KEEP OR REPAIR? CERTIFIED ACCEPTANCE FOR CONSERVATIVE QUERY REPAIR

Anonymous authors
Paper under double-blind review

ABSTRACT

Language-model repairers are a natural add-on to knowledge base question answering (KBQA) parsers, but once the parser is accurate, repair is governed by the damage it does rather than the errors it fixes. When upstream accuracy reaches 92%, as on our KQA Pro development cohort, even a small fraction of wrongly edited correct answers can outweigh gains on incorrect questions. We formalize *conservative query repair* and separate what to propose from what to accept. Our certified acceptance gate accepts a proposed query only when the repairer's own `KEEP`-sequence probability κ on it exceeds a threshold locked by a Clopper–Pearson bound; the rule provably controls the population regression rate, needs no training, and applies to repairers that expose `KEEP` token probabilities. On 11,797 KQA Pro questions, the gate turns a net-negative frozen repairer (net -18) into net $+57$, with a 95% upper bound of 0.35% on its regression rate against a 0.5% target, and κ achieves higher proposal-ranking AUROC than token likelihood or a prompted judge. Fine-tuned repairers achieve similar net gains ($+55$ to $+57$), raising `KEEP` recall from 71% to 88% without improving error-ranking AUROC. After target-domain recalibration on WebQSP and CWQ, the fine-tuned repairer's gate never exceeds its risk target across 1,000 random splits, but yields no positive net gains.

1 INTRODUCTION

Knowledge base question answering (KBQA) translates natural language questions into executable queries over a knowledge graph. Systems that combine language-model generation with schema retrieval and execution feedback now exceed 90% accuracy on compositional benchmarks (Cao et al., 2022; Luo et al., 2024; Huang et al., 2024b; Gao et al., 2025), which makes *repair*—a second model pass that fixes the remaining errors—a natural next step. Repair, however, also sees every query that is already correct, and each unnecessary edit risks a *regression*: a correct answer turned wrong.

At high accuracy this risk dominates. When 92% of upstream queries are already correct, as on our KQA Pro development cohort, even a small fraction of wrongly edited correct answers can outweigh the corrections made on the few incorrect ones: accuracy rises only if the correction rate on incorrect queries exceeds the regression rate on correct ones by a factor of about 11, and by more than 15 on the official validation set, where upstream accuracy is 94% (§2.1). Off-the-shelf models do not clear this bar: a 14B repairer fixes 108 official-set queries but breaks 126 (Figure 1a shows one), *lowering* accuracy. This is consistent with growing evidence that language models degrade their own outputs without reliable external feedback (Huang et al., 2024a; Madaan et al., 2023). Tool-augmented verification mitigates unreliable reasoning (Yao et al., 2023; Gou et al., 2024), and selective correction with fallback has been explored in KBQA to avoid regressions (Huang & Liu, 2025); what is missing is a way to decide *which* repairs to accept with a formal guarantee on how many correct answers are lost.

We therefore treat repair as a risk-control problem and separate proposing a repair from accepting it. The repairer emits either `KEEP` or a replacement query P', and we accept P' only when the repairer's own `KEEP`-sequence probability on the replacement, $\kappa(P')$, exceeds a threshold τ—in effect, we ask the model whether it would keep P' had P' been the upstream query. The threshold is locked on a calibration split as the most permissive value whose one-sided Clopper–Pearson bound

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.

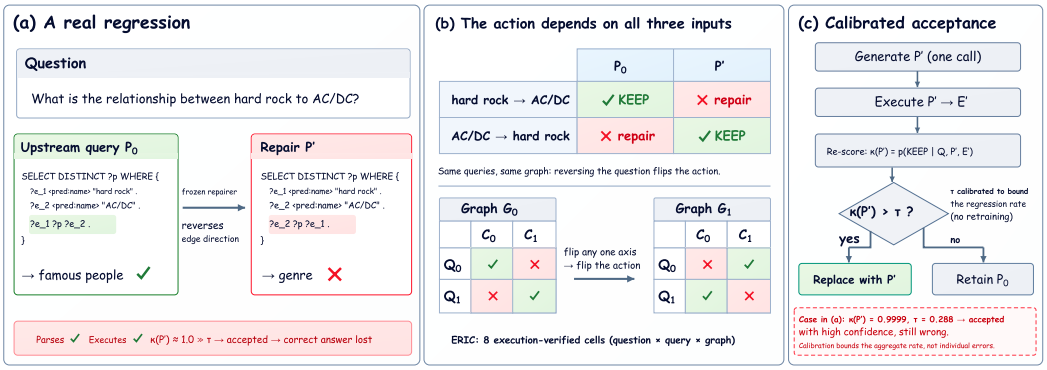

Figure 1: (a) A real regression on KQA Pro official validation (ID 420): the frozen repairer reverses the edge direction, the replacement passes every executability check and is accepted with $\kappa(P') = 0.9999 \gg \tau = 0.288$, and a correct answer is lost. (b) The correct action depends jointly on the question, the candidate, and the graph; reversing the question flips it (second row illustrative). (c) Calibrated acceptance: a replacement is accepted only when $\kappa(P')$ exceeds a threshold calibrated to bound the aggregate regression rate, which does not guarantee individual repairs, as (a) shows.

on the regression rate stays below a target α (Clopper & Pearson, 1934; Angelopoulos et al., 2021), and is never revisited on evaluation data. The gate needs no training and applies to repairers that provide KEEP-sequence probabilities. On the KQA Pro official set (11,797 questions) it turns the frozen repairer from net -18 into net $+57$ ($p = 3\times10^{-12}$), reducing the number of wrong answers by 8.0%, and its observed regression rate has a 95% upper bound of 0.35%, below the 0.5% target it was calibrated for; moreover, κ achieves higher proposal-ranking AUROC than the proposal's token likelihood and a separately prompted judge. We also derive the minimum calibration size the certificate needs; a 72-configuration repeated-split analysis examines how calibration size affects keep-all outcomes, and recalibrated on WebQSP and CWQ the gate never exceeds its risk target.

The natural alternative to an external gate is to fine-tune the repairer to be conservative itself, and we use fine-tuning as our main comparison. Fine-tuning achieves a similar net gain ($+55$), raises KEEP recall from 71% to 88%, and does not improve error-ranking AUROC (0.872 before, 0.865–0.870 after). These observations are consistent with greater conservatism rather than improved AUROC, but do not establish an identical decision rule. A supplementary study with execution-verified intervention groups, which isolate the joint dependence in Figure 1b, finds that richer supervision does not change this picture. These results motivate explicit control of acceptance, while limited proposal coverage constrains the gains available in our experiments. Our contributions:

1. **Formulation and gate**: conservative query repair with full-population correction–regression accounting, and a training-free acceptance gate with a finite-sample guarantee on the regression rate and a closed-form minimum calibration size (§2.1–§2.2).
2. **Evidence**: the gate turns a net-negative repairer into a net-positive one, achieves higher AUROC than token likelihood and a prompted judge, and keeps regressions within target under scarce calibration data and domain shift (§3.2, §3.4).
3. **Analysis**: fine-tuning increases conservatism without improving error-ranking AUROC, and adding the gate changes few outcomes in this setting (§3.3); group-structured supervision gives no measurable additional benefit (§3.5).

2 Method

2.1 Problem Setting

Given a question Q, a candidate query P_0 from an upstream system, and a knowledge graph G, the repairer outputs either KEEP or a replacement query. It receives $x = \text{Serialize}(Q, P_0, \mathcal{S}, \mathcal{F}, E_0, \rho)$, where $\mathcal{S}$ and $\mathcal{F}$ are visible schema and facts, $E_0 = \text{Exec}(P_0, G)$ is execution feedback, and ρ is a scope tag indicating the repair granularity (e.g., full query vs. single clause). Let $b_i, h_i \in \{0, 1\}$

indicate whether the original and final answers are correct. On N questions,

$$R = \textstyle\sum_i (1-b_i)h_i, \quad D = \textstyle\sum_i b_i(1-h_i), \quad \text{Net} = R - D. \tag{1}$$

Each correction offsets exactly one regression, but the two are drawn from populations of very different size. The expected net gain per question is $(1-\pi)\, r - \pi\, d$, where π is the fraction initially correct and r, d are the conditional correction and regression rates. Positive gain therefore requires $r > \frac{\pi}{1-\pi}\, d$: $r > 11.4\, d$ on our development cohort ($\pi = 0.919$) and $r > 15.5\, d$ on the KQA Pro official validation set ($\pi = 0.940$).

Acceptance decomposition. Every replacement must be *accepted* before it can change an answer. Let V indicate acceptance, with acceptance rates $\gamma_0 = \Pr(V{=}1 \mid b{=}0)$, $\gamma_1 = \Pr(V{=}1 \mid b{=}1)$ and conditional outcomes $\eta_0 = \Pr(h{=}1 \mid b{=}0, V{=}1)$, $\eta_1 = \Pr(h{=}0 \mid b{=}1, V{=}1)$. Then $r = \gamma_0\eta_0$, $d = \gamma_1\eta_1$, and

$$\mathbb{E}[h - b] = (1 - \pi)\, \gamma_0\, \eta_0 \;-\; \pi\, \gamma_1\, \eta_1. \tag{2}$$

Expanding a fixed acceptance set cannot decrease either γ_0 or γ_1; at high π, the regression term has the larger population weight. A useful repairer therefore needs an acceptance rule that keeps $\gamma_1\eta_1$ small while preserving as much of $\gamma_0\eta_0$ as possible; §2.2 provides such a rule with a certified bound on d.

2.2 Calibrated Acceptance

We separate the acceptance decision from the repair model. Let q denote the `KEEP` token sequence followed by EOS. The `KEEP`-sequence probability $\kappa(P') = p_\theta(q \mid x')$, where x' is the input re-serialized with the proposed replacement P' and its execution feedback $E' = \mathrm{Exec}(P', G)$ in place of P_0 and E_0, asks the repairer whether it would itself keep P' if P' were the upstream query. Computing it requires access to token probabilities, costs one scoring pass on x', and requires neither additional training nor a separate judging prompt. Because κ is the probability of the complete sequence, $\kappa > 1/2$ implies that greedy decoding emits `KEEP`, but not conversely (Appendix A.3); thresholds well below $1/2$ are therefore meaningful.

