al., 2024) placed inside the group-relative estimator, and an abstention-action instance of off-policy guidance (Yan et al., 2025); the one-sided treatment of off-policy negatives follows the tapering argument of Le Roux et al. (2025), and Che et al. (2026) analyse the complementary collapse of abstention actions.

Robustness and selective prediction. Group-DRO (Sagawa et al., 2020) and its label-free relaxation (Liu et al., 2021), CVaR-DRO (Duchi & Namkoong, 2021), and CVaR objectives in RL (Rockafellar & Uryasev, 2000; Tamar et al., 2015; Chow et al., 2015; Chaudhary et al., 2024) target worst-group risk. In our abstention setup, applying full-weight CVaR only within sampling groups does not create a positive abstention signal. Selective classification and learning with rejection or deferral (Chow, 1970; El-Yaniv & Wiener, 2010; Geifman & El-Yaniv, 2017; 2019; Cortes et al., 2016; Mozannar & Sontag, 2020), confidence baselines (Hendrycks & Gimpel, 2017; Jiang et al., 2018), and selective prediction for language models (Kamath et al., 2020; Whitehead et al., 2022; Yoshikawa & Okazaki, 2023; Kadavath et al., 2022; Lin et al., 2022; Kuhn et al., 2023; Ren et al., 2023; Mohri & Hashimoto, 2024) motivate the target. We compare these alternatives in a setting with explicit source-system evidence: estimator changes affect RL's use of abstention, but on this development-exposed benchmark gold-only supervised adaptation with confidence thresholding exceeds both the tested RL policies and explicit abstention targets. Conditional cell metrics and the evidence-free national-crosswalk slice show why this does not imply that every institutional failure is repaired (Section 6.4).

G Pre-specified Predictions

Before running CellTilt+X we recorded three predictions against the CVaR policy: (P1) worst-cell selective CVaR@10% rises from ≈ 0 to at least 0.05; (P2) coverage moves toward the structural-rule region and selective accuracy at matched coverage is at least that of SFT plus rules; (P3) the zero-variance-group fraction falls clearly below the CVaR run's. Outcome: P3 holds, P1 holds in sign (0.019, interval $[0.001, 0.058]$) but not at the pre-specified magnitude, and P2 is mixed: CellTilt+X improves matched-input confidence-threshold baselines, while structural rules remain useful post-hoc filters (Table 3). Everything added after these predictions is post-hoc and was not part of P1–P3: the augmented-SFT control with `ABSTAIN` targets, the gold-only control, their three-seed repetitions, the lenient-scoring, evidence-slice and dense-retrieval analyses, the worst-cell comparison between gold-only augmented SFT and plain-prompt SFT, and the controlled CellTilt+X runs, including one started from the gold-only model whose configuration and comparisons were

[2] Not to be confused with group-robust preference optimization (Ramesh et al., 2024), which shares the acronym.

fixed before launch but after the supervised results were known. The seed set, the paired design of the two supervised controls and the matched-coverage comparison were fixed before those runs were scored, but after the test split had been used for method development. None of these results is a pre-registered or fresh-test confirmation; the development-split check in Appendix J evaluates disjoint rows from overlapping systems and does not eliminate adaptive development bias. The dev-defined difficulty, seed-stability and review-budget analyses in Appendix L are also exploratory.

H Compute and Cost

Forced-choice GRPO runs took 10–18 GPU-hours each on a rented RTX 4090. The completed CellTilt+X and no-exploration runs took approximately 21.52 and 17.56 hours, respectively, according to their completion logs. The original augmented SFT control was trained and scored on a laptop (2.10 hours; 13.4 minutes for its ordinary test prediction pass). These are historical wall-clock durations, not estimates for a new device. The controlled runs of Appendix J were measured on one rented RTX 4090: each augmented-SFT run took 79–85 minutes, a 3,000-step CellTilt+X run took 16.4 hours, and an ordinary prediction pass or a confidence pass over the 20k test split took about 11 minutes each, including adapter merging and model loading. Zero-shot probes of base models up to 9B ran in 31 minutes total. Scoring and bootstraps use CPU resources; confidence evaluation also requires model inference. The closed-model pilot cost under $3.

I Additional Tables and Figures

Table 11 breaks the forced-choice ladder down by institutional complexity and income tier; Table 12 and Table 10 give the target-cell diagnostics and the information-source probes underlying Section 3; Table 13 and Table 18 report the closed-model pilot overall and by source category; Table 14 and Table 17 give selective behaviour by source category and by language; Table 16 is the outcome-taxonomy breakdown of Section 6; Table 19 tracks the 300-row check subset used to monitor training; Table 15 gives worked examples; and Figure 5 and Figure 6 are the per-cell and worst-decile-overlap views referenced from the main text.



Figure 5: Where the matched-context policy abstains. Each point is one of the 878 test cells (cells with no answered SFT row omitted), with a small fixed-seed jitter (± 0.012) to separate cells that land on the same rational fraction from having few rows; positions are for legibility only and are not used for any reported statistic. Coverage falls with the SFT model's per-cell accuracy for four source categories, but the national-crosswalk cells (triangles) stay near full coverage regardless of accuracy: the tested policy neither answers them correctly nor reliably detects that it should defer.

Table 6: Ablation on the 20k test split (matched inputs).

