PARTICLE-KGRAG: ANSWER-ALIGNED CONFIDENCE FOR KNOWLEDGE-GRAPH QUESTION ANSWERING

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge-graph question answering requires confidence scores that distinguish correct answers from retrieval failures, yet a score used to guide graph search need not predict answer correctness. We introduce Particle-KGRAG, a training-free construction of answer-aligned confidence. It samples independent paths from a Personalized-PageRank-guided graph prior, scores completed paths with a terminal LLM judge, and aggregates normalized weights by extracted answer. Confidence is the posterior mass assigned to the answer actually returned. Using the path prior as the proposal gives a self-normalized importance-sampling formulation with a fixed budget of 129 LLM calls per query. Across six benchmarks and three seeds, with post-registration analysis changes disclosed, mean correctness AUROC is 70.6 versus 63.7 for a ToG-style beam's search-derived score; improvements on HotpotQA, 2WikiMultiHopQA, and CWQ survive paired and query-clustered testing with false-discovery-rate correction. A separately pre-registered rerun correcting prompt rendering gives 66.4 versus 58.7 and preserves these three improvements. On archived 2WikiMultiHopQA runs, selective accuracy at 50% coverage increases from .337 to .573, while the full-coverage accuracy difference is not statistically established. A beam control with answer-level aggregation reduces the AUROC gap on that dataset to 1.3 points, leaving the contribution of independent sampling unresolved. There is no demonstrated answer-accuracy gain or corrected superiority over a perturbation-based uncertainty baseline, and MetaQA-3hop favors the beam. These results characterize the utility and limits of answer-aligned confidence from judge-weighted KG paths.

1 INTRODUCTION

Graph-based question answering must support two decisions: which answer to return and how much to trust it. LLM-guided search (Sun et al., 2024; Ma et al., 2025) produces scores for expanding, pruning, and stopping on reasoning paths, but these decisions do not directly evaluate the correctness of the final answer. A confident search trajectory can still produce an incorrect answer (on 2WikiMultiHopQA, a ToG-style beam's search-derived score ranks the correctness of its own answers below chance, AUROC 46.4; §5.1), while several distinct paths can support the same answer. This motivates constructing confidence at the level of the answer whose correctness will be evaluated.

Particle-KGRAG constructs this answer-level score from independently sampled KG paths. A Personalized-PageRank (PPR)-guided prior proposes the paths, and a terminal judge assigns a weight to each completed path without observing its extracted candidate answer. We normalize these weights and sum the mass of paths that yield the same answer. The system returns the answer of its highest-weight particle and reports the aggregated mass assigned to that answer. This construction instantiates self-normalized importance sampling (SNIS) over KG paths without task-specific training or answer-label supervision at inference time.

We evaluate whether this score ranks answer correctness and supports selective prediction across six KGQA benchmarks. We compare it with a ToG-style beam's search-derived confidence, perturbation-based uncertainty, and beam variants with answer-level aggregation. The distinction between these comparisons is central: an improvement over a search score does not by itself establish a benefit from independent path sampling. We therefore report both the confidence gains and

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.

the controls that restrict their interpretation, alongside answer accuracy, a prompt-rendering rerun, and the disclosed changes to the registered analysis.

Contributions.

1. **Answer-aligned confidence for KGQA.** We instantiate self-normalized importance sampling over PPR-guided graph paths and aggregate terminal judge weights by extracted answer. The mass assigned to the highest-weight particle's emitted answer supplies a training-free confidence score, explicitly linking path evidence to the answer being evaluated (§3).

2. **Evidence on correctness ranking and selective prediction.** Across six benchmarks and three seeds, the emitted-answer score improves AUROC over the evaluated beam score on HotpotQA, 2WikiMultiHopQA (2WikiMHQA), and CWQ under paired and query-clustered tests with Benjamini–Hochberg (BH) false-discovery-rate (FDR) correction. On archived 2WikiMHQA runs, selective accuracy at 50% coverage rises from .337 to .573, without an established full-coverage accuracy difference (§5.1–5.2).

3. **Controls that delimit the empirical contribution.** Comparisons with beam-side answer aggregation, fixed-answer confidence controls, and component ablations distinguish improvements over the search-derived score from conclusions about the sampler. Against a beam with answer-level aggregation (a separate post-hoc comparison family), the advantage is established on HotpotQA, MuSiQue, and CWQ, a different triple from the search-score comparison, while the 2WikiMHQA gap narrows to 1.3 AUROC points (§5.3, §5.5).

2 Related Work

Confidence for KG question answering connects graph search with sample-based uncertainty estimation. ToG (Sun et al., 2024) uses LLM-guided beam search to retrieve reasoning paths. Our study examines how evidence from completed paths can support confidence in the returned answer.

Confidence-weighted self-consistency (Taubenfeld et al., 2025) reweights sampled responses using model-reported confidence, while Emmenegger et al. (2026) pair posterior sampling of language-model outputs with a normalizing-constant estimate and conformal abstention. We treat weighted aggregation and posterior sampling as established ingredients. Particle-KGRAG instantiates them over independently sampled KG paths: a terminal judge scores each question–path pair without conditioning on the extracted candidate answer, and normalized weights are aggregated by answer identity. The confidence score is the mass assigned to the emitted answer.

UAG (Ni et al., 2025) and CPR (Lin et al., 2026) provide close KG-specific precedents for conformal reasoning with trained components (neither is run here, since both require training). CPR shares our diagnosis (naive calibration on path scores breaks exchangeability) but repairs it by training a residual value network and emits prediction sets; we stay training-free, calibrate one sampler statistic, the terminal effective sample size (ESS), in a separate gate that controls false abstention rather than errors among retained answers, and return point answers with abstention. EPERM (Long et al., 2025a) weights evidence paths with an LLM-based importance scorer distilled from answer-supervised targets, and RAR (Shen et al., 2026) learns latent reasoning-path alignment with answer-conditioned expectation-maximization; neither is training-free. RoG (Luo et al., 2024) is fine-tuned on WebQSP/CWQ train splits, so not directly comparable to a training-free method. HippoRAG2 (Gutiérrez et al., 2025) uses PPR to retrieve passages and defines no path posterior. R^2-KG (Jo et al., 2025) abstains inside a dual-agent iteration: a supervisor LLM judges evidence sufficiency under a hard iteration cap. R^2C (Soudani et al., 2026) estimates uncertainty through perturbations of retrieval-augmented reasoning, and an axiomatic analysis (Soudani et al., 2025) of why uncertainty quantification (UQ) falls short under retrieval motivates confidence upstream of the answer. COIN (Wang et al., 2026) gives provable risk control for selective QA on text, without graphs. Our contribution is the construction and controlled evaluation of answer-aligned confidence from terminally weighted KG paths, with beam aggregation and fixed-answer comparisons delimiting the conclusions about its components. Further related work (sampling-based uncertainty, KG tree search, and particle methods for LM inference) is in Appendix H.2.

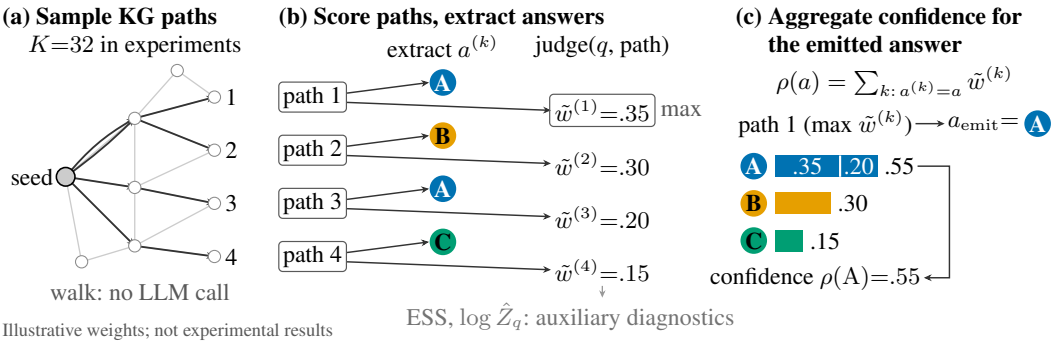


Figure 1: From graph paths to confidence in the returned answer. Independent PPR-guided paths receive terminal judge weights and per-path extracted answers. We aggregate normalized weights by answer identity and evaluate the resulting mass at the answer emitted by the highest-weight particle. In the illustrative example, path 1 has the highest weight, the returned answer is A, and its confidence is .35 + .20 = .55. The emitted answer need not be the mode of the aggregated distribution. The judge does not observe the extracted candidate answer. Example weights are schematic, not experimental results; the experiments use K=32 and 129 LLM calls per query. ESS and the normalizing-constant estimate are auxiliary diagnostics.

3 METHOD

Figure 1 summarizes the construction: sample paths, weight them at termination, and read confidence at the returned answer.