The threshold is locked on a calibration split S_{cal} and never revisited on evaluation data (protocol in Figure 4, Appendix A.2). For each candidate τ on a grid fixed in advance, we count the regressions $d(\tau)$ among the originally correct questions of S_{cal} and compute the one-sided $(1{-}\delta)$ Clopper–Pearson upper bound $\bar{p}(\tau)$ on the regression rate (Clopper & Pearson, 1934). We select the smallest τ with $\bar{p}(\tau) \le \alpha$ and $d(\tau) \le d_{\max}$, using $\alpha = 0.5\%$, $\delta = 5\%$, and $d_{\max} = 5$; for the frozen model this yields $\tau = 0.288$. The cap $d_{\max}$ is an additional conservative constraint that only shrinks the feasible set and does not affect the guarantee below. If no threshold is feasible, the rule accepts nothing and keeps every upstream query; otherwise it applies the smallest feasible threshold, with ties rejected by the strict inequality.

Because regressions can only decrease as the threshold rises, choosing the most permissive certified threshold needs no multiplicity correction: if calibration questions are i.i.d. draws from the deployment distribution and the scoring pipeline is fixed before calibration, $\Pr[\mathcal{R}(\hat{\tau}) > \alpha] \le \delta$, where $\mathcal{R}$ is the population regression rate (Proposition 2). The same analysis gives a closed-form requirement on calibration data: this Clopper–Pearson rule requires at least $\ln\delta / \ln(1-\alpha)$ originally correct questions, which is 598 at $\alpha = 0.5\%$ and $\delta = 5\%$ (Corollary 1); below that, the rule keeps every upstream query. The certificate bounds the aggregate regression rate, not the correctness of individual repairs.

2.3 Repairers and Inference

The gate requires a repairer that emits `KEEP` or a replacement and provides token probabilities. We gate an untuned (frozen) backbone and, as a comparison, repairers fine-tuned with the generation loss $\ell_{\mathrm{gen}}(x, y) = \frac{1}{2}\,\ell_1 + \frac{1}{2(T-1)} \sum_{t=2}^{T} \ell_t$, where $\ell_t = -\log p_\theta(y_t \mid x, y_{<t})$ for a target of length T; the first term weights the keep-or-repair decision equally with the body of the output. At inference the model generates one output o. If $o =$ `KEEP`, P_0 is retained. Otherwise, o is deterministically compiled into $\widehat{P} = \mathcal{C}(o; Q, P_0, \mathcal{F})$, which replaces P_0 only if the operational validator accepts it,

$\mathcal{V}_{\text{op}}(Q, P_0, \widehat{P}, G) = 1$, and, when the gate is active, $\kappa(\widehat{P}) > \hat{\tau}$. The validator checks parsing, execution, and answer format; the compiler handles namespace normalization; neither accesses reference answers.

3 EXPERIMENTS

We ask whether calibrated acceptance turns candidate repairs into net gains while bounding regressions, and which score it should use (**RQ1**, §3.2); how fine-tuning, our main comparison, reaches a similar outcome and why gating adds little on top of it (**RQ2**, §3.3); and whether the gate keeps regressions within target on other knowledge graphs (**RQ3**, §3.4). A supplementary analysis asks whether group-structured supervision improves on standard fine-tuning (§3.5). We report null and negative results alongside positive ones.

3.1 EXPERIMENTAL SETUP

Cohorts. Our primary evaluation uses the **KQA Pro** (Cao et al., 2022) official validation set: 11,797 questions, 11,084 initially correct and 713 initially incorrect (upstream accuracy 93.96%). We also use a 4,719-question internal development cohort (4,338 correct, 381 incorrect; 91.93%), split into a *selection half* (2,359) on which each model's acceptance rule is locked and a *report half* (2,360). Both KQA Pro cohorts were used during development. The reported acceptance thresholds are selected on the selection half and held fixed on the report half and official set. For transfer we use cached cohorts of **WebQSP** (Yih et al., 2016) (1,639 questions) and **CWQ** (Talmor & Berant, 2018) (3,531 questions). Every evaluation covers the full cohort, including questions the upstream parser already answers correctly. The internal cohort uses an independently trained *clean-source BART semantic parser*, initialized from `facebook/bart-base`, with fixed `merged_01` normalization and selection rules. Official-validation queries come from the upstream cache of the `merged_01` deployment pipeline; `merged_01` denotes the pipeline, not the internal parser.

Metrics and tests. We report R, D, and Net (Eq. 1), the repair rate RR $= R/N_w$ and the regression rate DR $= D/N_c$, where N_w and N_c are the numbers of initially incorrect and initially correct questions, and execution accuracy; on WebQSP and CWQ we additionally report answer F1, their standard metric, while R and D still count per-question correctness changes. Paired comparisons use two-sided exact McNemar tests on per-question correctness, with Holm correction within each comparison family, and question-level bootstrap confidence intervals, which do not include variation across training seeds.

Methods. All methods share the backbone (Qwen3-14B-AWQ; Yang et al., 2025), the cached upstream queries P_0, the retrieved context, the deterministic compiler, and the operational validator $\mathcal{V}_{\text{op}}$. *Frozen repair* prompts the untuned backbone; *Frozen+κ* adds the gate of §2.2. *Same-data SFT*, the fine-tuned comparison, trains with ℓ_{gen} on natural repair examples and on the cells of the intervention groups of §3.5; ERIC and *Random Groups* are the supplementary variants described there. Fine-tuned models use LoRA (Hu et al., 2022) (rank 16, scaling 32) on attention and MLP projections, AdamW (learning rate 2×10^{-5}, weight decay 0.01, gradient clipping 1), and 8 parameter updates, each on 8 natural examples and one complete eight-cell group; exposures are matched across arms within each seed (Appendix H). Acceptance rules are locked per model on the selection half with $\alpha = 0.5\%$, $\delta = 5\%$, $d_{\max} = 5$: the frozen model accepts when $\kappa(P') > 0.288$, Same-data SFT when $\kappa(P') > 0.0086$, and ERIC's selected threshold accepts every validated candidate in these evaluations, so ERIC and ERIC+κ coincide.

3.2 CALIBRATED CONSERVATIVE REPAIR (RQ1)

Calibrated acceptance turns a net-negative repairer into a net-positive one. Without filtering, the frozen model makes 108 corrections but 126 regressions on the official set (net -18, DR $=$ 1.14%; Table 1). Its selection-locked κ rule removes 98 regressions while giving up 23 corrections, yielding 85/28/+57 and raising accuracy from 93.96% to 94.44%. Paired against unfiltered frozen repair, Frozen+κ wins on 98 questions and loses on 23 (ΔAcc $= +0.64$ pp, 95% CI $[+0.45, +0.82]$; $p = 3.2\times10^{-12}$). In relative terms, the gate reduces the number of wrong answers from 713 to 656 (8.0%), whereas unfiltered repair increases it to 731. The observed regression rate of 0.253% has a

Table 1: KQA Pro official validation (11,797 questions; 713 initially incorrect, 11,084 initially correct). Acceptance thresholds were locked on the selection half and not re-selected. CP: one-sided 95% Clopper–Pearson upper bound on DR, reported for rules that must meet the target $\alpha = 0.5\%$ and for Random Groups. Lower block: fine-tuned comparisons (Same-data SFT) and supplementary variants (§3.5). Single training seed.

Method	$R\uparrow$	$D\downarrow$	Net↑	RR (%)	DR (%)↓	CP (%)	Acc. (%)↑
No Repair	0	0	0	0.00	0.000	–	93.96
Frozen Repair	108	126	−18	15.15	1.137	–	93.80
Frozen+κ (ours)	85	28	+57	11.92	0.253	0.346	94.44
Same-data SFT	89	34	+55	12.48	0.307	–	94.42
SFT+κ	88	33	+55	12.34	0.298	0.398	94.42
ERIC	90	33	+57	12.62	0.298	0.398	94.44
Random Groups	96	51	+45	13.46	0.460	0.581	94.34

Table 2: Acceptance scores for the frozen repairer. AUROC/AP: separating correcting from regressing proposals on the selection half. R/D/Net: report half (2,360 questions), with each score's threshold locked on the selection half. Fine-tuned models are in Appendix D.

Acceptance score	AUROC↑	AP↑	$R\uparrow$	$D\downarrow$	Net↑	DR (%)↓
KEEP-sequence probability $\kappa(P')$	0.839	0.863	38	4	+34	0.184
Mean token log-probability of P'	0.623	0.831	31	4	+27	0.184
$\kappa(P')$ minus mean log-probability	0.842	0.891	38	4	+34	0.184
Prompted judge	0.817	0.895	37	6	+31	0.276

one-sided 95% Clopper–Pearson upper bound of 0.346%, below the target $\alpha = 0.5\%$ for which the threshold was calibrated; this bound is computed on the official outcomes and complements, rather than replaces, the calibration guarantee of Proposition 2. The gate does not eliminate confident errors: the regression in Figure 1a is accepted with $\kappa(P') = 0.9999$ and is one of the 28 that remain, which is why the guarantee is stated for the aggregate rate. The same pattern holds on the report half, where Frozen+κ reaches 38/4/+34 against 43/21/+22 unfiltered (17 wins, 5 losses; raw $p = 0.017$, Holm $p = 0.10$).

The model's own `KEEP` probability is an effective acceptance score. Table 2 compares $\kappa(P')$ with three alternatives, each thresholded on the selection half under the same protocol. The generation likelihood of the proposal is a poor indicator of whether it helps (AUROC 0.623 vs. 0.839) and yields seven fewer net corrections. A separately prompted judge ranks proposals nearly as well (AUROC 0.817), but at its locked threshold it admits two more regressions and ends three questions lower, while requiring an additional prompt per proposal; these outcome differences are small, and the judge has higher AP (0.895 vs. 0.863), so neither score dominates both ranking metrics. Subtracting the mean log-probability from κ slightly improves ranking but leaves the outcome unchanged. The question the repairer is trained to answer, whether a query should be kept, is thus also the right question to ask of its own proposal.