Variant	Coverage	Sel. Acc@1	RL-excl. acc.	Worst CVaR@10%	Strict
Within-group CVaR + cell weights (no exploration)	79.1	53.7	56.0	0.002 [0.000, 0.019]	0.425
Augmented SFT + supervised ABSTAIN targets	58.7	84.2	90.9	0.274 [0.174, 0.374]	0.494
Mean advantages + cell weights (no exploration)	72.8	61.4	64.7	0.019 [0.002, 0.059]	0.447
Within-group CVaR + cell weights + two-sided forced ABSTAIN	diverged: on-policy abstention $\to$ 0 by step 170, KL $> 10^{12}$				
Mean + cell weights + one-sided forced ABSTAIN (ours)	70.8	63.3	67.4	0.019 [0.001, 0.058]	0.448

Table 7: Dataset statistics for the balanced experimental splits. "Country/entity labels" uses the same metadata label as Table 1; "Country/entity–systems" counts distinct (country/entity, source-system) pairs; and "Multi" is the fraction of rows with more than one official target code. Ninety-four country/entity–system pairs (2,139 test rows) are additionally excluded from all selective-RL training runs.

Split	Rows	Country/entity labels	Country/entity systems	Cells	Multi
Train (balanced)	50,000	198	946	946	38%
Dev	5,000	157	720	720	34%
Test	20,000	162	878	878	33%

J Supervised Controls across Seeds

Table 22 reports the controlled supervised runs (seeds 17, 29, 43; identical rows, prompts, initialization and optimizer settings, differing only in whether the target may be ABSTAIN). All numbers are on the 20k test split used during method development. Different seeds give different models: 1,137–1,235 of 20,000 predictions differ between gold-only seeds. The development-split check builds augmented prompts for the 5k development rows with the settings that reproduce the test prompts byte for byte, and scores the seed-17 gold-only model against plain-prompt SFT. Over all 720 development cells (median four rows) the worst-cell CVaR difference is -0.005; subsampling the test split to 5,000 rows gives a positive difference in only 81% of draws at that granularity, indicating limited sensitivity. The all-cell result remains inconclusive. Over the 322 development cells with at least five rows the difference is +0.038 $[+0.001, +0.077]$, in line with +0.045 $[+0.016, +0.074]$ when the test split is subsampled to the same size. Mean accuracy on the target cells present in the development split rises from 23.8% to 37.3%, less than on the test split where those cells were selected, and overall accuracy falls by 5.2 points.

Controlled RL runs. Table 23 reports the controlled repetitions of CellTilt+X and one post-hoc run that starts the same recipe from the seed-17 gold-only model (one seed; same test split). The 3 repetitions reproduce the historical run: 63.3–63.4% selective accuracy (mean 63.3%, sample SD 0.06) at 69.9–70.2% coverage with 59–65 fully abstained cells. At those coverages the same-seed gold-only model with a confidence threshold is 29.4–29.8 points higher. Started from the adapted model, the recipe does not abstain under greedy decoding (0 of 20,000 rows), although its sampled abstention rate rose to about 14% during training. Its strict accuracy is 74.5%, 0.34 points above its initialization (row bootstrap [0.14, 0.55]), a difference of the size seen between gold-only seeds. Thresholding its own confidence gives 96.3% at the ABSTAIN-target coverage, against 96.2% for its initialization (difference interval [−0.06, +0.26] points), and at most 0.51 points more at higher coverage. Worst-cell CVaR is 0.295 against 0.301. For this model and recipe, verifier training on the adapted policy adds no selectivity beyond post-hoc confidence.

K Confidence-Threshold Curves

Confidence-threshold baselines. Thresholding sequence log-probability, with each policy's own ABSTAIN outputs ranked least confident and counted as incorrect, gives retained-row AURC 0.326 for SFT + instruction, 0.307 for the Cell-CVaR policy, and 0.256 for CellTilt+X; at 70.6% coverage, SFT's threshold reaches 56.1% selective accuracy, below CellTilt+X's 63.4% in the confidence inference pass. First-token probability preserves the ordering but is worse for these three policies, and CellTilt+X's own $\pi(\texttt{ABSTAIN} \mid x)$ is worse still (AURC 0.355). Curves above a policy's native answering rate are retained-row diagnostics, not attainable answer coverage. The plain-prompt SFT deployment reference has AURC 0.030 (Appendix K, Figure 7).

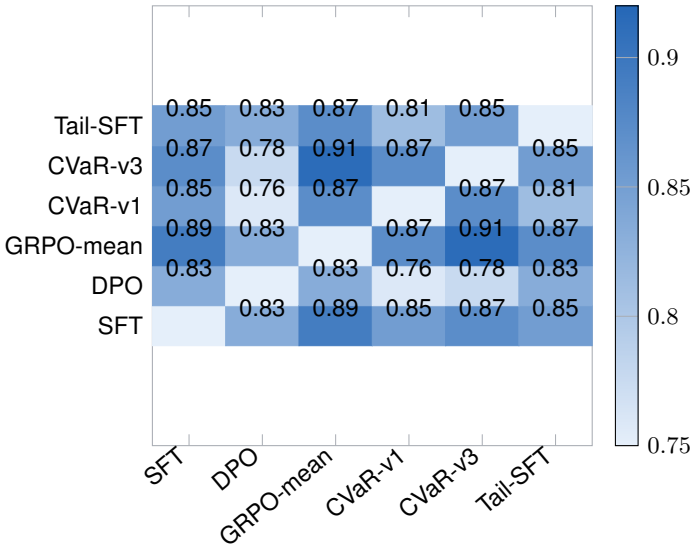


Figure 6: Jaccard overlap between the worst-decile cell sets (87 of 878 cells) of the six forced-choice checkpoints (SFT plus five variants) on the 20k test split. The trivial diagonal (self-overlap $= 1$) is masked; off-diagonal overlap is 0.76–0.91 and the color scale is stretched to that range. 73 cells are in the worst decile of all five early checkpoints (70% of their union) and define the target cells.