3.1 PATH SAMPLING AND TERMINAL WEIGHTING

Given a query q, we define a posterior over multi-hop reasoning paths $s_{1:T} = (s_1, \ldots, s_T)$ on the KG, each starting at a seed entity s_0:

$$p(s_{1:T} \mid q) \;\propto\; p_0(s_{1:T})\, \phi(s_{1:T}, q), \tag{1}$$

where p_0 is a PPR-guided path prior and ϕ a terminal LLM-judge potential. We draw K=32 independent paths $\{s_{1:T}^{(k)}\}_{k=1}^{K}$ from the prior itself ($q_{\text{prop}} \equiv p_0$), so each path's importance weight collapses to the terminal potential:

$$w^{(k)} \;=\; \frac{p_0(s_{1:T}^{(k)})\, \phi(s_{1:T}^{(k)}, q)}{q_{\text{prop}}(s_{1:T}^{(k)})} \;=\; \phi(s_{1:T}^{(k)}, q), \qquad \tilde{w}^{(k)} \;=\; \frac{w^{(k)}}{\sum_j w^{(j)}}, \tag{2}$$

which is self-normalized importance sampling (SNIS). "Posterior" is operational throughout: the normalized distribution induced by Eq. (1); we do not claim ϕ is a calibrated likelihood.

Path prior. Paths start at the query's seed entities. With teleport vector $\mathbf{e}_q$ (uniform over the query's seed entities), the Personalized-PageRank vector $\boldsymbol{\pi}_q$ solves

$$\boldsymbol{\pi}_q = \lambda\, \mathbf{e}_q + (1 - \lambda)\, \mathbf{A}\mathbf{D}^{-1}\boldsymbol{\pi}_q, \qquad \lambda = 0.15,$$

computed by power iteration ($\mathbf{A}$: undirected KG adjacency; $\mathbf{D}$: degree diagonal). At each step the proposal over the current vertex's unvisited neighbors is proportional to PPR mass, $q_{\text{prop}}(s_t \mid s_{1:t-1}) \propto \boldsymbol{\pi}_q[s_t] \cdot \mathbf{1}[s_t \in \mathcal{N}(s_{t-1}) \setminus s_{1:t-1}]$, up to the pre-registered depth cap of 6 unless dead-ended. The graph walk issues no LLM call. Because the proposal is the prior p_0, replacing PPR by uniform walks changes the target $p_0\phi$ as well; the ablation of §5.5 compares two path priors on their own answers, not two proposals for one target.

Terminal judge. Each completed path is scored only at termination: an LLM judge sees the question and the path and rates whether the path answers q; it never sees the answer later extracted from the path (§3.2), so an answerable path can still be mis-extracted. Across datasets, 6.0% to 44.4% of particles whose path contains a creditable entity extract an uncredited answer, and the terminal

judge score ranks particle-answer correctness at AUROC 50.4 to 59.0 (post-hoc, archived prompts; Appendix G.3). The rating is a first-token log-likelihood ratio (LLR, "Yes" vs. "No") for each of three query paraphrases, and we take the median, $\log \phi^{(k)} = \mathrm{median}_i \ \mathrm{LLR}(s_{1:T}^{(k)}, q^{(i)})$, to damp single-prompt variance.

3.2 Answer Aggregation and Emitted-Answer Confidence

From each particle we extract an answer $a^{(k)}$ and aggregate the normalized weights by answer (all quantities are per query):

$$\rho(a) = \sum_{k:\, a^{(k)}=a} \tilde{w}^{(k)}, \qquad k^\star = \arg\max_k \tilde{w}^{(k)}, \qquad a_{\mathrm{emit}} = a^{(k^\star)}, \qquad c(q) = \rho(a_{\mathrm{emit}}). \tag{3}$$

The equations describe the ordinary nonempty-answer case; empty-answer and failure-handling rules follow Appendix C. The returned answer is selected by the highest-weight particle; it need not be the mode of the aggregated answer distribution. Confidence is evaluated at that returned answer. Answers are grouped using the canonicalization rule in Appendix C. The posterior is operationally defined by the path prior and judge potential, and is not assumed to be calibrated for answer correctness. The emitted-answer mass was adopted after protocol registration: the registration scored the posterior mode, $\max_a \rho(a)$ (registered variants of every headline table: Appendix B, D6), and the emitted answer and the posterior mode coincide on 81.5% of answered queries. Under common key scoring, emitting the posterior mode instead changes Hit@1 by -1.6 to $+1.2$ points across datasets (-6.3 to $+1.0$ under archived-string scoring; post-hoc, archived prompts; Appendix F); against the beam, the mode rule's AUROC advantage is established on two of six datasets and reversed on zero.

3.3 Budget and Implementation Disclosure

For $K{=}32$ paths, the budget is one batched paraphrase-generation call, $32 \times 3 = 96$ terminal judge calls, and 32 answer extractions, totaling 129 LLM calls per query, fixed by construction; meters count the 128 scoring+extraction calls (paraphrase cached across seeds). This is a fixed call count, not a claim of lower monetary cost or latency; the ToG-style beam issues 145–218 calls per query (deduplicated) at the evaluated width, and a narrower beam issues fewer, so the comparison is width-specific (measured cost and latency: Appendix H.6). The archived run used an unrendered-template fallback that appended the filled fields. Appendix C specifies the as-sent prompts and verdict-token imputation. We retain these results as primary and separately report a rerun with corrected prompt rendering; the rerun also changes the beam's final extraction input to its top path (Appendix K).

3.4 Auxiliary Diagnostics

We additionally report the normalizing-constant estimate $\log \hat{Z}_q$ and an abstention gate on the terminal effective sample size, $\mathrm{ESS} = (\sum_k \tilde{w}^{(k)})^2 / \sum_k (\tilde{w}^{(k)})^2$, as auxiliary diagnostics. The normalizing constant (evidence) $Z_q = \mathbb{E}_{p_0}[\phi]$ has the unbiased importance-sampling estimator $\hat{Z}_q = \frac{1}{K}\sum_k \phi^{(k)}$, computed stably as $\log \hat{Z}_q = \mathrm{logsumexp}_k(\log \phi^{(k)}) - \log K$. By Jensen's inequality $\mathbb{E}[\log \hat{Z}_q] \le \log Z_q$, downward-biased in expectation; we report it as an auxiliary evidence signal, not a calibrated probability. The gate is calibrated by split-conformal prediction: on a held-out split we take the α-quantile of ESS among correctly answered queries as the abstention threshold τ (exact quantile index, tie rule, small-n_c rule and the certified unit: Appendix C). Validity assumes (ESS, correctness) pairs exchangeable across queries within a benchmark; conditioning on the correct subset makes this a label-conditional (one-class) split-conformal rule, checked over 30 random splits (Appendix H). Under exchangeability, the gate controls false abstention, $\mathbb{P}(\text{abstain} \mid \text{correct}) = \mathbb{P}(\mathrm{ESS} < \tau \mid \text{correct}) \le \alpha$, i.e., at most an α-fraction of would-be-correct answers is discarded (Vovk et al., 2005; Angelopoulos & Bates, 2021). It does not bound $\mathbb{P}(\text{correct} \mid \text{answered})$, which would require assumptions we do not make (Abbasi-Yadkori et al., 2024), and does not improve answered-set accuracy in our evaluation. The gate does abstain (up to $\sim$9% of queries; §5.2); error filtering falls to the confidence ranking (empirical checks: Appendices H and J.2).

4 EXPERIMENTAL SETUP

Benchmarks and knowledge graphs. Six KGQA benchmarks in two regimes: HotpotQA (Yang et al., 2018), MuSiQue (Trivedi et al., 2022), and 2WikiMHQA (Ho et al., 2020) with 300 queries each (Mode A), and WebQSP (Yih et al., 2016) and CWQ (Talmor & Berant, 2018) with 500 each and MetaQA-3hop (Zhang et al., 2018) with 300 (Mode B); each query set is sampled once at random (not first-N: a pre-launch check found first-N directionally harder on WebQSP, -18pp, Fisher's exact $p{=}.09$) and fixed across three algorithm seeds (42/123/456). Mode A: entity graphs we index from each dataset's source corpus with a GraphRAG pipeline (Edge et al., 2024) (17.7k–31.7k entities, 23.6k–47.5k relations per dataset); the KG is derived from text, and text-RAG baselines can additionally read the underlying passages. Mode B: curated KB subgraphs (Freebase for WebQSP/CWQ; MetaQA's movie KB) with no corpus, so only KG-native methods apply. Our method targets point-answer KGQA; the set-answer regime (MetaQA-3hop, median 6 golds) enters as a pre-registered stress test of that scope, not its home ground.

Backbone and cost accounting. All methods use the same backbone (GPT-4.1-mini) and the same graph access. Judge and extraction calls run at temperature 0; the exceptions are our paraphrase generation (one batched call at 0.7, fixed seed) and the sampling baselines (R^2C at its released 0.7/1.0, R^2-KG at 0.95). Every LLM call is cached and logged (call counts per query; dollar costs are character-based estimates, one estimator for all methods; appendix). Particle-KGRAG uses a fixed budget of 129 calls/query (§3.3).