Robustness of the acceptance rule. We repeat the calibration/test split 1,000 times for each model, calibration size (250, 500, 1,000, 2,359), target α (0.25%, 0.5%, 1%), and rule (empirical or certified); Figure 2 shows $\alpha = 0.5\%$. The empirical rule, which thresholds on the observed calibration regression rate alone, exceeds α in 4–21% of splits for the frozen model; the certified rule reduces this to at most 1% up to 1,000 calibration questions and 10.8% at 2,359, and never exceeds α for SFT or ERIC. At 2,359 calibration questions, the test split has about 2,170 initially correct questions; 11 regressions already exceed $\alpha = 0.5\%$. Thus finite-test variation can cause exceedance even when population risk is controlled; the observed 10.8% alone does not identify the cause. Exceedance on finite test splits from one question pool is therefore not the probability that the population-level guarantee fails (Appendix B). When calibration data are scarce, the certified rule fails safe: at $\alpha = 0.5\%$, 250 or 500 calibration questions provide fewer than the 598 originally correct questions that Corollary 1 requires, so every upstream query is kept; Table 3 compares this necessary condition with observed keep-all outcomes. At 1,000 calibration questions the certified

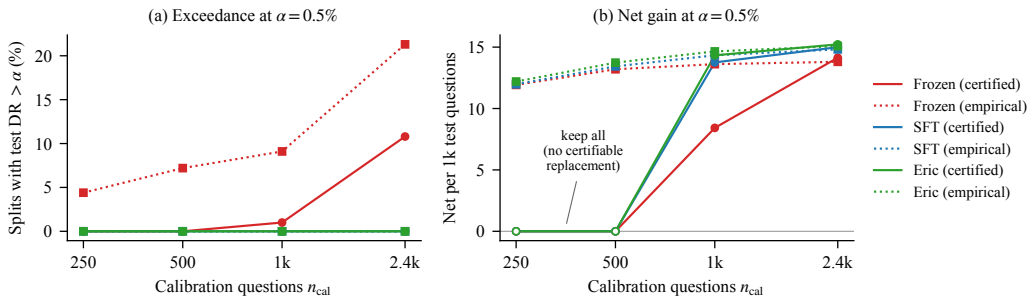

Figure 2: Certification reliability versus calibration size on 1,000 random calibration/test splits of the development cohort at $\alpha = 0.5\%$. (a) Fraction of splits whose test regression rate exceeds α. (b) Mean test Net per 1,000 test questions. Hollow markers: the certified rule keeps every upstream query in all splits. All 72 configurations are in Appendix B.

Table 3: Corollary 1 against observation. n_c: originally correct calibration questions (development: 2.5–97.5% range over 1,000 splits; transfer: expected value for a 30% split). Each cell gives the prediction (KA: below the necessary sample-size bound; —: not excluded by that bound) and whether the certified rule kept every query in all splits for all models.

Cohort	n_c	$\alpha = 0.25\%$ (need 1,197)	$\alpha = 0.5\%$ (need 598)	$\alpha = 1\%$ (need 299)
Dev, $n_{\text{cal}} = 250$	222–238	KA / KA	KA / KA	KA / KA
Dev, $n_{\text{cal}} = 500$	449–470	KA / KA	KA / KA	— / —
Dev, $n_{\text{cal}} = 1{,}000$	903–934	KA / KA	— / —	— / —
Dev, $n_{\text{cal}} = 2{,}359$	2,151–2,185	— / —	— / —	— / —
WebQSP (30%)	≈358	KA / KA	KA / KA	— / —
CWQ (30%)	≈777	KA / KA	— / —	— / —

rule recovers 96–98% of the empirical rule's net gain for fine-tuned models and 62% for the frozen model; at 2,359 it matches the empirical rule for all three.

3.3 Fine-Tuning as a Comparison (RQ2)

Fine-tuning matches the gate, and gating it changes little. Same-data SFT reaches 89/34/+55 on the official set, within two questions of Frozen+κ (Table 1). Its locked gate removes exactly one correction and one regression (SFT+κ: 88/33/+55), and on the report half gated and ungated SFT are identical question by question (40/3/+37, three questions above Frozen+κ; SFT wins 8 and loses 5; $p = 0.58$). The fine-tuned model nevertheless passes the certificate with margin (CP bound 0.398%), so the gate still adds a guarantee.

Fine-tuning moves the operating point. On the development cohort (Table 4), fine-tuning cuts regressions from 45 to 8–9 at a cost of 12 corrections. `KEEP` precision is essentially unchanged (97–98%), while recall rises from 71% to 88% and the number of accepted replacements with a wrong answer nearly halves (138 to 73–74). A query the model keeps cannot regress. Higher explicit `KEEP` recall reduces the fraction of initially correct queries available for editing, consistent with more conservative behavior.

Fine-tuning does not improve AUROC. We test this at two levels: *error ranking*, whether $\kappa(P_0)$ separates incorrect from correct upstream queries, and *proposal ranking*, whether $\kappa(P')$ separates correcting from regressing replacements (the quantity in Table 2). Neither AUROC improves in these comparisons; AP shows smaller, seed-dependent changes (Appendix F). Scoring every upstream query by $-\log \kappa(P_0)$ with originally incorrect queries as positives, the frozen model reaches AUROC 0.872 on the development cohort and 0.864 on the official set; across three training seeds, fine-tuned models reach 0.865–0.870 and 0.852–0.861 (Appendix F). The same holds for proposals: when the frozen model's 150 report-half proposals are filtered by each model's own locked κ rule,

Table 4: Internal development cohort (4,719 questions; 381 initially incorrect) under the retained inference pipeline, without κ filtering. Single training seed. KEEP-P/KEEP-R (%): precision and recall of explicit `KEEP` outputs. Acc. wrong: accepted replacements whose answer is wrong.

Method	$R\uparrow$	$D\downarrow$	Net$\uparrow$	DR (%)$\downarrow$	KEEP-P	KEEP-R	Acc. wrong$\downarrow$
No Repair	0	0	0	0.000	–	–	–
Frozen Repair	92	45	+47	1.037	98.37	70.98	138
Same-data SFT	80	9	+71	0.207	97.38	88.24	74
ERIC	80	8	+72	0.184	97.37	88.61	73

Table 5: Target-domain calibration of the ERIC repairer's gate. Exceed.: splits whose test regression rate exceeds α. Keep-all: splits in which no replacement is certifiable.

Dataset	α (%)	Exceed. (%)	Keep-all (%)	R	D	Net	DR (%)
WebQSP	0.25	0.0	100.0	0.00	0.00	0.00	0.000
WebQSP	0.5	0.0	100.0	0.00	0.00	0.00	0.000
WebQSP	1	0.0	0.0	1.19	1.76	-0.57	0.211
CWQ	0.25	0.0	100.0	0.00	0.00	0.00	0.000
CWQ	0.5	0.0	0.0	1.71	1.96	-0.26	0.108
CWQ	1	0.0	0.0	2.13	3.51	-1.39	0.194

the frozen model and all six fine-tuned checkpoints reach Net between $+33$ and $+35$, and fine-tuned κ separates correcting from regressing proposals no better than the frozen model's (AUROC 0.827 and 0.830 for seed-1 SFT and ERIC, vs. 0.839; Appendices E and D).

Two routes to a similar operating point. The gate selects a conservative threshold explicitly. Fine-tuning produces a similarly conservative outcome without increasing AUROC, and adding the gate changes few outcomes in this setting. The two routes are not identical question by question: ERIC and Frozen+κ disagree on 74 official questions, split exactly 37 to 37 ($p = 1$), so they reach the same net gain through different individual decisions.

3.4 RISK CONTROL UNDER DOMAIN SHIFT (RQ3)

Without target-domain calibration, the KQA Pro-trained ERIC repairer is slightly net-negative on both Freebase benchmarks: 2/4/-2 on WebQSP (answer F1 77.17% $\rightarrow$ 76.91%) and 3/5/-2 on CWQ (77.77% $\rightarrow$ 77.38%). Recalibrating the gate on 30% of each target cohort over 1,000 random splits, the test regression rate never exceeds α at any risk level on either benchmark (Table 5; Appendix G). The observed keep-all outcomes are consistent with Corollary 1 ($\alpha = 0.25\%$ on both benchmarks and $\alpha = 0.5\%$ on WebQSP, whose 30% split holds about 358 originally correct questions, versus 777 on CWQ); elsewhere it accepts a few replacements with mean test Net between -1.39 and -0.26. It thus controls regressions out of domain, but cannot create corrections the repairer does not propose. The limiting factor is retrieval coverage: among initially incorrect CWQ questions with reference queries, 601 require a relation that is new relative to P_0, and 595 of these relations are absent from the retrieved schema; for WebQSP the figures are 316 of 326.

3.5 SUPPLEMENTARY ANALYSIS: DOES GROUP-STRUCTURED SUPERVISION HELP?

Natural repair data observe each (question, candidate) pair under a single graph, although the correct action depends on all three (Figure 1b). We test whether supervision that isolates this joint dependence improves on standard fine-tuning. An execution-realizable intervention (ERIC) group combines two questions, two candidate queries, and two synthesized graphs, and is retained only when native execution verifies that flipping any single factor reverses the correct action; no pairwise-additive score can label all eight cells correctly (Proposition 4). ERIC adds a group penalty on the eight cells' `KEEP` probabilities to the same-data SFT objective; Random Groups keeps the cells, exposure, and penalty but assigns cells to groups at random, destroying the parity (construction, objective, and theory in Appendix J).