Table 8: Worst-cell CVaR@10% by minimum rows (forced choice) or answered rows (selective policies) per cell. Each threshold entry gives CVaR and the number of retained cells; selective policies have different support sets. The final columns count fully-abstained cells and report CVaR on the fixed set of all 878 cells with these cells assigned zero.

Policy	$n \geq 1$	$n \geq 5$	$n \geq 10$	$n \geq 20$	Fully abstained	Fixed-cell CVaR
SFT	0.221 (878)	0.189 (697)	0.141 (488)	0.111 (362)	0	0.221
DPO (self-hard negatives)	0.200 (878)	0.178 (697)	0.136 (488)	0.109 (362)	0	0.200
GRPO, mean advantage	0.224 (878)	0.190 (697)	0.145 (488)	0.116 (362)	0	0.224
GRPO, Cell-CVaR (v1)	0.225 (878)	0.191 (697)	0.138 (488)	0.111 (362)	0	0.225
GRPO, Cell-CVaR (v3, cell-reweighted)	0.229 (878)	0.200 (697)	0.146 (488)	0.118 (362)	0	0.229
Tail-cell SFT	0.231 (878)	0.201 (697)	0.152 (488)	0.120 (362)	0	0.231
SFT + instruction (selective)	0.000 (840)	0.103 (518)	0.095 (449)	0.095 (287)	38	0.000
Cell-CVaR selective	0.002 (842)	0.109 (515)	0.110 (447)	0.099 (281)	36	0.000
Mean + cell weights, no exploration	0.019 (828)	0.129 (505)	0.137 (431)	0.135 (214)	50	0.000
CELLTILT+X (selective)	0.019 (820)	0.127 (503)	0.159 (418)	0.127 (206)	58	0.000
Aug. SFT + ABSTAIN targets	0.274 (595)	0.399 (432)	0.380 (336)	0.216 (161)	283	0.000

Prediction and confidence passes. Operating points come from the ordinary prediction pass and curves from a separate confidence pass over the same checkpoint and prompts. Both use the same adapter merge, prompt formatting, batch size, left padding, greedy decoding and bfloat16 weights, and differ in how the merged model is placed on the device, so they are not bitwise identical. Across nine checkpoint pairs they disagree on 1.0–3.1% of rows. The disagreeing rows are low-confidence (median first-token probability 0.54–0.87, against at least 0.97 on agreeing rows), both passes are mostly wrong on them (14.5–23.7% correct), and strict accuracy differs by at most 0.13 points between passes. Where an operating point and a curve are compared, captions state which pass each comes from.

Table 9: Selective behaviour by World Bank income group of the source system (coverage / selective accuracy, %).

Income group	n	SFT + instruction	Cell-CVaR selective	CELLTILT+X
Low	917	60.4 / 60.6	59.9 / 62.1	52.1 / 72.8
Lower-middle	3240	74.8 / 64.1	74.3 / 67.3	70.8 / 76.6
Upper-middle	3174	81.7 / 65.9	81.3 / 68.3	77.9 / 76.2
High	5990	71.6 / 43.7	70.7 / 45.0	60.7 / 51.9
IGO / multilateral	6679	90.9 / 43.9	90.6 / 47.4	79.0 / 58.5

Table 10: Information-source probes on the 3,716 target-cell rows, using training data only. "Hit" is exact unit-group agreement (or gold contained in the candidate set for $K > 1$); "major" is major-group agreement.

Probe	Rows with a probe	Hit	Major hit / tail gain
Random same-system training row	74.2%	0.9%	17.9%
Nearest same-system code	74.2%	14.6%	55.9%
Code-neighbour candidates, $K = 10$	74.2%	24.5%	63.3%
TF-IDF title candidates, $K = 10$	74.2%	16.1%	47.6%
Agreement-gated official lookup	57 / 73 cells	2,310 rows	+27.3 CVaR points

Table 11: Slice strict accuracy (%) on the 20k test split: SFT absolute, other checkpoints as differences from SFT. ICS is the institutional-complexity tier of the source system and coincides with the source category in these splits (2: IGO multilingual, 3: national crosswalk, 4: survey codebook, 5: composed bridge, 6: structural expansion); tier 1, identity mappings from ISCO-08 to itself, is excluded from all experiments. Income groups follow the World Bank classification (IGO: intergovernmental sources).

Slice	n	SFT	DPO	GRPO-mean	CVaR-v1	CVaR-v3	Tail-SFT
ICS=2	3737	39.0	−2.2	+1.0	+0.9	+1.1	−1.8
ICS=3	1369	40.5	−6.6	+0.1	+0.9	+0.4	−6.1
ICS=4	1509	60.7	0.0	+0.7	−0.1	+0.9	+1.1
ICS=5	7313	97.2	−0.3	+0.1	+0.1	−0.1	−0.1
ICS=6	6072	95.0	−0.2	+0.1	−0.1	−0.3	+0.1
Income=L	917	63.7	−2.8	+1.0	0.0	+0.5	−1.7
Income=LM	3240	84.3	−0.8	+0.2	0.0	+0.2	0.0
Income=UM	3174	89.1	−1.6	+0.2	+0.3	−0.2	−1.7
Income=H	5990	53.9	−1.4	+0.6	+0.6	+0.4	−1.1
Income=IGO	6679	96.2	−0.2	+0.1	+0.1	0.0	0.0

Table 12: Target-cell diagnostics on the 3,716-row tail subset (strict Acc@1 with 95% row-bootstrap intervals, in %). Mistral-7B and Phi-3.5-mini fail to follow the output format on most rows, so their numbers reflect format failure rather than knowledge. "Context" rows inject same-system code neighbours and sibling or correspondence candidates without any training.