Baselines. (i) ToG-style beam search (our reimplementation): the primary same-interface comparator. Its confidence for AUROC comparison is σ(final top-path LLR), σ the logistic sigmoid; sequence-score confidence is an established selective-generation baseline (cf. Fadeeva et al., 2025), and UAG attaches conformal prediction to KG beam-search retrieval (Ni et al., 2025). ToG exposes no native confidence, so this derived proxy is the natural default, backed by beam-posterior, depth-adjusted, no-early-stop, and matched judge-call (relation text, clip) controls (appendix). We run it at nominal width 10, depth 3 (the original's default width is 3); our walks run to the pre-registered depth cap 6 unless dead-ended, so the comparison is not depth-matched. An audit found duplicate final paths collapsing the effective width on the Mode-B side (mean distinct paths 1.2–4.5), so all primary Mode-B accuracy and confidence comparisons use a deduplicated rerun under a protocol frozen before execution (two labelled pre-repair selective-prediction diagnostics excepted), which restored the effective width to 9.6–10.0, matching the Wikipedia side's 8.8–9.4 (original run and deltas: Appendix B, D8). (ii) ToG-2-style (Ma et al., 2025) (HotpotQA only, per pre-registration; its dense-retrieval coupling requires a source corpus, so it is undefined on the Mode-B side). (iii) R^2C (Soudani et al., 2026): perturbation-consistency UQ for RAG; Mode A only (it reads the corpus; Mode B provides none); run at its default $B{=}10$ perturbations and at a call-matched $B{=}40$ (registered as cost-matched). (iv) R^2-KG (Jo et al., 2025): dual-agent KG reasoning with supervisor-judged abstention; Mode B. (v) HippoRAG2 (Gutiérrez et al., 2025): the pre-registered corpus-reading upper reference (Mode A; no confidence signal of its own). (vi) Trivial floors: random-entity answers; every metric is reported against its trivial floor.

Metrics and testing. Accuracy: Hit@1 with contains-matching (RoG protocol (Luo et al., 2024); gold string contained in the normalized prediction). Confidence quality (primary axis): AUROC of each method's own confidence for predicting its own correctness (ours: $\rho(a_{\text{emit}})$, adopted post-registration over the registered top-1 mass; appendix D6). Selective prediction: risk–coverage curves, AURC, abstention rate. Efficiency: LLM calls and cost per query. Testing: paired bootstrap over matched (query, seed) pairs, per dataset, two-sided (the registered default; one-sided only where registered (C1, FC1), against-interest (toward the baseline), or a post-hoc diagnostic; appendix D7); no pooling across datasets (dataset identity confounds pooled AUROC); mean over 3 seeds; the six vs.-beam AUROC tests carry BH-FDR correction; deltas are rounded once from full precision.

Dataset	Ours	Beam	Δ (pp)	Verdict
HotpotQA	.354	.430	-7.6^{***}	beam
MuSiQue	.127	.148	-2.1^{*}	beam
2WikiMHQA	.381	.373	$+0.8$	tie
WebQSP	.629	.652	-2.3^{**}	beam
CWQ	.387	.428	-4.1^{***}	beam
MetaQA-3hop	.572	.813	-24.1^{***}	beam

Table 1: Hit@1: Particle-KGRAG (Ours) vs. the ToG-style beam (300–500 queries $\times$ 3 seeds per dataset; paired bootstrap). $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ (one-sided toward the beam). 95% CIs: all five deficits exclude zero; the tie's CI spans zero (non-significance, not equivalence). A query-clustered bootstrap (robustness; appendix) keeps the three largest deficits significant and widens MuSiQue and WebQSP to ties.

5 Results

5.1 Correctness Ranking Across Six Benchmarks

Table 2 evaluates each system's confidence against the correctness of its own answers. Particle-KGRAG reaches a mean AUROC of 70.6, compared with 63.7 for the ToG-style beam's search-derived score. Improvements on HotpotQA, 2WikiMHQA, and CWQ survive both paired and query-clustered tests with BH-FDR correction within the six-dataset comparison family; on 2WikiMHQA the beam's score is below chance (46.4). The advantage is not uniform: MetaQA-3hop favors the beam (68.0 vs. 58.8; paired two-sided p=.002, clustered two-sided p=.063, n.s.; appendix D7), and the MuSiQue ($+2.0$) and WebQSP (-1.9) differences are not established. A post-hoc cardinality breakdown localizes the MetaQA-3hop reversal: our AUROC collapses on large gold sets, 47.8 at 21+ golds (64.4 at 2–5 golds, 62.1 single-gold), while the beam holds (60.7–75.7); the posterior spreads mass across multi-modal answer sets, penalized by point-answer scoring (§6). A ToG-2-style baseline (HotpotQA only, §4) reaches Hit@1 .570, above both the beam and us, yet its confidence self-ranks at 61.3 (beam 65.6, ours 77.4): information access does not buy self-knowledge. The pre-registered criteria (C5 confidence floor and C6 evidence signal, met at 4/6 each), the C4 overlap analysis, and the held-out abstention check: Appendix H. The following sections examine selective prediction and whether these gains persist under beam-side answer aggregation.

What the beam's score tracks (post-hoc). The beam's derived confidence correlates with how early the search stopped (Pearson -0.51/-0.75 with depth on 2WikiMHQA/HotpotQA), and on 2WikiMHQA early stops are confident but wrong (depth-2 stops: confidence .92, accuracy .26). This is an observed association, not a demonstrated cause; neither disabling early stopping (2WikiMHQA 43.4) nor four label-free depth recalibrations (Table 27) repairs the score.

5.2 Selective Prediction and Answer Accuracy

On archived 2WikiMHQA runs, the full-coverage accuracy difference is not statistically established (ours .381 versus the beam's .373; Table 1), while selective accuracy at 50% coverage rises from .337 (beam) to .573 (ours; p<.001; Figure 2). The query-clustered difference is $+.237$, with a 95% confidence interval of $[+.155, +.317]$. This does not reflect higher answer accuracy: the beam has higher full-coverage accuracy on the other five benchmarks under the registered paired comparison (three of five under the query-clustered check). On the registered AURC, which compounds ranking quality with the base error rate, the beam leads on MuSiQue, MetaQA-3hop, and WebQSP, while Particle-KGRAG leads on 2WikiMHQA and CWQ and ties HotpotQA; the accuracy-adjusted excess AURC gives three wins under both tie rules and two losses (WebQSP, MetaQA-3hop; Appendix J.1). On HotpotQA, low coverage directionally reverses the accuracy deficit (.667 vs. .639 at 30%, n.s.); on the remaining four datasets the beam's higher raw accuracy carries its curve at half coverage (WebQSP: .733 vs. .671; Figure 3). One control cuts against us: at the conformal gate's operating point (coverage $\geq$ .91) the beam leads on 5/6 by absolute accuracy (pre-repair Mode-B beam; appendix); crossovers sit at lower coverage. At a fixed 40% target risk (same plug-in recipe both sides), ours answers 45% of 2WikiMHQA vs. the beam's 2% (30/30 splits; MetaQA-3hop reverses; post-hoc, pre-repair Mode-B beam; appendix). The corpus-reading upper reference

Panel A: archived run (primary)

	Particle-KGRAG vs. beam			Beam + answer aggregation	
Dataset	Particle-KGRAG	Beam	Δ	AUROC	Δ
HotpotQA	77.4 ± 1.7	65.6 ± 0.0	$+11.8^{*}$	68.1	$+9.3^{\S}$
MuSiQue	79.2 ± 2.1	77.2 ± 0.5	$+2.0$	70.9	$+8.3^{\S}$
2WikiMHQA	77.3 ± 1.3	46.4 ± 0.2	$+30.9^{*}$	76.0	$+1.3$
WebQSP	57.3 ± 2.1	59.3 ± 0.0	-1.9	53.8	$+3.5$
CWQ	73.3 ± 1.5	65.9 ± 0.0	$+7.4^{*}$	67.6	$+5.7^{\S}$
MetaQA-3hop	58.8 ± 3.9	68.0 ± 0.0	$-9.2^{\dagger}$	54.0	$+4.9$
mean	70.6	63.7		65.1	

Panel B: rendered-prompt rerun (pre-registered supplementary validation)

Dataset	Particle-KGRAG	Beam	Δ
HotpotQA	74.4 ± 1.5	61.1 ± 0.0	$+13.2^{*}$
MuSiQue	72.3 ± 1.0	66.7 ± 0.0	$+5.7$
2WikiMHQA	72.3 ± 1.5	36.8 ± 0.0	$+35.5^{*}$
WebQSP	56.5 ± 0.4	58.4 ± 0.0	-1.8
CWQ	70.1 ± 1.9	61.6 ± 0.0	$+8.5^{*}$
MetaQA-3hop	53.2 ± 2.0	67.7 ± 0.0	$-14.5^{\dagger}$
mean	66.4	58.7	

Table 2: Confidence quality (AUROC, correct vs. incorrect on each system's own answers). Particle-KGRAG: posterior mass of the emitted answer (post-registration primary, D6; registered variant: Table 3); Beam: search-derived σ(top-path LLR). Panel A, archived run: pooled AUROC over matched (query, seed) pairs; $\pm$: std over the three per-seed AUROCs (Table 10); the beam is near-deterministic. *two-sided $p < .05$ under both the paired and the query-clustered bootstrap, surviving BH-FDR under both (D7); †the single reversal (beam better): paired-significant two-sided, clustered n.s. in Panel A (D7), clustered-significant in Panel B. Beam + answer aggregation (post-hoc; re-paired Mode-B beam; Appendix J.4): an answer per beam path, softmax(LLR) mass, its top-mass answer emitted and labelled by that answer's Hit@1; seed-invariant, no std; its Δ forms a separate post-hoc family (§clustered two-sided $p < .05$, BH-FDR within it). Panel B (prompts rendered as written; Appendix K): computed as Panel A; the rendered beam's answers (hence labels) come from a top-path-only extraction. Mean rows: ours averages the per-seed AUROCs (the rendered pooled cells average 66.5). Δs are rounded once from full precision, so they can differ by 0.1 from the difference of the printed cells.

(HippoRAG2, Hit@1 .690/.353/.661) and the unmet exact-match parity criterion (C3) confirm the accuracy asymmetry (appendix). Table 1 reports all six accuracy results, and Figure 2 shows both the selective-prediction benefit and settings where the beam remains stronger. Every method clears the random-entity trivial floors (.000–.013).