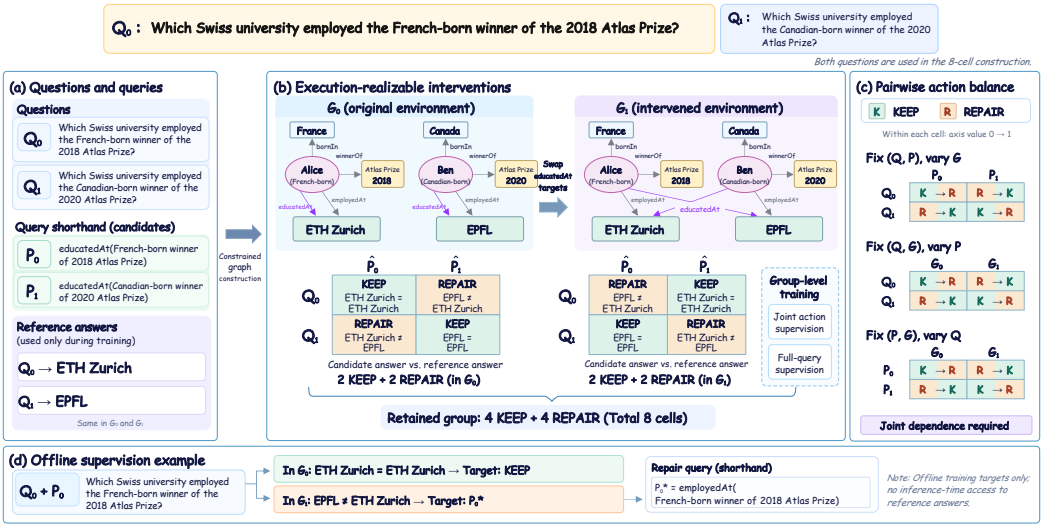

Figure 3: One execution-realizable intervention group. Two questions Q_0, Q_1, two candidate queries C_0, C_1, and two synthesized graphs G_0, G_1 yield eight cells. Native execution verifies the alternating parity pattern: flipping any single axis reverses the action label between KEEP and repair.

Findings. The group penalty gives no measurable benefit over same-data SFT: on the official set ERIC and SFT differ on four questions (90/33/+57 vs. 89/34/+55; 3 wins, 1 loss; $p = 0.63$), on the development cohort on one (80/8/+72 vs. 80/9/+71), and on the report half on none; averaged over three seeds their error-ranking AUROC differs by 0.0003 (development) and 0.0011 (official), and on the fixed proposal pool by one net question per seed (Appendix E). Parity structure matters only relative to random grouping: Random Groups makes six more corrections but 18 more regressions (96/51/+45; ERIC wins 19, loses 7, raw $p = 0.029$, single seed) and is the only trained model whose bound exceeds α (Table 1); an earlier ablation shows the same ordering (Appendix C). On 246 held-out, source-disjoint groups, the diagnostics provide no evidence of learned three-way structure: cell accuracy is at chance (49.2–50.7%), no group is solved in all eight cells, and mean contrast D_g lies between -0.009 and 0.052, far below the pairwise-additive ceiling of 6 (Table 13). Same-data SFT already produces the observed conservative behaviour; a penalty gradient that vanishes on confidently mislabelled cells and a budget of 8 of the 1,085 available groups are further candidate explanations that we do not isolate (Appendix J.5).

4 DISCUSSION

Gate or fine-tune? The two routes reach similar net gains on KQA Pro but need different resources. The gate needs no training, only sequence probabilities from the repairer, but it needs labelled calibration questions from the deployment distribution: at least 598 originally correct ones at $\alpha = 0.5\%$ (Corollary 1), and in practice about a thousand questions for a useful certificate (Figure 2). Fine-tuning needs repair supervision and a training run, and it changes inference behaviour: the same-data SFT model emits an explicit KEEP on 83% of development questions instead of 66% and generates 47% fewer output tokens, although wall-clock generation time in our batched logs fell by only 2.6% (Appendix H). Because fine-tuned models pass the certificate with margin, gating them changes almost no outcomes but adds a guarantee. When labelled target-distribution data are available, the gate is therefore a sensible default, and fine-tuning an optional step that shortens outputs.

When does the gate matter most? For a fixed proposal pool, gating changes Net by the number of avoided regressions minus discarded corrections. A high regression rate leaves room for improvement, but gains also depend on the score distinguishing these outcomes. The data follow this pattern: gating the frozen model (DR 1.14%) changes correctness on 121 official questions and

turns net -18 into $+57$, whereas gating Same-data SFT (DR 0.31%) changes correctness on two questions and leaves Net unchanged (Table 1). Holding conditional rates fixed, larger π increases the weight of regressions in Eq. 2; changes in parser accuracy alone do not determine gating gains.

Where the remaining headroom lies. After gating, the frozen repairer corrects 85 of 713 initially incorrect official questions (11.9%). An earlier generator found a correct candidate among eight samples for only 51 of 191 initially incorrect questions (Appendix I); this does not isolate the current pipeline's bottleneck. Better proposals and retrieval are promising directions, with the gate bounding regression risk.

5 Related Work

Knowledge base question answering. KBQA benchmarks over Freebase and Wikidata (Bollacker et al., 2008; Vrandečić & Krötzsch, 2014) range from simple and multi-hop questions (Berant et al., 2013; Yih et al., 2016; Talmor & Berant, 2018) to compositional and generalization-oriented settings (Gu et al., 2021; Cao et al., 2022). Systems have moved from enumeration, ranking, retrieval, and logical-form generation (Ye et al., 2022; Shu et al., 2022; Yu et al., 2023) to language-model pipelines with in-context learning, structured interfaces, schema guidance, and execution feedback (Li et al., 2023; Jiang et al., 2023; Luo et al., 2024; Huang et al., 2024b; Gao et al., 2025; Tian et al., 2025); see Pan et al. (2024) for a survey. Closest to our setting, RKBQA revises generate-then-retrieve outputs once with a fallback (Huang & Liu, 2025); we take such systems as given and study the acceptance decision itself, with an explicit bound on regressions over the full question population.

Self-correction and verification. Intrinsic self-correction often fails to improve, and can degrade, reasoning outputs (Huang et al., 2024a), while iterative self-feedback helps on some generation tasks (Madaan et al., 2023); tools, retrieval, and environment interaction make correction more reliable (Yao et al., 2023; Schick et al., 2023; Gou et al., 2024; Asai et al., 2024), and correction behaviour can also be trained directly (Kumar et al., 2025). Our repairer uses execution feedback and an operational validator, neither of which sees reference answers; the gate adds what they lack, a population-level bound on how many correct answers accepted edits may break.

Selective prediction and risk control. Selective classification abstains when confidence is low (Geifman & El-Yaniv, 2017), calibration studies ask whether such confidence can be trusted (Guo et al., 2017), and distribution-free risk control selects thresholds on held-out data so that a user-specified risk is bounded with high probability (Bates et al., 2021; Angelopoulos et al., 2021; 2024). Our gate is a simple instance of this framework, a single threshold on the repairer's own `KEEP`-sequence probability selected with an exact binomial bound (Clopper & Pearson, 1934); what is specific to repair is the choice of risk, regressions among initially correct queries, and the finding that the model's own probability suffices to control it. Our supplementary intervention groups relate to contrast sets (Gardner et al., 2020) and to denotation-based supervision in executable semantic parsing (Pasupat & Liang, 2016; Zhong et al., 2020).

6 Conclusion

Conservative query repair separates proposing edits from accepting them. Our training-free gate thresholds the repairer's `KEEP` probability using a Clopper–Pearson bound, controlling population regression risk for a fixed pipeline and i.i.d. calibration data from the deployment distribution. On KQA Pro, it changes the frozen repairer's net gain from -18 to $+57$, comparable to fine-tuning. Fine-tuning increases `KEEP` recall without improving error-ranking AUROC, and the group penalty adds no measurable benefit. On WebQSP and CWQ, target-domain recalibration keeps observed regressions within target but yields no positive net gain, motivating better proposals alongside explicit acceptance control.

LLM USAGE STATEMENT

Apart from the language models studied as repairers (§3.1), we used large language models to assist with writing and checks of mathematical arguments and numerical consistency. All research ideas, methods, experiments, analyses, and conclusions are the authors' own, and the authors take full responsibility for the entire content of the paper.

REPRODUCIBILITY STATEMENT

Section 3.1 specifies cohorts, backbone, fine-tuning hyperparameters, matched exposures, and the locked acceptance thresholds, and §2.2 gives the calibration procedure. All thresholds are selected on the selection half and never re-selected on evaluation data. Appendix A contains complete proofs, Appendix B reports all 72 repeated-split configurations, and Appendices D–I report every additional analysis cited in the main text, including the data-construction statistics.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then Test: Calibrating predictive algorithms to achieve risk control. *arXiv preprint arXiv:2110.01052*, 2021. URL https://arxiv.org/abs/2110.01052.

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=hSyW5go0v8.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1533–1544, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1160/.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, pp. 1247–1250. ACM, 2008. doi: 10.1145/1376616.1376746.

Shulin Cao, Jiaxin Shi, Liangming Pan, Lunyiu Nie, Yutong Xiang, Lei Hou, Juanzi Li, Bin He, and Hanwang Zhang. KQA pro: A dataset with explicit compositional programs for complex question answering over knowledge base. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6101–6119, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.422. URL https://aclanthology.org/2022.acl-long.422/.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.2307/2331986.

Shengxiang Gao, Jey Han Lau, and Jianzhong Qi. Beyond seen data: Improving KBQA generalization through schema-guided logical form generation. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8753–8772, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.442. URL https://aclanthology.org/2025.emnlp-main.442/.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.findings-emnlp.117. URL https://aclanthology.org/2020.findings-emnlp.117/.

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

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. CRITIC: Large language models can self-correct with tool-interactive critiquing. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Sx038qxjek.

Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond I.I.D.: Three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pp. 3477–3488. ACM, 2021. doi: 10.1145/3442381.3449992.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

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

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=IkmD3fKBPQ.

Wenbo Huang and Lihui Liu. RKBQA: Revise-once correction for generate-then-retrieve KBQA. In *2025 IEEE International Conference on Big Data (BigData)*, pp. 679–686, 2025. doi: 10.1109/BigData66926.2025.11401949. URL https://doi.org/10.1109/BigData66926.2025.11401949.

Xiang Huang, Sitao Cheng, Shanshan Huang, Jiayu Shen, Yong Xu, Chaoyun Zhang, and Yuzhong Qu. QueryAgent: A reliable and efficient reasoning framework with environmental feedback based self-correction. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5014–5035, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.274. URL https://aclanthology.org/2024.acl-long.274/.