Model or input	Strict Acc@1	Major-group Acc@1	Invalid
Qwen2.5-3B-Instruct (base)	0.5 [0.3, 0.8]	8.2	44.3
Qwen2.5-7B-Instruct (base)	0.8 [0.5, 1.0]	14.5	44.9
Qwen3-8B (base)	1.6 [1.2, 2.0]	13.3	64.7
Qwen3.5-9B (base)	2.2 [1.7, 2.6]	29.9	35.8
Phi-3.5-mini (base)	1.3 [1.0, 1.7]	11.4	75.5
Mistral-7B-Instruct (base)	0.1 [0.0, 0.1]	0.3	99.1
SFT-3B, plain prompt	12.0 [11.0, 13.1]	60.1	8.1
Base-3B + 5 code neighbours	9.1 [8.2, 10.0]	39.0	29.7
Base-3B + neighbours + siblings	12.5 [11.4, 13.5]	44.3	18.5
Base-3B + routed context	13.0 [11.9, 14.1]	44.1	20.1
SFT-3B + 5 code neighbours	11.7 [10.7, 12.7]	56.9	12.3
SFT-3B + neighbours + siblings	14.7 [13.6, 15.9]	58.9	6.9
SFT-3B + routed context	15.8 [14.6, 17.0]	59.8	7.4

Table 13: Closed-model pilot on 50 rows sampled uniformly from the target cells (seed 0), scored by the first valid ISCO-08 code in the response; brackets are 95% Wilson intervals. The pilot is diagnostic, not a benchmark comparison; closed models A and B agree on 28 of 50 rows and at least one is right on 33. A thinking-enabled run of closed model C completed only 24 of 50 rows before timeouts and is omitted.

Model	Correct	Strict Acc@1
Closed model A (minimal reasoning)	31/50	62% [48, 74]
Closed model B (low effort)	30/50	60% [46, 72]
Closed model C (thinking off)	19/50	38% [26, 52]
Closed model D (thinking off)	14/50	28% [17, 42]
Closed model A + routed context	28/50	56% [42, 69]
SFT-3B, plain prompt	9/50	18% [10, 31]
SFT-3B + routed context	12/50	24% [14, 37]
Qwen3.5-9B (base)	0/50	0% [0, 7]

Table 14: Per-source selective behaviour on the 20k test split (coverage / selective accuracy, %).

Source category	n	SFT + instruction	Cell-CVaR	CELLTILT+X	Aug. SFT
composed bridge	7313	92.5 / 43.8	92.0 / 48.4	81.4 / 61.3	64.7 / 94.2
structural expansion	6072	71.9 / 73.2	71.6 / 74.8	68.2 / 80.6	58.1 / 97.3
igo multilingual	3737	66.1 / 39.8	64.9 / 41.1	52.2 / 48.9	42.9 / 63.6
survey codebook	1509	63.7 / 43.9	62.8 / 44.8	49.5 / 52.7	33.9 / 76.0
national crosswalk	1369	100.0 / 41.4	99.9 / 40.9	99.8 / 45.5	99.9 / 42.8

Table 15: Qualitative examples (randomly drawn within each pattern). A: SFT answers wrongly, CELLTILT+X abstains. B: SFT abstains, CELLTILT+X answers correctly. C: national crosswalk, both answer and both are wrong. D: survey codebook, SFT answers wrongly, CELLTILT+X abstains.

	Country	System	Code	Title	Gold	SFT	CELLTILT+X
A	Latvia	Profesiju klasifikators (EURES)	3432.10	Teātra DEKORĒTĀJS	3323/3432/3435	3522	ABSTAIN
B	India	NCO 2015	1120.2600	Director, Airport	1120	ABSTAIN	1120
C	UK	SOC 2020	8111/04	Maker, snuff	7516	7317	8111
D	Togo	EHCVM 2022 AFRISTAT second-job pro	470	maçon - tâcheron	7112	9313	ABSTAIN

Table 16: Outcome taxonomy on the 20k test split (% of rows). "Same major" is a valid code in the gold major group but the wrong unit group. The last column is the share of a policy's abstentions that fall on rows the plain-prompt SFT model answers correctly.

Policy	Correct	Same major	Wrong major	Invalid	Abstain	Abst. on plain-correct
SFT, plain prompt	79.0	11.0	7.4	2.6	0.0	–
SFT + instruction	40.7	26.0	7.4	5.6	20.4	67.3
Cell-CVaR selective	42.5	24.4	7.1	5.1	20.9	67.3
Augmented SFT + supervised `ABSTAIN`	49.4	5.3	2.7	1.3	41.3	73.3
CELLTILT+X	44.8	18.0	4.4	3.6	29.2	68.2

Table 17: Selective behaviour by primary source language (eight most frequent; coverage / selective accuracy, %).

Language	n	SFT + instruction	Cell-CVaR selective	CELLTILT+X
German	4402	86.4 / 40.3	85.7 / 42.4	69.1 / 52.9
English	3245	80.8 / 40.7	80.1 / 43.1	76.2 / 49.6
Multilingual	1769	95.8 / 52.1	95.6 / 57.3	90.2 / 70.1
Spanish	1687	77.6 / 65.9	77.1 / 68.4	72.1 / 76.1
French	1342	55.2 / 49.0	53.7 / 51.0	42.2 / 63.3
Arabic	903	86.6 / 62.7	86.7 / 67.6	85.5 / 78.6
Polish	694	82.7 / 46.0	82.4 / 46.3	68.7 / 53.7
Malay	665	89.3 / 85.9	89.3 / 86.2	88.1 / 92.7

Table 18: Closed-model pilot by source category (correct / rows on the same 50 rows).