5.3 What Changes When the Beam Also Aggregates Answers?

A search-derived score is only one confidence baseline. Applying Eq. (2) to the beam's own ten paths does not repair its confidence (2WikiMHQA 44.7; Table 26). We therefore equip the beam's archived paths with per-path answer extraction and answer-level aggregation (Beam + answer aggregation; three seeds; repaired Mode-B beam; Appendix J.4). The resulting comparison establishes advantages for Particle-KGRAG on HotpotQA, MuSiQue, and CWQ within this post-hoc six-comparison family (clustered two-sided $p \leq .023$, BH-FDR; a different triple from §5.1), but reduces the 2WikiMHQA difference from 30.9 to 1.3 points (95% CI -3.8 to $+6.1$; WebQSP $+3.5$, -1.0 to $+7.9$): no difference larger than this design can resolve (pre-registered margins 10.0 and 8.5 AUROC points), which bounds the difference without establishing practical equivalence. On 2WikiMHQA the gap thus shrinks by at least 20.9 of 30.9 points (13.2 at the gap's CI lower bound), without isolating aggregation as the cause. This comparison changes both answers and confidence: the aggregated beam emits its posterior mode, whereas our method emits the highest-weight particle's answer. The path sets, depth (six vs. three hops), and judging (the beam's own per-hop single-question LLRs)

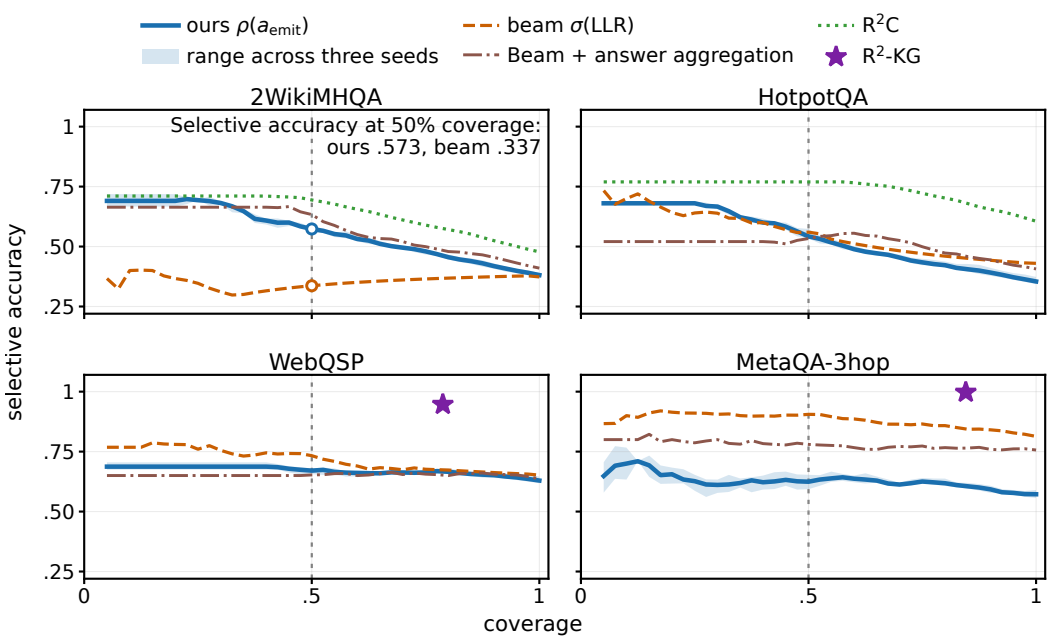


Figure 2: Archived run: selective accuracy vs. coverage on four of the six datasets (all six: Figure 3), including two where the beam's higher accuracy leads. Each method ranks its own answers by its own confidence. Band: range across three seeds (not a confidence interval). 2WikiMHQA markers: selective accuracy at 50% coverage, ours .573 vs. the beam's .337; there the R^{2}C curve lies above ours, and the Beam + answer aggregation curve lies above ours except at low coverage. Dotted vertical line: 50% coverage. R^{2}C: Wikipedia datasets only; R^2-KG: one operating point, answered-subset basis, Mode B; Beam + answer aggregation: post-hoc, Appendix J.4, mass saturated at 1.0 on 13–49% of rows per panel, prorated as ties. Prorated tie rule (the reported AURC uses the block rule, Appendix J.1); crossovers post-hoc. The ESS gate's false-abstention bound is not a property of these curves.

are also not fully matched. Scoring the beam instead by the unweighted answer frequency over its final paths (in the spirit of Fadeeva et al., 2025; post-hoc, archived prompts), our advantage is established on three of six datasets and reversed on zero; on the others it is bounded within the inherited margin of Appendix J.4 (two) or not established (one). The three are HotpotQA, MuSiQue, and MetaQA-3hop; on 2WikiMHQA the point estimate favors the beam (81.6 vs. 77.3; clustered p=.066). Grafting our posterior onto the beam's own answers repairs its worst confidence failure relative to σ(LLR) (2WikiMHQA $+30.8$; on the repaired Mode-B beam, CWQ $+2.1$ is not established and MetaQA-3hop reverses, -28.3), but at that same fixed answer and label, no difference between our posterior's mass and the beam's own aggregation mass is established on any dataset (post-hoc; query-clustered two-sided p, BH-FDR within that family; Appendices J.4–J.5). These controls delimit the empirical gain without isolating independent sampling or judge weighting as its cause.

5.4 Comparison with Perturbation-Based Uncertainty

Against R^{2}C on the three Wikipedia benchmarks (it requires a corpus), all three default-budget AUROC differences are positive ($+0.7$ to $+3.7$ on the emitted-answer score; $+1.2$ to $+3.6$ on the registered top-1 mass, where the pre-registered directional C1 criterion holds; Table 34), but none survives BH-FDR correction, and all sit below the 80%-power minimum detectable effect (4.8–6.6 points). The registered claim is therefore directional consistency at strictly lower information access, not superiority: R^{2}C reads the source passages, whereas Particle-KGRAG operates on the derived KG, and R^{2}C has higher answer accuracy (Hit@1 .606/.182/.478 vs. ours .354/.127/.381, surviving EM re-scoring). Quadrupling R^{2}C's perturbation budget to our call budget (B=40) barely moves its confidence quality on two of three benchmarks: HotpotQA AUROC 73.7 $\rightarrow$ 74.2 (advantage

$+3.7$, p=.039 → $+3.2$, p=.055), MuSiQue 76.5 → 76.8 ($+2.7$ → $+2.5$), with accuracy essentially unchanged. On 2WikiMHQA the extra samples do help R^2C (76.6 → 77.4), closing our advantage to a tie (paired, post-hoc: $\Delta+0.7$, p=.37 → -0.1, p=.53); R^2C does so by re-running its corpus-reading pipeline 40× per query. A P(True)-style self-check (Kadavath et al., 2022) on our final answers (post-hoc) scores 73.2/59.0 on HotpotQA/2WikiMHQA vs. our 77.4/77.3 ($\Delta+4.2$, one-sided p=.016; $\Delta+18.3$, p<.001). Per-dataset B=40 results and tests: Table 12, Appendix B.

5.5 Robustness and Component Analyses

A separately pre-registered rerun with corrected prompt rendering yields mean AUROC 66.4 versus 58.7 and retains the three positive findings on HotpotQA, 2WikiMHQA, and CWQ; MetaQA-3hop favors the beam (Table 2, Panel B). The archived run remains the primary analysis. Rendering also changes the beam's final extraction input to the top path; Appendix K reports this consequence and the secondary extraction-form check.

Component ablations do not establish the PPR prior or judge weighting as the source of the headline gains. Against a uniform walk prior (pre-registered; a different target, §3.1), top-1 AUROC gains reach one-sided p=.046 on MuSiQue ($+3.3$) and p=.040 on WebQSP ($+2.3$), two-sided n.s. (D7), and are null elsewhere. A uniform vote over the same particles (post-hoc; cache replay) lowers AUROC on 5/6 datasets (significant on three, -7.7/-2.7/-1.9; MetaQA-3hop to chance; exception 2WikiMHQA, $+1.5$ n.s.) and accuracy on 5/6, but the weights' significant gains are disjoint from the three headline AUROC wins (weighting n.s. there). The post-hoc median-versus-single-paraphrase comparisons (the declared necessity rule holds on 0/18 dataset–seed cells, without an equivalence margin) likewise do not establish equivalence or validate the cheaper 65-call judge. Per-dataset effects, testing details, and the K and backbone checks: Tables 30 and 19, Appendices I and H.6–H.8.