Jinhao Jiang, Kun Zhou, Zican Dong, Keming Ye, Wayne Xin Zhao, and Ji-Rong Wen. StructGPT: A general framework for large language model to reason over structured data. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9237–9251. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.emnlp-main.574. URL https://aclanthology.org/2023.emnlp-main.574/.

Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, JD Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, Lei M. Zhang, Kay McKinney, Disha Shrivastava, Cosmin Paduraru, George Tucker, Doina Precup, Feryal Behbahani, and Aleksandra Faust. Training language models to self-correct via reinforcement learning. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=CjwERcAU7w.

Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhu Chen. Few-shot in-context learning on knowledge base question answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6966–6980. Association for Computational Linguistics, July 2023. doi: 10.18653/v1/2023.acl-long.385. URL https://aclanthology.org/2023.acl-long.385/.

Haoran Luo, Haihong E, Zichen Tang, Shiyao Peng, Yikai Guo, Wentai Zhang, Chenghao Ma, Guanting Dong, Meina Song, Wei Lin, Yifan Zhu, and Anh Tuan Luu. ChatKBQA: A generate-then-retrieve framework for knowledge base question answering with fine-tuned large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2039–2056, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.122. URL https://aclanthology.org/2024.findings-acl.122/.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=S37hOerQLB.

Shirui Pan, Linhao Luo, Yufei Wang, Chen Chen, Jiapu Wang, and Xindong Wu. Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, 36(7):3580–3599, 2024. doi: 10.1109/TKDE.2024.3352100.

Panupong Pasupat and Percy Liang. Inferring logical forms from denotations. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23–32, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-1003. URL https://aclanthology.org/P16-1003/.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://openreview.net/forum?id=Yacmpz84TH.

Yiheng Shu, Zhiwei Yu, Yuhan Li, Börje F. Karlsson, Tingting Ma, Yuzhong Qu, and Chin-Yew Lin. TIARA: Multi-grained retrieval for robust question answering over large knowledge base. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 8108–8121. Association for Computational Linguistics, December 2022. doi: 10.18653/v1/2022.emnlp-main.555. URL https://aclanthology.org/2022.emnlp-main.555/.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 641–651, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/N18-1059. URL https://aclanthology.org/N18-1059/.

Yuhang Tian, Dandan Song, Zhijing Wu, Pan Yang, Changzhi Zhou, Jun Yang, Hao Wang, Huipeng Ma, Chenhao Li, and Luan Zhang. CompKBQA: Component-wise task decomposition for knowledge base question answering. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 293–309, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.16. URL https://aclanthology.org/2025.emnlp-main.16/.

Denny Vrandečić and Markus Krötzsch. Wikidata: A free collaborative knowledgebase. *Communications of the ACM*, 57(10):78–85, 2014. doi: 10.1145/2629489.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui

Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. RNG-KBQA: Generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6032–6043, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.417. URL https://aclanthology.org/2022.acl-long.417/.

Wen-tau Yih, Matthew Richardson, Chris Meek, Ming-Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–206, Berlin, Germany, August 2016. Association for Computational Linguistics. doi: 10.18653/v1/P16-2033. URL https://aclanthology.org/P16-2033/.

Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Wang, Zhiguo Wang, and Bing Xiang. DecAF: Joint decoding of answers and logical forms for question answering over knowledge bases. In *International Conference on Learning Representations*, 2023. URL https://www.amazon.science/publications/decaf-joint-decoding-of-answers-and-logical-forms-for-question-answering-over-knowledge-bases.

Victor Zhong, Mike Lewis, Sida I. Wang, and Luke Zettlemoyer. Grounded adaptation for zero-shot executable semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6869–6882. Association for Computational Linguistics, November 2020. doi: 10.18653/v1/2020.emnlp-main.558. URL https://aclanthology.org/2020.emnlp-main.558/.

A Proofs for Certified Acceptance

This appendix collects the formal statements behind the certified acceptance gate. Section A.1 proves the accounting identities of §2.1; Section A.2 proves that the acceptance rule of §2.2 controls the regression rate and derives the minimum calibration size; Section A.3 relates κ to greedy decoding. Results on the supplementary intervention groups are in Appendix J.

A.1 Net Accounting and the Acceptance Decomposition

Lemma 1 (Pointwise accounting)**.** *For $b, h \in \{0, 1\}$, $h - b = (1 - b)h - b(1 - h)$. Consequently, on N questions the change in accuracy is $\Delta\mathrm{Acc} = (R - D)/N = \mathrm{Net}/N$.*

Proof. Expanding the right-hand side gives $h - bh - b + bh = h - b$. Summing over questions and using the definitions of R and D in Eq. 1 gives $\sum_i (h_i - b_i) = R - D$, and dividing by N gives the change in accuracy. □

Proposition 1 (Acceptance decomposition)**.** *Assume that a question whose replacement is not accepted keeps its upstream query, so that $V = 0$ implies $h = b$. With π, $\gamma_0, \gamma_1, \eta_0, \eta_1$ as in §2.1, $r = \gamma_0\eta_0$, $d = \gamma_1\eta_1$, and $\mathbb{E}[h - b] = (1 - \pi)\gamma_0\eta_0 - \pi\gamma_1\eta_1$. If $\gamma_0 = 0$ (resp. $\gamma_1 = 0$), the corresponding product is defined as zero.*

Proof. By Lemma 1, $\mathbb{E}[h-b] = \Pr(b{=}0)\Pr(h{=}1 \mid b{=}0) - \Pr(b{=}1)\Pr(h{=}0 \mid b{=}1) = (1-\pi)r - \pi d$. Conditioning on V, $r = \Pr(V{=}1 \mid b{=}0)\Pr(h{=}1 \mid b{=}0, V{=}1) + \Pr(V{=}0 \mid b{=}0)\Pr(h{=}1 \mid b{=}0, V{=}0)$. The last factor is zero because $V = 0$ implies $h = b = 0$, so $r = \gamma_0\eta_0$. The same argument with $b = 1$ gives $d = \gamma_1\eta_1$. When an acceptance rate is zero the conditional probability is undefined, but the corresponding term of r or d is zero, which the convention reproduces. □

The decomposition makes explicit that raising the frequency of edits is not the same as raising the correction rate: expanding a fixed acceptance set cannot decrease either γ_0 or γ_1, and at $\pi \approx 0.92$ the regression term carries more than eleven times the weight of the correction term.

A.2 Validity of the Certified Acceptance Rule

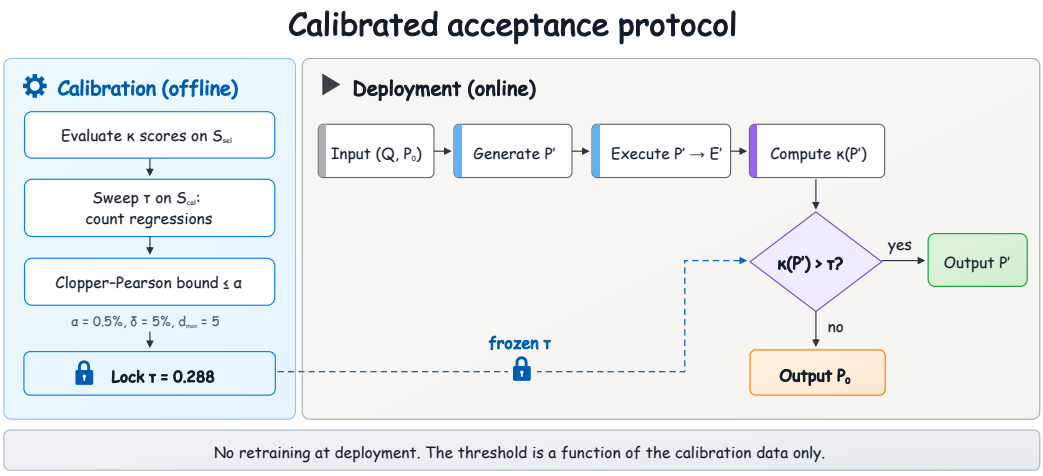

Figure 4: Calibrated acceptance protocol. The threshold τ is chosen on a calibration split so that the Clopper–Pearson upper bound on the regression rate stays below α, then frozen for evaluation.

Setting. Fix the upstream and repair pipelines, including generation, scoring, compilation, and operational validation, before calibration. For an originally correct question, let $\kappa \in [0, 1]$ be the score of its validated replacement ($\kappa = 0$ if there is none) and let $\beta \in \{0, 1\}$ indicate that accepting that replacement would make the answer wrong ($\beta = 0$ if there is none). At threshold τ the question regresses exactly when $\kappa > \tau$ and $\beta = 1$. The population regression rate is $\mathcal{R}(\tau) = \Pr(\kappa > \tau, \beta = 1 \mid b = 1)$. The calibration split contains n_c originally correct questions; conditional on n_c, their pairs (κ_i, β_i) are i.i.d. draws from the same conditional distribution as at deployment. Let $d(\tau) = \sum_{i=1}^{n_c} \mathbb{1}[\kappa_i > \tau, \beta_i = 1]$.

Clopper–Pearson bound. For $k < n$, let $U(k; n, \delta)$ be the unique p with $\Pr[\mathrm{Bin}(n, p) \le k] = \delta$, and let $U(n; n, \delta) = 1$. Two standard facts hold (Clopper & Pearson, 1934). (i) *Validity*: if $K \sim \mathrm{Bin}(n, p)$ then $\Pr[p > U(K; n, \delta)] \le \delta$. (ii) *Monotonicity*: $U(k; n, \delta)$ is nondecreasing in k.

Selection rule. Let $\Lambda \subset \mathbb{R}$ be a finite grid of thresholds fixed before calibration, with $0 < \alpha, \delta < 1$. If $n_c = 0$, define the rule to keep every query; otherwise use the following selection rule. The rule of §2.2 selects

$$\hat{\tau} = \min\{\tau \in \Lambda : \; U(d(\tau); n_c, \delta) \le \alpha, \; d(\tau) \le d_{\max}\}, \tag{3}$$

and keeps every upstream query if the set is empty.