Source category	n	Closed A	Closed B	Closed C	Closed D	SFT-3B plain
igo multilingual	28	15/28	17/28	12/28	7/28	3/28
survey codebook	12	9/12	9/12	5/12	4/12	4/12
national crosswalk	8	6/8	4/8	2/8	3/8	2/8
structural expansion	2	1/2	0/2	0/2	0/2	0/2

Table 19: Intermediate checkpoints on a fixed 300-row check subset (coverage / selective accuracy, %). The subset is easier than the full split; it was used to monitor training, not to select checkpoints.

Policy	Coverage	Sel. Acc@1
SFT + instruction	81.3	57.0
Cell-CVaR selective, step 300	81.0	57.6
Cell-CVaR selective, step 2,100	81.7	60.4
CELLTILT+X, step 1,500	76.0	69.7

Table 20: Code-space locality by source group on the target-cell rows with a same-system training neighbour (grouping of the audit script, which separates the West-African regional list and ISCO-88-derived systems from their source categories). "Major" is the share of rows whose nearest same-system code shares the gold major group, against a random same-system row; "K" is the share of rows whose gold code is among the top-K candidates by code proximity versus TF-IDF title similarity.

Group	n	Major: nearest / random	K=1: code / title	K=10: code / title
IGO multilingual (EURES bridges)	1732	48.2 / 18.2	13.4 / 8.9	23.1 / 15.9
West-African regional list	684	68.9 / 17.7	18.4 / 7.5	27.6 / 14.2
ISCO-88-derived systems	184	55.4 / 20.7	12.0 / 5.4	15.8 / 12.0
Structural expansion	96	88.5 / 11.5	8.3 / 20.8	41.7 / 36.5
Survey codebook	44	84.1 / 18.2	18.2 / 15.9	22.7 / 20.5
Composed bridge	18	66.7 / 5.6	33.3 / 22.2	38.9 / 22.2
All rows with a neighbour	2758	55.9 / 17.9	14.6 / 8.9	24.5 / 16.1

Table 21: Where gold-only augmented SFT gains and loses against plain-prompt SFT (strict accuracy on the 20k test split, mean of three seeds; both policies answer every row). "Evidence block" marks prompts that carry same-system reference records or ISCO-08 candidates.

Source category	Evidence block	n	Plain-prompt SFT	Gold-only aug. SFT	Change
composed bridge	yes	7,313	97.2%	84.7%	−12.4
igo multilingual	yes	3,519	41.1%	47.0%	+5.9
igo multilingual	no	218	4.6%	3.4%	−1.2
national crosswalk	no	1,369	40.5%	36.8%	−3.7
structural expansion	yes	5,993	95.0%	94.0%	−1.0
structural expansion	no	79	96.2%	95.4%	−0.8
survey codebook	yes	1,476	59.8%	51.1%	−8.7
survey codebook	no	33	100.0%	100.0%	+0.0
All rows	yes	18,301	82.7%	77.8%	−4.9
All rows	no	1,699	39.7%	36.5%	−3.2

Table 22: Controlled supervised runs on the augmented prompt. "Gold-only at same cov." ranks the paired gold-only model's answers by sequence log-probability and cuts at the `ABSTAIN`-target model's native coverage. Worst-cell CVaR@10% is over all 878 cells with fully abstained cells scored as zero.

Policy (seed)	Coverage	Sel. acc.	Correct, all rows	Gold-only at same cov.	Worst-cell CVaR
Gold-only targets (17)	100.0%	74.2%	74.2%	–	0.301
Gold-only targets (29)	100.0%	74.2%	74.2%	–	0.296
Gold-only targets (43)	100.0%	74.4%	74.4%	–	0.292
`ABSTAIN` targets (17)	59.0%	84.0%	49.5%	96.2%	0.000
`ABSTAIN` targets (29)	58.5%	84.3%	49.3%	96.1%	0.000
`ABSTAIN` targets (43)	58.8%	84.1%	49.4%	96.2%	0.000

Table 23: Controlled CELLTILT+X runs (GPU prediction pass, 20k test split). "Gold-only at same cov." cuts the same-seed gold-only model to the run's native coverage by sequence log-probability. The last row starts the recipe from the seed-17 gold-only model (post hoc, one seed).

Policy (seed)	Coverage	Sel. acc.	Sel. acc., lenient	Correct, all rows	Gold-only at same cov.
CELLTILT+X (17)	70.1%	63.3%	68.8%	44.4%	92.9%
CELLTILT+X (29)	70.2%	63.4%	69.0%	44.5%	92.8%
CELLTILT+X (43)	69.9%	63.3%	68.8%	44.2%	93.1%
CELLTILT+X from gold-only SFT (17)	100.0%	74.5%	76.3%	74.5%	74.2%