6 Limitations and Scope

These comparisons evaluate each method's confidence against its own answers, so answer accuracy is not held fixed. Particle-KGRAG establishes no answer-accuracy improvement over the beam (one tie, five losses; Table 1), and no AUROC advantage over R^2C survives multiplicity correction; text-RAG answers better (R^2C .606 vs. .354 on HotpotQA). With the beam's answer and correctness label held fixed, no difference between our posterior and the beam's own answer aggregation is established on any dataset. Our AUROC falls as our Hit@1 rises across the six benchmarks (Tables 1–2; the beam's pairing is not monotone); no accuracy-matched control is reported (closest: Appendix J.5). The path sets, depth, judging, and emission rules are not fully matched across systems. Pre-registered criteria not met: accuracy parity (ties-or-beats on ≥4/6; met on 1/6) and the 0.88 coverage target of the conformal gate (FC2; provable-coverage claim downgraded, D2). Confidence also falls below the pre-registered floor of 65 on WebQSP and MetaQA-3hop (57.3/57.0 and 58.8/58.7, emitted/registered), although C5 passes at 4/6. Confidence is uncalibrated (ECE worse than the beam on 4/6, post-hoc), and the auxiliary ESS gate bounds false abstention rather than answered-set error: it filters only 1–10% of errors at α=.1, is vacuous on low-accuracy folds (MuSiQue, 9/30 splits at the $\tau\geq1$ ESS floor), and lowers answered-set accuracy on all 18 dataset–seed systems (0.1–2.2 points; Appendix J.2). Set-valued gold (MetaQA-3hop, median 6 golds) penalizes the point-answer design (set-F1 .237 vs. R^2-KG .722). The main study uses one backbone (GPT-4.1-mini), with limited post-hoc checks on another model (Qwen2.5-7B), reimplemented baselines, and three seeds; protocol deviations (including the registered name "twisted SMC," reported here as SNIS; D1) and prompt-rendering effects are disclosed (Appendices B and K).

7 Conclusion

Particle-KGRAG provides an explicit route from judge-scored graph paths to confidence in a returned answer. Without task-specific training, the construction improves correctness ranking over the evaluated beam score on three of six benchmarks (BH-FDR-corrected) and supports higher selective accuracy on archived 2WikiMHQA runs. Beam aggregation and fixed-answer comparisons delimit these benefits and leave their attribution to independent sampling unresolved. The study makes answer-level aggregation a concrete design and evaluation target for confidence-aware KGQA.

AI USE STATEMENT

In this work, we used generative AI tools (LLM-based coding and writing assistants, with independent cross-checks by other commercial LLM assistants) for the following tasks with required disclosure. *Implementing methods*: the method, the baseline re-implementations, and the analysis and table-generation scripts were drafted with AI assistance, then reviewed, tested, and executed by the authors; the numbers reported in the paper were regenerated by those scripts from the archived per-query results and LLM cache, which the supplement does not include: it ships the code, the aggregated results, the pinned analysis outputs, and the matched evaluation frames, and supports offline recomputation of the headline confidence metrics, the AURC tie-convention check, and the supplement-table file. *Designing or providing feedback on research methodology or experiments*: AI-generated critiques, pre-mortems, and mock reviews of the pre-registered protocol and of the manuscript were used as input, and the authors adjudicated each point. *Proposing or refining hypotheses*: AI tools assisted in drafting and stress-testing the pre-registered protocol (hypotheses, decision rules) that the authors finalized and froze. *Cleaning and reformatting datasets*: the benchmark-preprocessing and graph-construction code was written with AI assistance. *Interpreting results*: AI tools produced descriptive summaries and consistency checks of the archived results, which the authors consulted; the interpretations and all claims are the authors' own. *Developing conceptual frameworks*: AI tools were used to check the correspondence between the implemented sampler and its description as self-normalized importance sampling; the paper makes no new theoretical claims. *Assisting with translation*: the pre-registration was written with interleaved notes in the authors' working language; for the anonymized copy in the code supplement those notes were translated into English line-for-line with AI assistance and then checked by the authors, leaving the registered field names, numbers, and thresholds unchanged. We did not use generative AI tools to generate synthetic datasets or to formulate or prove mathematical claims; qualitative or thematic data analysis is not applicable to this work. Additionally, we used generative AI tools for tasks with recommended disclosure: drafting and editing parts of the paper to improve readability, creating and modifying figures and tables through generated code, summarizing and identifying relevant literature, formatting references, suggesting the paper's structure, and brainstorming. Separately from the research process, LLMs are components of the studied system (the judge and answer-extraction model, and a second backbone in a cross-backbone check), as described in §3, §4, §6, and Appendix H.8. We have reviewed all AI-assisted work: LLM-generated code was verified and tested by the authors, every reported number was checked by scripts against the archived results (the shipped checker covers the headline metrics; the remaining checks ran on the unshipped per-query archive), every citation was verified against its original source, and AI-suggested literature and critiques were checked manually. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies confidence estimation for knowledge-graph question answering on six public benchmarks with public knowledge graphs and Wikipedia-derived corpora; it involves no human subjects, no collection of personal data, and no crowd annotation. LLM inference used commercial and open-weight models under their terms of use. Because the method emits a confidence score and an abstention rule, over-trust is the main foreseeable risk; the paper therefore states explicitly that the method does not improve answer accuracy, that its conformal guarantee is one-sided (it bounds false abstention, not the error rate of answered queries), and that the confidence advantage reverses on one benchmark. We see no direct harmful applications beyond those of knowledge-graph question-answering systems in general.

REPRODUCIBILITY STATEMENT

The full experimental protocol (sample sizes, seeds, metrics, decision and falsification criteria) was pre-specified and locally frozen before the main run, as a tagged commit in the authors' repository rather than a deposit with an external time-stamping service; Appendix B reproduces the pre-registered rules, enumerates every deviation, states the SHA-256 digest of the frozen protocol file, and says what that digest does and does not certify. Appendix C lists the implementation details that affect reproduction (backbone and decoding settings, per-query call count, judge prompt and log-probability handling, proposal and scoring fallbacks, cost account-

ing, caching, software versions). Appendix K reports a pre-registered rerun of the two headline configurations with the judge and extraction prompts rendered as written (Table 2, Panel B). The models were accessed through their public APIs as `gpt-4.1-mini` (OpenAI; primary backbone for every method, main-run calls in June–July 2026, the rendered-prompt rerun's in September 2026), `Qwen/Qwen2.5-7B-Instruct-Turbo` (Together; cross-backbone check), and `text-embedding-3-small` (OpenAI; graph indexing); no dated snapshot beyond each provider's default at run time was pinned. All experiments ran on a single laptop (Apple M4 Max, CPU-only) against these APIs. The anonymized code supplement (supplementary material) contains the method, all baseline re-implementations, the analysis scripts, the aggregated per-dataset results and pinned analysis outputs, an anonymized copy of the frozen protocol, the rendered-prompt rerun's registration, scripts, gate records and readout, and the matched per-query confidence frames behind the headline numbers together with a checker script that recomputes the reported AUROCs, the six-dataset means, and the 2WikiMultiHopQA selective-accuracy pair from the archive alone. Analyses marked as cache replays in the appendix were computed from archived LLM responses without new calls; the archive ships results and frames, not the LLM cache, so re-running the pipeline end to end requires API access to the same backbone.

REFERENCES

Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating LLM hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024. URL https://arxiv.org/abs/2405.01563.

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021. URL https://arxiv.org/abs/2107.07511.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. From local to global: A Graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL https://arxiv.org/abs/2404.16130.

Nicolas Emmenegger, Theo X. Olausson, Armando Solar-Lezama, and Chara Podimata. Conformal language modeling via posterior sampling. *arXiv preprint arXiv:2606.03731*, 2026. URL https://arxiv.org/abs/2606.03731.

Ekaterina Fadeeva, Maiya Goloburda, Aleksandr Rubashevskii, Roman Vashurin, Artem Shelmanov, Preslav Nakov, Mrinmaya Sachan, and Maxim Panov. Don't throw away your beams: Improving consistency-based uncertainties in LLMs via beam search. *arXiv preprint arXiv:2512.09538*, 2025. URL https://arxiv.org/abs/2512.09538.

Edwin Fong and Chris C. Holmes. Conformal Bayesian computation. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/97785e0500ad16c18574c64189ccf4b4-Abstract.html.

Giorgio Giannone, Mustafa Eyceoz, Shabana Baig, Shivchander Sudalairaj, Anna C. Doris, Faez Ahmed, Akash Srivastava, and Kai Xu. Intrinsic selection and particle resampling for inference-time scaling beyond domain verifiability. *arXiv preprint arXiv:2606.08850*, 2026. URL https://arxiv.org/abs/2606.08850.

Bernal Jiménez Gutiérrez, Yiheng Shu, Weijian Qi, Sizhe Zhou, and Yu Su. From RAG to memory: Non-parametric continual learning for large language models. In *International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 21497–21515, 2025. URL https://arxiv.org/abs/2502.14802. arXiv:2502.14802.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, 2020.

Sumin Jo, Junseong Choi, Jiho Kim, and Edward Choi. R2-KG: General-purpose dual-agent framework for reliable reasoning on knowledge graphs. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2025*, pp. 486–509, 2025. URL https://aclanthology.org/2025.findings-ijcnlp.29/. arXiv:2502.12767.

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. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Shuhang Lin, Chuhao Zhou, Xiao Lin, Zihan Dong, Kuan Lu, Zhencan Peng, Jie Yin, and Dimitris N. Metaxas. Conformal path reasoning: Trustworthy knowledge graph question answering via path-level calibration. *arXiv preprint arXiv:2605.08077*, 2026. URL https://arxiv.org/abs/2605.08077.

Xiao Long, Liansheng Zhuang, Aodi Li, Minghong Yao, and Shafei Wang. EPERM: An evidence path enhanced reasoning model for knowledge graph question and answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12282–12290, 2025a. URL https://doi.org/10.1609/aaai.v39i12.33338. arXiv:2502.16171.