Lemma 2 (Monotonicity in the threshold)**.** *For $\tau_1 \le \tau_2$, $d(\tau_2) \le d(\tau_1)$ on every calibration sample and $\mathcal{R}(\tau_2) \le \mathcal{R}(\tau_1)$.*

Proof. $\mathbb{1}[\kappa > \tau_2] \le \mathbb{1}[\kappa > \tau_1]$ pointwise, so each summand of d and the event defining $\mathcal{R}$ can only shrink as τ grows. □

Proposition 2 (Certified regression rate)**.** *Under the setting above,* $\Pr\big[\mathcal{R}(\hat{\tau}) > \alpha\big] \le \delta$*, where the probability is over the calibration sample and the keep-all outcome is assigned* $\mathcal{R} = 0$*.*

Proof. The claim is immediate for $n_c = 0$, so assume $n_c \ge 1$. Let $B = \{\tau \in \Lambda : \mathcal{R}(\tau) > \alpha\}$. If B is empty there is nothing to prove. Otherwise let $\tau' = \max B$; it depends only on the population, not on the calibration sample. Suppose $\mathcal{R}(\hat{\tau}) > \alpha$. Then the rule did not keep every query, so $\hat{\tau}$ belongs to the feasible set in Eq. 3 and $U(d(\hat{\tau}); n_c, \delta) \le \alpha$; moreover $\hat{\tau} \in B$, hence $\hat{\tau} \le \tau'$. By Lemma 2, $d(\tau') \le d(\hat{\tau})$, and by fact (ii), $U(d(\tau'); n_c, \delta) \le U(d(\hat{\tau}); n_c, \delta) \le \alpha < \mathcal{R}(\tau')$. Thus the event $\{\mathcal{R}(\hat{\tau}) > \alpha\}$ is contained in $\{\mathcal{R}(\tau') > U(d(\tau'); n_c, \delta)\}$. Since τ' is fixed and $d(\tau') \sim \mathrm{Bin}(n_c, \mathcal{R}(\tau'))$, fact (i) bounds the probability of the latter event by δ. □

The argument is the fixed-sequence principle of Learn-then-Test (Angelopoulos et al., 2021) specialized to a monotone risk: because regressions can only decrease as the threshold rises, choosing the most permissive certified threshold requires no multiplicity correction. The constraint $d(\tau) \le d_{\max}$ only shrinks the feasible set and is not used in the proof.

Corollary 1 (Minimum calibration size)**.** *Under the Clopper–Pearson selection rule above, no non-trivial acceptance threshold is feasible unless* $n_c \ge \ln\delta / \ln(1-\alpha)$*. This is a necessary, not sufficient, condition. For* $\delta = 5\%$ *this requires* $n_c \ge 1{,}197$*,* 598*, and* 299 *originally correct calibration questions at* $\alpha = 0.25\%$*,* 0.5%*, and* 1%*.*

Proof. For $n_c \ge 1$, by fact (ii), the smallest attainable bound is $U(0; n_c, \delta)$, which solves $(1-p)^{n_c} = \delta$, i.e. $U(0; n_c, \delta) = 1 - \delta^{1/n_c}$. The condition $1 - \delta^{1/n_c} \le \alpha$ is equivalent to $n_c \ln(1-\alpha) \le \ln\delta$, i.e. $n_c \ge \ln\delta / \ln(1-\alpha)$. If it fails, every threshold is infeasible and the rule keeps every upstream query. Evaluating the bound at $\delta = 0.05$ gives $1{,}196.8$, 597.6, and 298.1, rounded up. □

Remark (What the certificate does and does not say). Proposition 2 bounds the *population* regression rate with probability $1 - \delta$ over the calibration draw. It does not bound the regression rate observed on a finite test split: with about 2,170 originally correct test questions, a rule whose population rate lies just below α exceeds α on the test split with substantial probability, which is why the exceedance frequencies of Figure 2 and Appendix B are not estimates of δ. The guarantee also presumes that calibration and deployment questions come from the same distribution; the KQA Pro official set is a different cohort from the selection half, so its own Clopper–Pearson bound (Table 1) is reported as separate evidence. Finally, the certificate concerns the aggregate rate, not any individual repair.

A.3 KEEP Probability and Greedy Decoding

Proposition 3. *Let* $q = (q_1, \ldots, q_m)$ *be the* `KEEP` *token sequence followed by EOS and* $\kappa = \prod_{l=1}^{m} p_\theta(q_l \mid x, q_{<l})$*. If* $\kappa > 1/2$*, unconstrained greedy decoding on* x *outputs exactly* q*. The converse is false.*

Proof. Each factor lies in $[0, 1]$, so $\kappa > 1/2$ forces every factor to exceed $1/2$. A token with conditional probability above $1/2$ is the unique arg-max at its step, so greedy decoding selects q_l at every step l, including EOS. For the converse, let the first `KEEP` token have probability 0.4 and two other tokens 0.3 each, and let the rest of q have conditional probability one: greedy decoding outputs `KEEP` although $\kappa = 0.4 < 1/2$. □

Consequently, three quantities should be evaluated separately: whether `KEEP` probabilities exceed $1/2$, the action actually decoded, and the correctness of a generated repair. It also explains why acceptance thresholds far below $1/2$, such as the locked SFT threshold $\kappa(P') > 0.0086$, are meaningful.

B Robustness of the Acceptance Rule

For each model, calibration size $n_{\mathrm{cal}} \in \{250, 500, 1{,}000, 2{,}359\}$, target $\alpha \in \{0.25\%, 0.5\%, 1\%\}$, and rule (empirical or Clopper–Pearson certified), we draw 1,000 random calibration/test splits of the 4,719-question development cohort, lock the threshold on the calibration split, and apply it to the remaining questions. Table 6 reports how often the test regression rate exceeds α, and Table 7 reports

mean test Net. Test splits are finite and drawn from one question pool, so exceedance frequencies are not estimates of the probability that the population-level guarantee fails (Appendix A.2). Corollary 1 explains keep-all outcomes when the realized number of originally correct calibration questions is below its necessary bound (Table 3). The empirical rule always accepts some replacements but often exceeds α for the frozen model; fine-tuned models, whose regressions are rarer, seldom exceed α under either rule.

Table 6: Test-split exceedance: percentage of 1,000 random selection/test splits in which the locked rule's test regression rate exceeds α. Emp.: empirical rule; Cert.: Clopper–Pearson certified rule ($\delta = 5\%$, $d_{\max} = 5$). KA: the rule keeps every upstream query in all splits (no replacement is certifiable). Development cohort; the test split is the complement of the calibration split.

		Frozen		SFT		ERIC	
n_{cal}	α (%)	Emp.	Cert.	Emp.	Cert.	Emp.	Cert.
250	0.25	21.8	KA	0.0	KA	0.0	KA
250	0.5	4.4	KA	0.0	KA	0.0	KA
250	1	0.0	KA	0.0	KA	0.0	KA
500	0.25	31.5	KA	0.0	KA	0.0	KA
500	0.5	7.2	KA	0.0	KA	0.0	KA
500	1	0.0	1.9	0.0	0.0	0.0	0.0
1,000	0.25	27.2	KA	0.0	KA	0.0	KA
1,000	0.5	9.1	1.0	0.0	0.0	0.0	0.0
1,000	1	0.0	6.3	0.0	0.0	0.0	0.0
2,359	0.25	19.8	5.3	8.0	7.5	1.6	7.3
2,359	0.5	21.3	10.8	0.0	0.0	0.0	0.0
2,359	1	0.0	10.0	0.0	0.0	0.0	0.0

Table 7: Mean test-split Net ($R - D$) over the same 1,000 splits. Test splits contain 4,469, 4,219, 3,719, and 2,360 questions for the four calibration sizes, so Net is comparable within a row but not across rows (Figure 2 normalizes per 1,000 questions). KA as in Table 6 (Net $= 0$).

		Frozen		SFT		ERIC	
n_{cal}	α (%)	Emp.	Cert.	Emp.	Cert.	Emp.	Cert.
250	0.25	47.1	KA	52.8	KA	53.8	KA
250	0.5	53.3	KA	53.5	KA	54.5	KA
250	1	53.8	KA	53.5	KA	54.5	KA
500	0.25	54.1	KA	56.3	KA	57.8	KA
500	0.5	55.7	KA	56.7	KA	58.0	KA
500	1	55.8	46.3	56.7	60.5	58.0	62.4
1,000	0.25	50.0	KA	53.1	KA	54.4	KA
1,000	0.5	50.6	31.3	53.3	51.2	54.5	53.3
1,000	1	50.7	50.4	53.3	55.9	54.5	56.7
2,359	0.25	33.3	23.1	35.0	32.4	35.6	34.1
2,359	0.5	32.6	33.4	35.0	35.4	35.6	35.9
2,359	1	32.5	31.3	35.0	35.4	35.6	35.9

C HISTORICAL ABLATION

Before the inference pipeline of §2.3 was fixed, we ran the Random Groups ablation on the development cohort under an earlier compilation pipeline (Table 8). These numbers come from a separate run with different checkpoints and are not comparable to Tables 1 or 4; we report them because they show the same ordering.

D ACCEPTANCE SCORES FOR ALL MODELS

Table 9 extends Table 2 to the fine-tuned models (training seed 1). All candidates are the frozen model's proposals; each score is thresholded on the selection half under the protocol of §2.2 and evaluated on the report half (2,360 questions; 190 initially incorrect, 2,170 initially correct). AUROC

Table 8: Earlier ablation under a different compilation pipeline (development cohort, 4,719 questions; 381 initially incorrect).

Method	$R\uparrow$	$D\downarrow$	Net↑	RR (%)	DR (%)↓	Acc. (%)↑
Same-data SFT	47	15	+32	12.34	0.346	92.60
Random Groups	57	27	+30	14.96	0.622	92.56
ERIC	47	14	+33	12.34	0.323	92.63

and AP measure how well a score separates correcting from regressing proposals on the selection half.

Table 9: Alternative acceptance scores (report half).