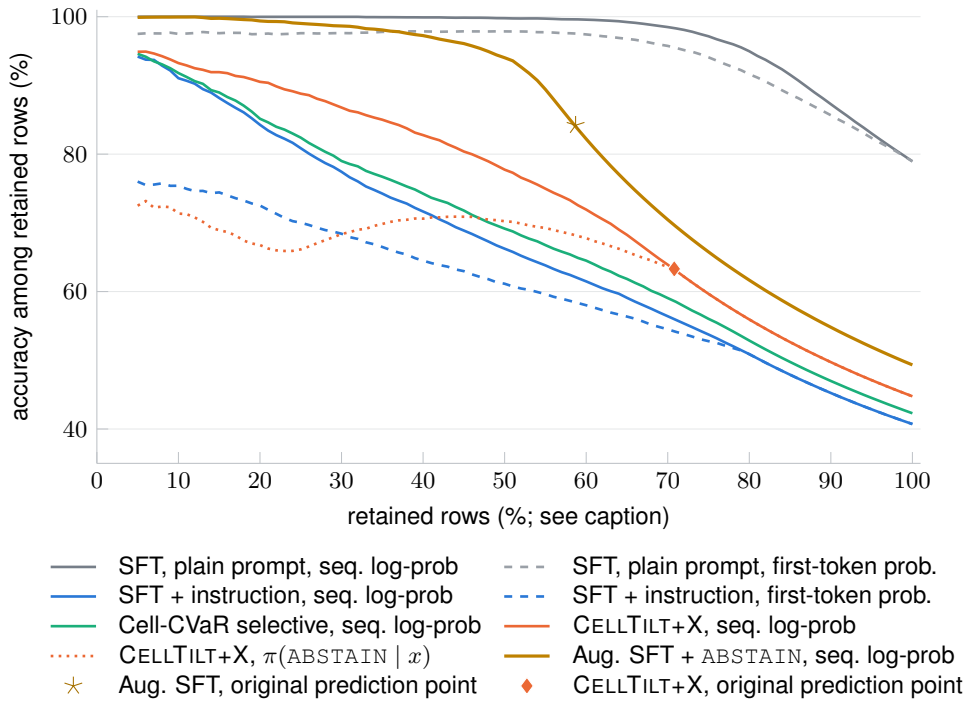


Figure 7: Confidence-ranked retained-row diagnostics on the 20k test split. Native ABSTAIN outputs are ranked last and scored as errors. Up to a confidence pass's own answering rate, the curve is a realizable selective risk–coverage curve; beyond that point, retaining abstentions does not make the model answer them. Figure 2 restricts curves to realizable coverage. Solid lines use sequence log-probability; dashed lines use first-token probability for the two original SFT baselines. Markers come from the original prediction passes, which differ from the confidence passes; curve endpoints and AURC use the latter. Plain-prompt SFT is a separate deployment reference. Full retained-row AURC (lower is better), sequence log-probability / first-token probability: plain-prompt SFT 0.030 / 0.052; SFT + instruction 0.326 / 0.389; Cell-CVaR selective 0.307 / 0.370; CellTilt+X 0.256 / 0.302; CellTilt+X's own $\pi(\texttt{ABSTAIN} \mid x)$ 0.355, worse than either token-probability signal. The abstention-probability sweep is non-monotone at low coverage because $\pi(\texttt{ABSTAIN} \mid x)$ separates "should not answer" from "should answer" but does not rank the answered rows among themselves.

Augmented SFT with supervised ABSTAIN targets has retained-row AURC 0.166 / 0.184 (sequence log-probability / first-token probability); its confidence-run native coverage is 58.5%.

L Difficulty, Evidence Type, and Review Budgets

These analyses reuse the development-exposed test split. They characterize heterogeneity; they are not independent confirmation of a selected method.

Difficulty defined on different records. We define cell difficulty from plain-SFT predictions on the development split, requiring at least five development records per cell. Strict accuracy below 50% defines difficult cells, 50–90% intermediate cells, and at least 90% easy cells; cells with insufficient development support remain a separate stratum. The test records in the 39 difficult cells improve from 14.3% to 24.1–24.2% under gold-only adaptation, while easy cells fall from 99.3% to 91.3–91.6% (Table 24). Across the same three seeds, 789 individual test rows are repaired in every seed and 1,558 rows are damaged in every seed relative to plain SFT. This establishes repeatable heterogeneity, not its causal mechanism.

Table 24: Test accuracy (%) within strata defined from development records. Gold-only ranges span seeds 17, 29 and 43. All test rows remain included.

Dev-defined stratum	Test rows	Cells	Plain SFT	Gold-only
Difficult	2,655	39	14.3	24.1–24.2
Intermediate	1,326	24	68.7	66.3–66.7
Easy	9,414	246	99.3	91.3–91.6
Insufficient dev support	6,605	569	78.1	71.3–71.9

Evidence types. The augmented prompt uses either same-system neighbours with expanded candidates, an official ISCO-88 correspondence candidate block, or neither. Official correspondence blocks count as evidence throughout the corrected analysis. These are different source/input regimes, not randomized treatments (Table 25); their contrasting accuracies cannot alone identify the effect of adding evidence to an individual record.

Table 25: Strict accuracy (%) by evidence regime of the augmented input. Plain SFT is evaluated on its own plain prompt for the same records.

Evidence regime	Rows	Plain SFT	Gold-only
Neighbours / expanded candidates	13,951	79.1	72.4–72.5
Official correspondence candidates	4,350	94.2	94.8–95.2
No evidence block	1,699	39.7	36.2–36.7

Equal review budgets do not imply equal institutional coverage. At 70% automatic answering, confidence-ranked plain SFT leaves 10.6 automatic errors per 1,000 inputs and completely defers 59 cells. The gold-only models leave 48.5–49.6 errors and completely defer 5–8 cells. All policies defer 300 records per 1,000; no assumption is made about the correctness of subsequent human review. Broader institutional coverage is therefore not a free reliability gain. Each comparison uses the answers and confidence from the same inference pass. Ranking within the test batch fixes review volume but is not a development-calibrated deployment threshold.

Sensitivity to dependence between cells. We resample entire country/entity or literal-system-name blocks, preserving their constituent cells and row counts, and recompute the worst-decile mean for both policies in each paired replicate. The gold-only tail differences remain positive under both alternatives (Table 26). These metadata blocks are dependence proxies, not verified independent source families. Resampling the five source categories gives intervals crossing zero; with only five blocks, this is a coarse stress test, not a calibrated population interval. None of these checks removes development exposure.