Xiao Long, Liansheng Zhuang, Chen Shen, Shaotian Yan, Yifei Li, and Shafei Wang. Enhancing large language models with reward-guided tree search for knowledge graph question and answering. *arXiv preprint arXiv:2505.12476*, 2025b. URL https://arxiv.org/abs/2505.12476.

Xiayin Lou and Peng Luo. Weighted Bayesian conformal prediction. *arXiv preprint arXiv:2604.06464*, 2026. URL https://arxiv.org/abs/2604.06464.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2310.01061.

Shengjie Ma, Chengjin Xu, Xuhui Jiang, Muzhi Li, Huaren Qu, Cehao Yang, Jiaxin Mao, and Jian Guo. Think-on-graph 2.0: Deep and faithful large language model reasoning with knowledge-guided retrieval augmented generation. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://arxiv.org/abs/2407.10805. arXiv:2407.10805.

Bo Ni, Yu Wang, Lu Cheng, Erik Blasch, and Tyler Derr. Towards trustworthy knowledge graph reasoning: An uncertainty aware perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 12417–12425, 2025. URL https://arxiv.org/abs/2410.08985. arXiv:2410.08985.

Isha Puri, Shivchander Sudalairaj, Guangxuan Xu, Abhishek Bhandwaldar, Kai Xu, and Akash Srivastava. Rollout roulette: A probabilistic inference approach to inference-time scaling of LLMs using particle-based Monte Carlo methods. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. URL https://arxiv.org/abs/2502.01618. arXiv:2502.01618.

Xiangqing Shen, Fanfan Wang, Zinong Yang, Bing Wang, Wenli Du, Chengqing Zong, and Rui Xia. Reason-align-respond: Aligning LLM reasoning with knowledge graphs for KGQA. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(7):7340–7353, 2026. doi: 10.1109/TPAMI.2026.3665645. URL https://arxiv.org/abs/2505.20971. arXiv:2505.20971.

Xiaozhuang Song, Shufei Zhang, and Tianshu Yu. ReKG-MCTS: Reinforcing LLM reasoning on knowledge graphs via training-free Monte Carlo tree search. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9288–9306, 2025. URL https://aclanthology.org/2025.findings-acl.484/.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 16596–16616, 2025. URL https://aclanthology.org/2025.findings-acl.852/.

Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. Uncertainty quantification for retrieval-augmented reasoning. In *ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pp. 1665–1676, 2026. URL https://arxiv.org/abs/2510.11483. arXiv:2510.11483.

Jiashuo Sun, Chengjin Xu, Lumingyuan Tang, Saizhuo Wang, Chen Lin, Yeyun Gong, Lionel M. Ni, Heung-Yeung Shum, and Jian Guo. Think-on-graph: Deep and responsible reasoning of large language model on knowledge graph. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2307.07697. arXiv:2307.07697.

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*, 2018.

Amir Taubenfeld, Tom Sheffer, Eran Ofek, Amir Feder, Ariel Goldstein, Zorik Gekhman, and Gal Yona. Confidence improves self-consistency in LLMs. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 20090–20111, 2025. URL https://aclanthology.org/2025.findings-acl.1030/. arXiv:2502.06233.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2023. URL https://arxiv.org/abs/2305.14975. arXiv:2305.14975.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics (TACL)*, 10:539–554, 2022.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023.

Zhiyuan Wang, Jinhao Duan, Qingni Wang, Xiaofeng Zhu, Tianlong Chen, Xiaoshuang Shi, and Kaidi Xu. Coin: Uncertainty-guarding selective question answering for foundation models with provable risk guarantees. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33764–33772, 2026. URL https://ojs.aaai.org/index.php/AAAI/article/view/40667. arXiv:2506.20178.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

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, 2016.

Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander J. Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 6069–6076, 2018.

Stephen Zhao, Rob Brekelmans, Alireza Makhzani, and Roger Baker Grosse. Probabilistic inference in language models via twisted sequential Monte Carlo. In *International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 60704–60748, 2024. URL https://arxiv.org/abs/2404.17546. arXiv:2404.17546.

A Overview

This appendix provides per-seed detail, robustness checks, and post-hoc controls for the headline aggregate numbers of the main paper. Each caption states whether the analysis was pre-registered or post-hoc (all post-hoc analyses are additionally labeled in our archived analysis artifacts, shipped with the code supplement). Table 10 gives per-seed confidence-quality detail behind main Table 2; Table 11 the HippoRAG2 corpus-reading upper reference and the C3/FC3 verdicts (C3: registered EM parity with the corpus reader, unmet; FC3: falsification had our total cost exceeded $30\times$ HippoRAG2's, not triggered at $0.44\times$); Table 12 the full call-matched control including the 2WikiMHQA tie disclosed in main §5.4; Table 13 the held-out false-abstention check behind the conformal abstention rule, including its abstention recall; Table 14 the answer-cardinality mechanism behind the MetaQA failure; Tables 15 and 16 the P(True) same-answer baseline and the K-sensitivity sweep; Table 17 threshold-free AUPRC plus the random-ranking AURC floor; Table 18 excess AURC (the base-error-corrected AURC referenced in main §5.2); Table 19 the vote-share ablation isolating the judge weights' contribution; Tables 20 and 21 the alternative-signal comparison (entropy variants) and the search-vs-selection oracle decomposition; Table 22 the query-clustered bootstrap; Table 23 the matched-coverage control at our conformal operating point; Table 24 the cumulative top-k mass control (the natural multi-answer repair; it does not help); Table 25 the particle posterior grafted onto the beam's answers as a drop-in confidence layer (repeated with the answer held fixed: Appendix J.5); Table 26 the beam-path-posterior control (Eq. (2) applied to the beam's own ten paths); its Mode-A cells show that Eq. (2) on the beam's own paths does not close the gap, while on 2WikiMHQA its answer-aggregation variant, extended to three seeds and the repaired Mode-B beam in Appendix J.4, leaves no difference larger than this design can resolve (main §5.3), its pre-repair Mode-B cells being diagnostic only; Table 27 the depth-adjusted beam-confidence control (four label-free depth repairs; none closes the gap); Table 28 fixed-target-risk operating points (the identically calibrated plug-in threshold on both sides, per split); Table 29 the round-18 robustness companion (query-clustered accuracy bootstrap, strict-EM re-scoring, and the 2WikiMHQA selective-accuracy interval); and Table 30 the full PPR-proposal ablation. The registered evidence-signal criterion C6 (Appendix H) is scored per dataset on $\log \hat{Z}_q$ AUROC (tie-safe canonical aggregation): 64.0/62.7/69.5 (HotpotQA/MuSiQue/2WikiMHQA) and 62.6 (CWQ) clear the registered bar of 60, while WebQSP (55.9) and MetaQA-3hop (50.6) do not, giving the 4/6 pass reported there. Figure 3 gives the six-dataset risk–coverage curves behind main-text Figure 2. Appendix I (added in this version) reports a post-hoc component ablation of the paraphrase median (Tables 35–36). Every control in Appendices F–J replays the archived run (judge and extraction prompts as sent). Appendix K (added in this version) reports the pre-registered rerun with both prompts rendered as written (Table 2, Panel B).

B Pre-Registration and Deviations

The full protocol (sample sizes, seeds, metrics, decision rules, falsification criteria) was frozen before the main run; its decision rules are reproduced in the main paper (§4) and its deviations are enumerated below. The frozen protocol file embeds identifying metadata (researcher, timestamps, an internal decision log); the code supplement therefore ships an anonymized copy (`preregistration/` in the archive): verbatim except in-place redaction (marked `[redacted-*]`) of identities, venue names, and the two schedule details that would name a venue without naming it (its deadlines and its page limit), plus a header noting the redaction; its interleaved working-language notes are translated into English line-for-line in this release (the header lists the translated lines; registered field names, numbers, and thresholds are unchanged). The unredacted original is release-tagged in the repository accompanying the camera-ready. As a verifiable commitment, the unredacted frozen file's SHA-256 digest is `8ad4b4c3 38701bc1 728d2d02 ee9e4ca3 fcd4ed43 bc9dc75e 4835a96f 50b2f256` (spaces are line-breaking aids only; strip them to verify); the camera-ready release must reproduce this digest exactly (the shipped

anonymized copy differs from it only by the redaction header, the redaction markers, and the line-for-line translation of the working-language notes). Two limits of this commitment are stated plainly. The freeze is local: the protocol was frozen as a tagged commit in the authors' version-controlled repository, not deposited with an independent time-stamping service, so the digest certifies which file was frozen, and the repository history its position relative to the main run, but a reader cannot verify the freeze date independently of the authors. And because the digest is that of the unredacted file, it cannot be checked against the shipped anonymized copy at review time; it binds the camera-ready release. The code supplement additionally ships the matched per-query confidence frames behind the headline numbers (6,600 rows: dataset, query id, seed, both methods' confidence and correctness label; no text) with a checker that recomputes the per-dataset AUROCs, the six-dataset means, and the 2WikiMHQA selective-accuracy pair from the archive alone and checks them against the shipped pins (`scripts/verify_headline_from_frames.py`). Every deviation between the registration and this paper is listed here.

D1: Sampler naming (twisted SMC → SNIS). The registration described the sampler as twisted SMC with ESS-adaptive resampling. Under the registered uninformative twist ($\psi_t \equiv 1$) the intermediate ESS equals K identically, so the resampling trigger is structurally unreachable and the algorithm that actually ran is self-normalized importance sampling. The paper follows the code (main §3.1). No numbers are affected; this is a naming correction, disclosed rather than silently rewritten.