Scoring model	Score	AUROC	AP	R	D	Net	DR (%)
Frozen	$\kappa(P')$	0.839	0.863	38	4	+34	0.184
Frozen	Mean log-prob.	0.623	0.831	31	4	+27	0.184
Frozen	κ − mean log-prob.	0.842	0.891	38	4	+34	0.184
Frozen	Prompted judge	0.817	0.895	37	6	+31	0.276
SFT (seed 1)	$\kappa(P')$	0.827	0.858	36	2	+34	0.092
SFT (seed 1)	Mean log-prob.	0.780	0.906	37	4	+33	0.184
SFT (seed 1)	κ − mean log-prob.	0.762	0.799	35	2	+33	0.092
SFT (seed 1)	Prompted judge	0.810	0.893	39	6	+33	0.276
ERIC (seed 1)	$\kappa(P')$	0.830	0.859	36	3	+33	0.138
ERIC (seed 1)	Mean log-prob.	0.781	0.907	37	4	+33	0.184
ERIC (seed 1)	κ − mean log-prob.	0.762	0.797	35	2	+33	0.092
ERIC (seed 1)	Prompted judge	0.811	0.893	39	6	+33	0.276

E GATING A FIXED PROPOSAL POOL

To separate the quality of the acceptance score from the quality of the proposals, we fix the candidate pool to the frozen model's 150 report-half proposals and filter it with each model's own κ and its selection-locked threshold. All seven scoring models reach Net between $+33$ and $+35$ (Table 10); within each seed, ERIC and SFT differ on at most three questions (Table 11).

Table 10: Fixed proposal pool, report half (2,360 questions).

Scoring model	R	D	Net	DR (%)	Acc. (%)
Frozen	38	4	+34	0.184	93.39
SFT (seed 1)	36	2	+34	0.092	93.39
ERIC (seed 1)	36	3	+33	0.138	93.35
SFT (seed 2)	36	1	+35	0.046	93.43
ERIC (seed 2)	36	2	+34	0.092	93.39
SFT (seed 3)	36	2	+34	0.092	93.39
ERIC (seed 3)	34	1	+33	0.046	93.35

F ERROR RANKING BY THE KEEP PROBABILITY

Each upstream query is scored by $-\log \kappa(P_0)$, with originally incorrect queries as positives (8.07% of the development cohort and 6.04% of the official set). Fine-tuning leaves the ranking essentially unchanged (Table 12). Per-seed ERIC−SFT differences have question-level bootstrap intervals that exclude zero in some seeds but change sign across seeds.

Table 11: Paired comparison of ERIC against SFT on the fixed pool (exact McNemar; Holm-adjusted $p = 1$ for all seeds).

Seed	ERIC wins	ERIC losses	p
Seed 1	0	1	1
Seed 2	0	1	1
Seed 3	1	2	1

Table 12: Error ranking by $-\log \kappa(P_0)$. Seeds 1–3 are the three training seeds of each arm.

	Development (4,719)		Official (11,797)	
Model	AUROC	AP	AUROC	AP
Frozen	0.872	0.480	0.864	0.444
SFT (seed 1)	0.870	0.511	0.861	0.460
ERIC (seed 1)	0.870	0.512	0.861	0.462
SFT (seed 2)	0.867	0.470	0.859	0.436
ERIC (seed 2)	0.868	0.494	0.859	0.449
SFT (seed 3)	0.867	0.496	0.855	0.446
ERIC (seed 3)	0.865	0.488	0.852	0.438
SFT (mean±SD)	0.868±0.001	0.492±0.020	0.858±0.003	0.447±0.012
ERIC (mean±SD)	0.868±0.003	0.498±0.012	0.857±0.005	0.450±0.012

G TARGET-DOMAIN CALIBRATION

For WebQSP and CWQ we lock the certified threshold on a random 30% of the labelled cohort and evaluate on the remaining 70%, repeated over 1,000 splits. This uses target-domain labels and is therefore not a zero-shot result. Table 5 reports means over splits, so Net can be fractional.

H DATA CONSTRUCTION AND INFERENCE COST

Intervention groups. Group synthesis starts from 1,513 originally incorrect source questions, of which 790 are eligible for synthesis. After 1,553 synthesis attempts, 732 sources (48.4% of all incorrect seeds) yield at least one group, for 1,331 execution-verified groups (10,648 cells). Of these, 1,085 are available for training and 246 are held out, source-disjoint, for the diagnostics of Table 13. The natural repair pool contains 18,727 unique examples (1,365 originally incorrect, 17,362 originally correct).

Table 13: Held-out intervention diagnostics on 246 source-disjoint groups (1,968 cells). All-eight: groups with every cell correct. $D_g > 6$: groups beyond the pairwise-additive bound (Proposition 6).

Method	Cell Acc. (%)	All-eight	Mean D_g	D_g>6
Same-data SFT	49.39	0/246	0.003	0/246
Random Groups	50.66	0/246	0.052	0/246
ERIC	49.24	0/246	−0.009	0/246

Training exposure. For each of the three seeds, SFT and ERIC receive 8 parameter updates with matched exposure: 64 natural examples and 8 complete groups (64 cells). Exposure counts are not unique-source counts, and the available pools are much larger than what is consumed.

Inference cost. On the development cohort, the frozen model emits an explicit `KEEP` for 3,130 of 4,719 questions (66.3%) and same-data SFT for 3,931 (83.3%). Output tokens fall from 162,390 to 85,972 (−47.1%), but historical generation wall-clock time falls only from 7,351 s to 7,164 s (−2.6%); replaying the actual batches, estimated padded decode slots fall from 507,964 to 363,138 (−28.5%). These are estimates from generation logs, not end-to-end latency measurements. Scoring $\kappa(P')$ happens after generation and does not reduce generation cost.

I PROPOSAL COVERAGE

In an earlier data version, we sampled eight candidates for each of the 191 initially incorrect selection-half questions (1,528 candidates) with an execution-verified minimal-edit generator that differs from the repairer evaluated in the main text. Raw pass@8 coverage was 47 questions (24.6%), rising to 51 (26.7%) after the operational pipeline; compilation recovered four questions and no question with a correct raw candidate was lost to the acceptance checks. Of the remaining questions, 107 had executable candidates none of which was correct, 32 received `KEEP` in all eight samples, and one had no successfully executing candidate. Because this version fell short of a preset 40% coverage threshold, we did not train on it; we report it as a coverage limitation of that earlier pipeline, without isolating the current pipeline's bottleneck.

J EXECUTION-REALIZABLE INTERVENTION GROUPS

This appendix describes the supplementary intervention groups of §3.5: their construction, the group penalty used by ERIC, and the theory behind both.

J.1 CONSTRUCTION

The correct action is a function of all three inputs (Q, P_0, G), but natural repair data observes each (question, candidate) pair under a single graph and so never isolates their joint effect. We therefore construct groups in which every single-factor change is execution-verified to flip the label (Figure 3).

Each group contains two questions Q_i, two candidate queries C_j, and two finite execution graphs G_k $(i, j, k \in \{0, 1\})$. A reference program $P_i^\star$ specifies the intended answer for Q_i:

$$a_{ik}^\star = A_{G_k}(P_i^\star), \qquad h_{ijk} = \mathrm{Hit}\big(A_{G_k}(C_j),\, a_{ik}^\star\big). \tag{4}$$

We synthesize finite graphs and retain a group only when native execution verifies

$$H_0 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad H_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \tag{5}$$

so that $s_{ijk} = 2h_{ijk} - 1 = (-1)^{i+j+k}$: fixing any two axes and varying the third flips the label. The target is $y_{ijk} = \texttt{KEEP}$ when $h_{ijk} = 1$, and $P_i^\star$ otherwise; each group has four `KEEP` and four repair cells.

Proposition 4 (Pairwise-additive separation)**.** *For any $f_{ijk} = u_{ij}+v_{ik}+w_{jk}+c$ and any threshold τ, $\sum_{i,j,k} s_{ijk}\, f_{ijk} = 0$. No fixed threshold classifies all eight cells correctly.*

Proof. $\sum_{i,j,k}(-1)^{i+j+k}u_{ij} = \sum_{i,j}(-1)^{i+j}u_{ij}\sum_k(-1)^k = 0$; the v_{ik}, w_{jk}, and constant terms vanish analogously. Perfect classification would require every $s_{ijk}(f_{ijk}-\tau) > 0$, contradicting a zero sum. □

The tight accuracy bound is $7/8$, a rule that sees only two of the three factors is exactly at chance, and any pairwise-additive logit incurs at least chance-level logistic loss on every group (Appendix J.3). For a pairwise-additive action logit z, the bounded contrast $D_g = \sum s_{ijk}\tanh(z_{ijk}/2)$ satisfies $\sup D_g = 6$, compared to 8 for unrestricted joint logits (Proposition 6). These results characterize a function class; they do not imply that SFT-trained language models are limited to pairwise-additive representations.

J.2 GROUP PENALTY AND TRAINING OBJECTIVE

With $\kappa_\theta(x)$ from §2.2 and its logit $z_\theta(x) = \log[\kappa_\theta/(1-\kappa_\theta)]$, we define within each group g a penalty on probability errors:

$$E_g = \textstyle\sum_{i,j,k} |\kappa_\theta(x_{ijk}) - h_{ijk}|, \qquad J_g = \mathrm{softplus}(2E_g - 4). \tag{6}$$

Since $D_g = 8 - 2E_g$, this equals $\mathrm{softplus}(4 - D_g)$ (Lemma 6; gradient and curvature in Appendix J.5). The full objective is

$$\mathcal{L} = \underbrace{\frac{1}{|\mathcal{B}|}\sum_{(x,y)\in\mathcal{B}} \ell_{\mathrm{gen}}}_{\text{natural}} + \underbrace{\frac{\beta}{8|\mathcal{G}|}\sum_{g}\sum_{i,j,k} \ell_{\mathrm{gen}}(x_{ijk}, y_{ijk})}_{\text{group cells}} + \underbrace{\frac{\lambda}{|\mathcal{G}|}\sum_{g} J_g}_{\text{group penalty}}, \tag{7}$$

with ℓ_{gen} as in §2.3, $\beta = 1$, and $\lambda = 0.1$; Same-data SFT uses the same generation supervision with $\lambda = 0$, and Random Groups uses $\lambda = 0.1$ with cells assigned to groups at random.

J.3 Parity Groups and Pairwise-Additive Scores

Throughout, $s_{ijk} = (-1)^{i+j+k}$, and a score f labels cell (i, j, k) correctly at threshold τ when $s_{ijk}(f_{ijk} - \tau) > 0$.