Unused training data as reference material. The existing 50,000 training records are a subset of 962,029 eligible records in the official training split after removing identity-source and trivial rows. Expanding the reference pool preserves $k = 5$, code-distance retrieval, the original tie order, candidate expansion, and the official correspondence route. No development or test record ID occurs in the pool; this check does not certify semantic or source-lineage independence. On

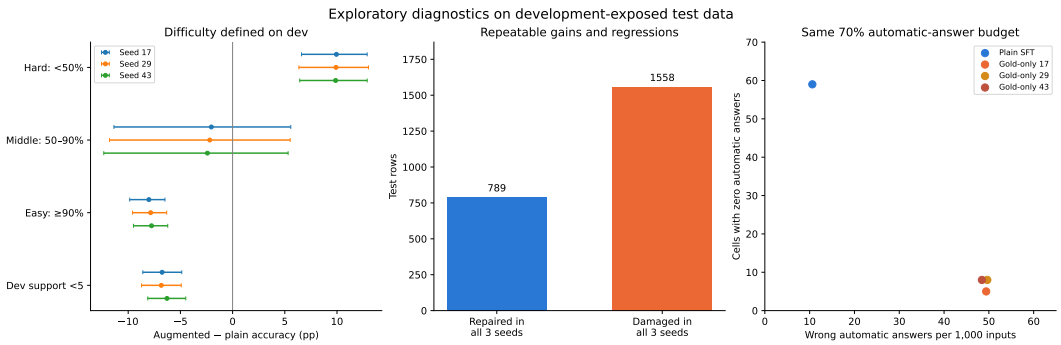


Figure 8: Exploratory heterogeneity diagnostics. Left: paired cell-bootstrap intervals within dev-defined strata, separately for three training seeds. Centre: rows repaired or damaged in every seed. Right: automatic errors and fully deferred cells at equal 70% answering coverage. Cell sizes and source dependencies limit population-level interpretations.

Table 26: Paired bootstrap 95% intervals for gold-only minus plain-SFT CVaR@10%, in percentage points; 2,000 replicates per resampling unit. Training seeds are reported separately.

Resampling unit	Blocks	Seed 17	Seed 29	Seed 43
Institutional cell	878	[+3.9, +12.3]	[+3.4, +11.6]	[+3.3, +11.0]
Country/entity	162	[+2.1, +12.8]	[+1.6, +12.3]	[+2.1, +11.3]
Literal system name	807	[+3.6, +12.3]	[+3.0, +11.7]	[+3.1, +11.2]
Source category (stress test)	5	[-1.3, +13.6]	[-1.7, +13.2]	[-1.1, +11.6]

the same 5,000 development queries, candidate inclusion and nearest-reference correctness improve (Table 27). These are retrieval diagnostics, not improvements in a trained policy's accuracy; gold inclusion does not establish that the reference evidence is sufficient or unambiguous.

Table 27: Training-only reference-pool expansion on dev5k (% of all queries). Candidate hit means the candidate block contains the primary gold code; first-reference match counts queries without neighbours as misses.

Reference pool	Has neighbours	Candidate hit	First-reference match
50,000 training references	69.08	54.36	17.98
962,029 training references	69.56	67.30	35.04

M PROSPECTIVE NOL SOURCE/VERSION TRANSFER

We evaluate previously trained, frozen checkpoints on the official New Zealand National Occupation List V1.0.0 to ISCO-08 direct correspondence.[3] The protocol, input files, checkpoint chains and inference code were hashed before the first query prediction. The 1,288 codes with valid official targets comprise 1,219 singletons and 69 multi-target cases; four further codes have no equivalent target. The primary endpoint uses singletons only, retaining a complete exclusion inventory. Of the 1,219 singleton codes, 243 are fixed support examples and 976 are queries, split by a code-only hash. No query answer enters the reference pool. Five numerically nearest support codes provide the reference condition; the no-reference condition has the same abstention instruction. Plain SFT additionally uses its original instruction.

All 15 conditions use the same local MPS device and software environment, greedy decoding, batch size eight and a 16-token generation limit. These are separate inference passes from the earlier CUDA benchmark results. We report all prespecified checkpoints and conditions in Table 28. Primary strict scoring retains historical code normalization; the raw-format column additionally requires a response of exactly four digits. Invalid non-abstention responses count as answered errors.

Table 28: Frozen-checkpoint NOL transfer on the same 976 singleton queries. All metrics are percentages; raw format is accuracy over all queries, not a conditional score.

Checkpoint	Input	Strict	Coverage	Selective	Raw format
Plain SFT	Original instruction	11.8	100.0	11.8	11.8
Plain SFT	No references	6.8	100.0	6.8	6.8
Plain SFT	Five support references	23.2	68.9	33.6	23.2
Gold-only 17	No references	9.6	100.0	9.6	9.6
Gold-only 17	Five support references	29.2	100.0	29.2	29.2
Gold-only 29	No references	9.1	100.0	9.1	9.1
Gold-only 29	Five support references	29.6	100.0	29.6	29.6
Gold-only 43	No references	9.9	100.0	9.9	9.9
Gold-only 43	Five support references	28.9	100.0	28.9	28.9
ABSTAIN-target 17	No references	6.1	100.0	6.1	6.1
ABSTAIN-target 17	Five support references	27.4	49.0	55.9	27.4
RL unadapted 17	No references	4.5	100.0	4.5	4.5
RL unadapted 17	Five support references	20.9	57.5	36.4	20.9
RL adapted 17	No references	9.4	100.0	9.4	9.4
RL adapted 17	Five support references	29.3	100.0	29.3	29.3

Table 29: Within-checkpoint effect of five labeled support references versus no references, in strict-accuracy percentage points. Intervals use 2,000 paired row-bootstrap replicates.