D2: Coverage claim downgraded (criterion C2; falsification consequence FC2). The registration (criterion C2) targeted empirical conformal coverage ≥ 0.88 at $\alpha = 0.1$; the observed ESS-based rule missed this target. Per the registered falsification consequence, the planned provable-coverage claim was downgraded to AURC-based selective-prediction ranking, and ESS is reported as a diagnostic. The direction that is certified ($\mathbb{P}(\text{abstain} \mid \text{correct}) \leq \alpha$) is stated precisely in main §3.4. A post-hoc diagnosis of why the bar was out of reach: with base accuracy .127–.629 across benchmarks (main Table 1), a $\geq$.88-conditional-accuracy operating point exists only at near-zero answer rates: under the identical plug-in calibration of Table 28, a 10% target risk leaves our confidence with no feasible threshold on 20–30 of 30 splits per benchmark (mean coverage $\leq$.034; the beam likewise, except on its highest-accuracy dataset, MetaQA-3hop). The registered bar, not the abstention mechanism, was miscalibrated to this accuracy regime.

D3: Beam-parity rule not met. The registered accuracy-parity rule (ties-or-beats the beam on $\geq$4/6) failed at 1/6 (the deduplicated Mode-B beam rerun of D8; 3/6 against the archived original run). All headline claims are therefore confined to confidence quality and selective prediction; the paper states this explicitly in §5.2 and §6.

D4: Post-hoc analyses. All analyses not in the registration are labeled POST-HOC in their table captions and in the archived artifacts: the excess-AURC correction, the K-sensitivity sweep, the P(True) same-answer baseline, the cardinality decomposition, the query-clustered bootstrap, the matched-coverage control, the cross-backbone (Qwen) check, the set-F1 re-scoring, the leakage screen (5 of 200 sampled correct answers contain a gold alias in the question text; all benign paraphrases, none prompt-copied), the beam-confidence dissection and its depth-adjusted repair control, the vote-share, alternative-signal, and cumulative top-k-mass ablations, the beam-graft and beam-path-posterior controls, the answer–confidence alignment analysis, and the oracle-recall decomposition; the ECE comparison (main §6) is prose-only and equally post-hoc. One measurement convention was fixed after the runs: AUROC ranking rounds confidences to 10^{-9} (the "tie-safe" convention referenced in table captions; serialization-level float noise otherwise creates spurious ranks inside full-consensus ties, moving AUROC by up to ± 0.4; no registered verdict or significance conclusion changes under either convention, and the sensitivity is archived). None of them alters a registered verdict; they diagnose mechanisms or tighten robustness.

D5: Registered baselines not run at full scope. RoG is fine-tuned on WebQSP/CWQ train splits and was excluded from the training-free comparison set (disclosed in main §2); ToG-2 was registered and run on HotpotQA only, as its dense-retrieval coupling requires a source corpus that the Mode-B side lacks.

D6: Primary confidence score (registered top-1 mass → emitted-answer mass). The registration (C5) defines the confidence proxy as the top-1 (mode) mass of the answer posterior, $\max_a \rho(a)$. Following external pre-submission feedback (2026-07-16), the paper's primary per-query confidence is instead the posterior mass of the emitted answer, $\rho(a_{\text{emit}})$: the confidence attached to the answer the system actually returns, closing the semantic gap analyzed in §F. The switch is post-hoc, and its retention depended on a pre-stated guard that was evaluated on the observed results: revert to the registered top-1 mass unless $\geq$3/6 datasets remain significant under both the paired and the query-clustered bootstrap with BH-FDR. The guard held (3/6: HotpotQA, 2WikiMHQA, CWQ). The emitted-mass comparisons are therefore post-hoc analyses with a robustness check, not an independently registered confirmation. All registered criteria (C1, C5) remain evaluated on the registered top-1 mass and their verdicts are unchanged; Table 3 preserves the registered variant of main Table 2. An initial computation of $\rho(a_{\text{emit}})$ passed the emitted answer through a non-canonical normalizer (a punctuation-deleting variant, run against posterior keys stored in canonical form), zeroing the emitted mass on 443 of 6,597 answered rows (148 of them correct) and depressing the first-circulated emitted numbers (pooled mean 68.7). Under the fixed pipeline (posterior keys compared as stored, the emitted string normalized once with the canonical normalizer), no spuriously zeroed row remains, the guard re-passes with the same three survivors, and the switch is essentially free: pooled mean AUROC 70.6 emitted vs. 70.4 registered top-1 ($+0.1$; rounded once from full precision, so the printed pair appears 0.2 apart). Two apparent costs dissolve with the fix: the C1 2WikiMHQA comparison no longer reverses and the HotpotQA P(True) margin is significant (both main §5.4).

D7: Significance reporting (paired → paired + query-clustered). The registered testing protocol is the per-dataset paired bootstrap over matched (query, seed) pairs. Because a query's three seed rows share the graph neighborhood, every starred claim in the main text additionally requires survival under a query-clustered bootstrap (qids resampled with all their seed rows, B=2000, add-one p) with BH-FDR; paired-only p-values are retained in the appendix tables. Strictly more conservative; no registered verdict changes. Sidedness: the registration's default test family is two-sided (one-sided tests were registered only for C1 and FC1). Earlier drafts reported one-sided p in each claim's declared direction: a direction fixed only after the effects were observed; the final revision therefore reports the primary vs.-beam confidence family two-sided (main §4). The switch is immaterial for the three starred advantages (two-sided clustered $p \leq .005$, all significant after BH-FDR). The MetaQA reversal keeps paired significance two-sided (p=.002) but its clustered check does not reach two-sided significance (p=.063; one-sided .031): main Table 2 keeps the distinct † marker and the prose reports the clustered check as n.s. (the reversal counts against our method and is reported in full). Remaining one-sided values, itemized: (i) registered directions (C1, FC1); (ii) against-interest directional tests, where significance strengthens the negative result: the Table 1 accuracy stars and the C3 EM-parity p's (both toward the baseline); (iii) tests whose direction was fixed post hoc, reported as computed: the registered proposal ablation (p=.046/.040), the beam-path-control bound (6/6 at $p\leq$.049), and the P(True) margins (p=.016, $p<$.001), the answer-aggregation control of Appendix J.4 being two-sided throughout and so outside this list; (iv) the post-hoc tie-convention checks of Appendix J.1 (AURC and excess AURC under prorated ties), tested one-sided for ours lower, as computed. A two-sided reading at most doubles an add-one bootstrap p: that moves the ablation gains and the $\leq$ .049 bound across .05, with no main-text conclusion moving, since the ablation paragraph's claim is deflationary (the proposal adds at most a small gain) and the answer-aggregation control is already two-sided (advantage established on three datasets within its post-hoc family, clustered $p\leq$.023; Appendix J.4); the P(True) HotpotQA margin stays significant (.032). Among the tie-convention checks, doubling moves the HotpotQA query-clustered excess-AURC value (.044) across .05, which Appendix J.1 already flags (its clustered interval includes zero), while its paired value stays significant; the 2WikiMHQA and CWQ values stay significant; MuSiQue's nominal block-end value is not counted in any reading.

D8: ToG-1 Mode-B beam width repair (protocol-frozen deduplicated rerun). A 2026-07-17 pre-submission audit found the beam's Mode-B runs collapsing to effective width far below K_{beam}: the runner seeded all K_{beam} slots with a single topic entity and never deduplicated expansion candidates (MetaQA-3hop: 281/300 seed-42 queries scored ten identical paths, mean distinct 1.16; WebQSP 3.92, CWQ 4.46; Mode A healthy at 8.8–9.4). A rerun with candidate deduplication, its protocol frozen before execution (`tog1_dedup`; every other component unchanged), restores width to 9.6–10.0 and is canonical for every Mode-B beam cell in the main text. The repaired beam is

stronger: Hit@1 rises on all three datasets (MetaQA-3hop .637→.813), moving the accuracy tally to one statistical tie, five losses; its MetaQA-3hop confidence AUROC moves 73.2→68.0 (still significantly above ours), the six-dataset beam AUROC mean to 63.7, and its call count shrinks to the faithful accounting (145–218/query), under which the beam is also cheaper per query than our fixed budget on four of six datasets (Appendix H.6). The three BH-FDR confidence survivors are unchanged. Appendix control tables that replay the archived beam paths analyze the pre-repair run and say so in their captions (the repaired-beam answer-aggregation control of Appendix J.4 is the exception: it reads the deduplicated run); Table 10 and the D6 companion below use the rerun.

Dataset	top-1 mass	Beam	Δ
HotpotQA	77.3	65.6	$+11.6^{*}$
MuSiQue	78.3	77.2	$+1.1$
2WikiMHQA	77.9	46.4	$+31.5^{*}$
WebQSP	57.0	59.3	-2.2
CWQ	73.4	65.9	$+7.5^{*}$
MetaQA-3hop	58.7	68.0	$-9.3^{\dagger}$
mean	70.4	63.7	

Table 3: D6 companion: the registered top-1–mass variant of main Table 2 (confidence AUROC vs. the beam; $^{*}p < .05$ paired bootstrap, the three starred advantages survive BH-FDR; †the significant reversal, beam better, marked as in main Table 2). The main text's primary is $\rho(a_{\text{emit}})$; the registered metric is preserved for comparison with the registration, with Mode-B beam cells from the deduplicated rerun (D8), matching main Table 2.