Lemma 3 (Two observed factors are uninformative)**.** *Let a rule observe only two of the three factors of a cell, say (i, j), and receive no other information about the third. Under uniform weight on the eight cells, its expected accuracy is exactly $1/2$, whether it is deterministic or randomized.*

Proof. For fixed (i, j) the two cells $k = 0, 1$ carry opposite labels, $s_{ij0} = -s_{ij1}$, and receive the same prediction distribution. Whatever label is predicted, exactly one of the two cells is correct, so the conditional accuracy given (i, j) is $1/2$; averaging over (i, j) gives $1/2$. The same argument applies to the pairs (i, k) and (j, k). □

Lemma 3 concerns rules that see two factors only. A pairwise-additive score $f_{ijk} = u_{ij} + v_{ik} + w_{jk} + c$ combines all three pairs and is strictly more expressive; Proposition 4 and the results below characterize it.

Lemma 4 (Tie handling)**.** *Proposition 4 continues to hold when ties $f_{ijk} = \tau$ are resolved by a fixed rule in either direction.*

Proof. Suppose ties are labelled `KEEP`. Correct labelling then requires $f_{ijk} - \tau \geq 0$ on the four cells with $s_{ijk} = +1$ and $f_{ijk} - \tau < 0$ on the four cells with $s_{ijk} = -1$, so every term of $\sum s_{ijk}(f_{ijk} - \tau)$ is nonnegative and four are strictly positive. The sum is therefore strictly positive, contradicting $\sum s_{ijk} f_{ijk} = 0$ and $\sum s_{ijk} = 0$. Labelling ties as repair is symmetric. □

Proposition 5 (Tight accuracy bound)**.** *Every pairwise-additive score labels at most seven of the eight cells correctly, and seven is attained.*

Proof. The upper bound is Proposition 4 together with Lemma 4. For attainment, take $f_{ijk} = -7+6(i+j+k)-4(ij+ik+jk)$, which is pairwise additive, and $\tau = 0$. For $i+j+k = 0, 1, 2, 3$ the value is $-7, -7+6 = -1, -7+12-4 = 1$, and $-7+18-12 = -1$, while $s_{ijk} = +1, -1, +1, -1$. Only cell $(0, 0, 0)$ is misclassified. □

Corollary 2 (Logistic loss of pairwise-additive logits)**.** *For any pairwise-additive logit table f, the average binary logistic loss over a group satisfies $\frac{1}{8}\sum_{i,j,k} \mathrm{softplus}(-s_{ijk} f_{ijk}) \geq \log 2$, with equality at $f \equiv 0$.*

Proof. softplus is convex, so by Jensen's inequality the average is at least $\mathrm{softplus}\big(-\frac{1}{8}\sum s_{ijk} f_{ijk}\big) = \mathrm{softplus}(0) = \log 2$, using Proposition 4. □

A pairwise-additive logit therefore cannot beat a constant predictor in logistic loss on any single group. This binary loss is a diagnostic of the action decision; it is not the full query-generation loss used in training.

J.4 The Bounded Contrast

Lemma 5 (Two properties of tanh)**.** *(a)* tanh *is concave on* $[0,\infty)$*, so for* $x_1,\dots,x_n \geq 0$*,* $\sum_l \tanh(x_l) \leq n\tanh\big(\frac{1}{n}\sum_l x_l\big)$*. (b) For* $a,b \geq 0$*,* $\tanh(a)+\tanh(b) \geq \tanh(a+b)$*, and by induction* $\sum_l \tanh(x_l) \geq \tanh(\sum_l x_l)$ *for* $x_l \geq 0$*.*

Proof. (a) $\tanh''(x) = -2\tanh(x)\operatorname{sech}^2(x) \leq 0$ for $x \geq 0$; the inequality is Jensen's. (b) $\tanh(a+b) = \frac{\tanh a+\tanh b}{1+\tanh a\tanh b} \leq \tanh a + \tanh b$ because the denominator is at least one when $a,b \geq 0$. □

Proposition 6 (Bounded contrast)**.** *Let* $D_g = \sum_{i,j,k} s_{ijk}\tanh(z_{ijk}/2)$*. If* z *is pairwise additive, then* $D_g \leq 6$*, and* $\sup D_g = 6$*. For unrestricted* z*,* $\sup D_g = 8$*.*

Proof. Index cells by u and let $t_u = s_u z_u$. Since tanh is odd and $s_u \in \{\pm 1\}$, $s_u\tanh(z_u/2) = \tanh(t_u/2)$, so $D_g = \sum_u \tanh(t_u/2)$. By Proposition 4, $\sum_u t_u = 0$. If every $t_u = 0$ then $D_g = 0 \leq 6$. Otherwise some t_u are positive and some negative. Let P be the set of cells with $t_u > 0$, $n = |P| \in \{1,\dots,7\}$, and $T = \sum_{u\in P} t_u > 0$; the remaining cells satisfy $\sum_{u\notin P}|t_u| = T$. By Lemma 5(a), $\sum_{u\in P}\tanh(t_u/2) \leq n\tanh(T/2n)$. By Lemma 5(b), $\sum_{u\notin P}\tanh(|t_u|/2) \geq \tanh(T/2)$, so $\sum_{u\notin P}\tanh(t_u/2) \leq -\tanh(T/2)$. Hence $D_g \leq \phi_n(T) = n\tanh(T/2n) - \tanh(T/2)$. Its derivative is $\phi_n'(T) = \frac{1}{2}\big[\operatorname{sech}^2(T/2n) - \operatorname{sech}^2(T/2)\big] \geq 0$, because $T/2n \leq T/2$ and sech^2 is decreasing on $[0,\infty)$. So $\phi_n(T) \leq \lim_{T\to\infty}\phi_n(T) = n-1 \leq 6$, and $D_g \leq 6$.

For the supremum, let $z = af$ with f from Proposition 5 and $a > 0$. On the seven correctly labelled cells $s_u f_u = 1$, and on cell $(0,0,0)$ $s_u f_u = -7$, so $t_u = a$ seven times and $-7a$ once. Then $D_g = 7\tanh(a/2) - \tanh(7a/2) \to 6$ as $a \to \infty$. For unrestricted z, the choice $z_u = as_u$ gives $t_u = a$ on all cells and $D_g = 8\tanh(a/2) \to 8$; since each term is below one, $D_g < 8$ always. □

J.5 Probability Errors and the Group Penalty

Let $\kappa_u = \sigma(z_u)$, $h_u = (s_u+1)/2$, and $e_u = |\kappa_u - h_u|$, so that $E_g = \sum_u e_u$.

Lemma 6 (Contrast and probability error)**.** *For every cell,* $s_u\tanh(z_u/2) = 1-2e_u$*. Hence* $D_g = 8-2E_g$ *and* $J_g = \operatorname{softplus}(2E_g-4) = \operatorname{softplus}(4-D_g)$*.*

Proof. $\tanh(z/2) = 2\sigma(z)-1$, so $s_u\tanh(z_u/2) = s_u(2\kappa_u-1)$. If $h_u = 1$ this is $2\kappa_u - 1 = 1-2(1-\kappa_u) = 1-2e_u$; if $h_u = 0$ it is $1-2\kappa_u = 1-2e_u$. Summing over the eight cells gives $D_g = 8-2E_g$. □

Proposition 7 (Gradient, curvature, and sign errors)**.** *(a)* $\partial J_g/\partial z_u = -2\,\sigma(4-D_g)\,s_u\,\kappa_u(1-\kappa_u)$*. (b) In the error coordinates* $e = (e_u)_u$*,* $\nabla_e^2 J_g = 4\,\sigma(2E_g-4)\big(1-\sigma(2E_g-4)\big)\mathbf{1}\mathbf{1}^\top$*, a rank-one matrix. (c) The number* M_g *of cells whose probability is on the wrong side of* $1/2$ *(or equal to it) satisfies* $M_g \leq 2E_g = 8-D_g$*; hence* $D_g > 7$ *implies that every cell is classified correctly, whereas* $D_g > 6$ *does not.*

Proof. (a) $\frac{d}{dz}\tanh(z/2) = \frac{1}{2}\big(1-\tanh^2(z/2)\big) = \frac{1}{2}\big(1-(2\kappa-1)^2\big) = 2\kappa(1-\kappa)$, and $\frac{d}{dD}\operatorname{softplus}(4-D) = -\sigma(4-D)$; the chain rule gives the result. (b) $\partial J_g/\partial e_u = 2\sigma(2E_g-4)$ for every u, and differentiating again with respect to e_v gives $4\sigma'(2E_g-4) = 4\sigma(2E_g-4)(1-\sigma(2E_g-4))$ for every pair (u,v). (c) A cell is misclassified exactly when $e_u \geq 1/2$, so $M_g \leq \sum_u 2e_u = 2E_g$. If $D_g > 7$ then $M_g < 1$. For $D_g > 6$, take one cell with $e_u = 0.6$ and seven with $e_u = 0.05$: $E_g = 0.95$, $D_g = 6.1$, and one cell is misclassified. □

Part (a) shows that the penalty's gradient carries the factor $\kappa_u(1-\kappa_u)$, which vanishes both on confidently correct cells and on confidently *mislabelled* ones; the generation terms of Eq. 7 are unaffected. Part (b) shows that the penalty couples the eight cells only through their total error. The softplus transition of J_g is centered at $D_g = 4$, below the pairwise-additive ceiling of 6, so the penalty does not specifically reward crossing that ceiling; a larger margin would change the optimization pressure but would guarantee neither crossing it nor gains in natural repair. Nothing here implies that the overall network objective is convex.

J.6 Scope of the Execution Certificate

For every retained group, native execution replays $h_{ijk} = \mathrm{Hit}(A_{G_k}(C_j), A_{G_k}(P_i^\star))$ and confirms Eq. 5. This certifies the action labels relative to the reference programs; it does not establish that the natural-language question and the program are equivalent in every environment, and finite-domain synthesis does not guarantee that a group exists for every query. Because the repair target $P_i^\star$ depends only on the question axis, full content supervision is present, but the three-way structure resides in the action label rather than in the content to be generated; action-level structure does not by itself imply that content generation has learned the corresponding interaction.