Checkpoint	Difference	95% interval
Plain SFT	+16.4	[+13.4, +19.3]
Gold-only 17	+19.6	[+16.4, +22.8]
Gold-only 29	+20.5	[+17.2, +23.7]
Gold-only 43	+19.0	[+15.7, +22.1]
ABSTAIN-target 17	+21.2	[+18.3, +24.2]
RL unadapted 17	+16.4	[+13.6, +19.2]
RL adapted 17	+19.9	[+16.6, +23.2]

Scope and exposure. One English institutional cell sharing ANZSCO/OSCA lineage cannot validate cross-institution CVaR gains. The title audit finds 426 query titles in train50k, 144 elsewhere in audited history, 26 near matches, and 380 without observed matches; this does not certify pretraining novelty. Labeled support is not zero-shot. Row intervals condition on queries, excluding institution and training-seed uncertainty. The artifact retains all prespecified contrasts, exposure strata and realizable matched-coverage diagnostics. Intervals are descriptive, without multiplicity adjustment.

[3] Statistics New Zealand ARIA, official direct correspondence: `https://aria.stats.govt.nz/aria/#ConcordanceView:uri=http://stats.govt.nz/cms/ConcordanceVersion/JIT3Ftp7ssjxszAh`.

Post-hoc support-only baselines. After observing the frozen model results, we added two deterministic diagnostics using exactly the same 243 labeled support codes. Copying the nearest numeric code's target yields 35.8% strict accuracy at full coverage, versus 27.6% for plurality among the same five references (ties resolved by nearest occurrence). Neither rule uses query labels. Every trained checkpoint therefore improves with references but none outperforms the simplest nearest-code rule: all seven fall 6.1–14.9 points below it, and every descriptive paired row interval excludes zero on the negative side. These are additional post-hoc comparisons, not part of the original 15-condition protocol. The correct target occurs among the five support labels for 48.6% of queries; this oracle inclusion diagnostic uses query gold and is not a deployable policy or an upper bound on unrestricted model predictions.

Table 30: Post-hoc reference-label availability diagnostic. The correct target occurs among the five support labels for 474 queries and is absent for 502. Accuracy and coverage are percentages.

Checkpoint	Present: strict	Coverage	Absent: strict	Coverage
Plain SFT	45.1	77.4	2.4	60.8
Gold-only 17	53.8	100.0	6.0	100.0
Gold-only 29	55.7	100.0	5.0	100.0
Gold-only 43	53.6	100.0	5.6	100.0
ABSTAIN-target 17	55.9	69.8	0.4	29.3
RL unadapted 17	41.8	66.9	1.2	48.6
RL adapted 17	54.6	100.0	5.4	100.0

These gold-defined strata are descriptive, not deployable selection rules or a causal error decomposition. Target inclusion does not certify sufficient or unambiguous evidence; its absence does not prevent answering from other knowledge. Both strata and all seven checkpoints are retained.

Adaptation against the original deployment. The prespecified comparison against the plain model in its own prompt format is adverse: the adapted gold-only checkpoints score 1.8–2.7 points below the plain SFT model they continue from, with 2 of 3 intervals excluding zero. On this source the adaptation recipe does not reproduce its benchmark advantage, and the reference gain above is a within-checkpoint input effect, not evidence that the recipe is preferable here.

N FIXED-WEIGHT EVIDENCE INTERVENTIONS

We fix development IDs, checkpoint weights, decoding and software, and vary the prompt evidence. References and candidates form a paired two-by-two intervention; official ISCO-88 correspondence blocks are candidate evidence. A fifth condition removes only the optional abstention instruction. All conditions are evaluated on dev5k and are exploratory development analyses. Table 31 reports all five conditions for plain SFT and gold-only seed 17. Input removal can induce distribution shift; cross-checkpoint differences additionally include continued training and cannot alone identify the contribution of training evidence.

Table 31: Fixed-weight input interventions on dev5k (%). Every condition retains all 5,000 records.

Input	Plain strict	Plain coverage	Gold strict	Gold coverage
Full evidence	41.1	80.2	72.8	100.0
References only	48.2	81.0	67.7	100.0
Candidates only	38.6	78.1	66.0	100.0
No evidence	73.3	100.0	74.3	100.0
Full, no abstain instruction	45.1	100.0	71.0	100.0

Table 32: Fixed-model effects of expanding the training reference pool from 50,000 to 962,029 eligible rows. Accuracy is in percent and differences in percentage points; 95% intervals resample paired institutional cells.

Checkpoint	50k pool	Full pool	Difference	Interval
Plain SFT	41.1	53.2	+12.1	[+10.0, +14.2]
Gold-only 17	72.8	75.9	+3.1	[+1.0, +5.5]
Gold-only 29	72.6	75.9	+3.3	[+1.3, +5.8]
Gold-only 43	72.9	76.0	+3.0	[+1.0, +5.4]

Every gold-only seed is retained. The reference expansion changes neither weights nor query labels. Gold inclusion is a retrieval diagnostic rather than a guarantee of sufficient evidence. Source-specific metrics, overall repaired and damaged rows, strict per-cell CVaR with abstentions counted as errors, raw-format sensitivity, and reference/candidate interaction contrasts are reported in the artifact. An additional metadata audit finds no exact query-ID overlap, but 17 of 5,000 development queries share country, system and source code with the 50k pool, increasing to 223 with the expanded pool. These title-variant records flag pool-level dependence despite distinct IDs. The retriever excludes identical source-code strings; an audit of actual displayed references found no normalized query-code matches. Pool overlap therefore does not establish direct same-code answer exposure, and the expansion is not a test of entirely novel source systems.