D9: Multiplicity families. The registration specified Benjamini–Hochberg FDR within per-table comparison families as the primary correction (for the main table, Particle-KGRAG against each baseline across its accuracy and confidence metrics) and, as the supplementary family-wise control, a Bonferroni correction over all hypotheses across the tables (about 150 in the registration's count) for headline-claim verification. The reported families are narrower: the six per-dataset vs.-beam confidence comparisons of Table 2 (also the family of the D6 guard; Mode B against the deduplicated beam, D8) and the three Mode-A R^2C comparisons (the registered C1 family). The broader per-table families were not evaluated as single families, and no Bonferroni check was reported in earlier drafts. As a post-hoc sensitivity check on the reported six-comparison family, a Bonferroni threshold of $\alpha/6 = .0083$ leaves the three advantages significant under both the paired p (each at the bootstrap floor: no resampled difference at or below zero in $B{=}1000$) and the two-sided query-clustered p (HotpotQA .004, 2WikiMHQA .001, CWQ .005; two-sided values double the smaller of the two add-one tails, so the two negative differences, WebQSP and MetaQA-3hop, are never counted as advantages), so at this family size the check changes no verdict. This six-comparison check does not implement the registered global correction: at the registered family size the threshold ($\alpha/150 \approx .0003$) lies below the Monte Carlo resolution of the retained bootstraps ($1/B$ with $B{=}1000$ paired; $1/(B{+}1)$ with $B{=}2000$ clustered), the bootstraps were not recomputed at a larger budget, and no comparison is claimed to survive it. The BH-adjusted significance reported in the paper applies only to the narrower declared families, not to the originally planned collection of hypotheses; we record the family narrowing as a deviation (the Bonferroni script and its pin are in the code supplement). The post-hoc repaired-beam answer-aggregation control (Appendix J.4) adds 108 reported p-values (three scoring variants; per dataset the three per-seed paired, the pooled paired, the query-clustered, and the TOST two-sided values of Tables 40–42); each six-comparison family is BH-corrected as its plan declares, none enters the registered global correction, which remains unexecuted, and no comparison is claimed to survive it. The fixed-answer analysis (Appendix J.5) adds 48 reported p-values under the same rule (six clustered per contrast; eighteen per-seed for the primary contrast). The rendered-prompt rerun (Appendix K) adds 54 (per read-out dataset, the paired and clustered two-sided values of the primary and of the registered variant and the Hit@1 test on the rendered rows; the paired and clustered two-sided values of secondaries S1 and S2); its marks are BH-corrected over the read-out datasets as its plan declares, and none enters the registered global correction. The answer-frequency rescoring of the beam (Tables 43–44) adds 72 reported p-values (two scoring variants, each with the per-dataset set of the answer-aggregation control) and the posterior-mode what-if (Appendix F) adds 30 (eighteen per-seed and six clustered against the beam,

six clustered against the key-scored emitted answer); each six-comparison family is BH-corrected as the plan declares, and none enters the registered global correction. The fragmentation and judge-score analyses report no p-values.

Registration status at a glance. Table 4 lists where each reported result stands relative to the registration.

Result	Status	Where
C1 against R^2C (registered top-1 mass)	registered	Table 34
C2/FC2 coverage, C5/C6 criteria, beam-parity rule	registered	D2, D3; Appendix H
Emitted-answer mass $\rho(a_{\text{emit}})$ as the primary score	post-registration	D6; Tables 2, 3
Query-clustered tests	post-hoc (robustness)	D7; Table 22
Deduplicated Mode-B beam	post-hoc; protocol frozen before execution	D8
Answer-aggregation and fixed-answer beam controls	post-hoc; plans frozen before execution	Appendices J.4, J.5
Other analyses	post-hoc	D4; the captions of Appendices F–J
Rendered-prompt rerun	pre-registered supplementary validation	Appendix K; Table 2
Answer frequency, mode emission, key fragmentation, judge-score bins	post-hoc; plan frozen before computation	Appendices J.4, F, C, G.3

Table 4: Registration status of the reported results.

C Implementation Details Affecting Reproduction

Frozen constants and fallbacks that the main text's method section does not enumerate, disclosed for exact reproduction (all present in the shipped code and configuration):

- **Software stack.** macOS on Apple silicon (M4 Max, CPU-only, 36 GB unified memory; cf. the reproducibility statement), Python 3.11; exact library versions ship in the code supplement's `requirements.txt`.
- **Sampling temperatures and per-query call count.** Judge and answer-extraction calls run at temperature 0. Query-paraphrase generation is one batched call per query at temperature 0.7 with a fixed seed (42), cached, so it is shared across the three experimental seeds, which vary graph-walk randomness only. Design total: $1 + 32 \times 3 + 32 = 129$ LLM calls per query; the archived per-query meter counts the 128 scoring+extraction calls (main §3.3).
- **Cost accounting is an estimate.** Reported dollar costs approximate token counts as characters$/4$ and apply static list prices; no provider-metered usage was recorded. The same estimator applies to every method, so cross-method comparisons are like-for-like, but absolute dollar figures are estimates, not invoices.
- **Per-dataset sample sizes.** 300 randomly sampled queries each on HotpotQA, MuSiQue, 2WikiMHQA, and MetaQA-3hop; 500 each on WebQSP and CWQ (the main text's "300–500"); $\times$ 3 seeds throughout.
- **Judge prompt and LLR (as-sent form).** The runner's scoring template is `Question: {query} / Reasoning path: {path_text} / Does this path support the answer? Answer yes or no:` (newlines shown as /), but the archived wrapper's `.format` call supplies different placeholder names, so rendering falls back to a legacy append path: the template survives intact, braces included, and the filled fields plus a fixed yes/no suffix are appended. Every archived judge call therefore received, verbatim: `Question: {query} / Reasoning path: {path_text} / Does this path support the answer? Answer yes or no: / Question: ⟨paraphrased question⟩ / Evidence: ⟨empty⟩ / Path: ⟨path text⟩ / / Answer with exactly one word: Yes or`

`No. / Answer:`. Only the paraphrased question and the path text are inserted; no candidate answer enters the prompt, so "the answer" carries no injected referent (main §3.1). The answer-extraction prompt is rendered by the same fallback, and the ToG-style beam's judge and final-answer calls use the same wrapper and templates, so the as-sent template is identical across compared methods; the cache archives every sent prompt byte-exactly. A preregistered rerun with both templates rendered as written, the beam re-traversed under the rendered judge, is reported in Appendix K. The inserted path text is not identical, however. Ours writes every edge with a label slot, `e -[r]-> e'`, whereas the beam writes entities only, `e -> e'`. The slot carries a relation only on the Mode B side: on the Wikipedia side the walker receives no relation map, so every edge of our Mode A judge prompts reads `_undirected_` (2,119,869 edges in 460,620 prompts; census of every archived judge prompt of the primary backbone, all configurations pooled), and on the Freebase and MetaQA side the label is looked up in the stored direction only, so an edge walked against it reads `_unknown_` (30.5% of 3,125,775 Mode B edges; 23.7% on the Freebase datasets, 48.4% on MetaQA-3hop; pin `ours_path_relation_census_2026-09-19.json`). Whether this label text, and the judge clip below, change the beam's confidence ranking is tested by re-scoring its archived paths under matched conditions (Appendix J.3). The LLR is the difference of the independent "Yes"/"No" first-token log-probabilities, $\mathrm{LLR} = \log p(\text{Yes}) - \log p(\text{No})$, read from the returned top-k list; the two are not assumed to normalize to one.

- **Judge log-probability clipping and token-absence imputation.** Per-token judge log-probabilities are floored at -10 (`logprob_clip`) in our judge calls, so log-weights $\log \phi$ live in $[-10, 10]$. The ToG-style beam's judge calls pass no floor and inherit the wrapper default of -20, the API's own minimum, which can then bind only where the token-absence imputation pushes a value below it: on the beam's archived final paths the -20 floor binds on 2.5% of the calls and the -10 floor would bind on 18.5% (Appendix J.3). If a yes/no token is absent from the returned top-k logprobs, imputation precedes the floor: with both tokens absent, both log-probabilities are set to $\log 0.5 \approx -0.69$ (LLR 0); with one absent, the missing side is set to the present side's value minus 2.0. The one-sided branch is not rare: a cache-wide replay of every archived judge scoring call (all configurations, ablations included; `judge_token_absence_2026-07-18.json`) finds it firing, in the archived run, on 36.4% of the primary backbone's 1,511,580 calls (550,490, all but 86 of them with "Yes" absent): when the judge answers "No" confidently, "Yes" often falls outside the returned top-k and that call's LLR becomes the offset value -2.0, well inside the ± 10 clip. The both-absent branch never fires (0 calls); the cross-backbone Qwen check is more extreme (79.2% "Yes"-absent). Every reported number is computed under this frozen rule. Gap sensitivity (POST-HOC). A cache replay (`llr_imputation_sensitivity_2026-07-18.json`, seed 42, answered rows) resets the Yes-absent LLR from -2 to $-\delta$ for $\delta \in \{0.5, 1, 2, 4, 10\}$ and recomputes $\rho(a_{\text{emit}})$ (δ=2 reproduces the pinned score, parity $< 10^{-6}$ on every row). The confidence AUROC moves by at most 3.8 points on any dataset (2WikiMHQA 1.5, HotpotQA 2.4, WebQSP 2.1, MuSiQue 3.0, MetaQA-3hop 3.8, CWQ 3.8), smaller than every significant vs.-beam gap in the main paper (the smallest starred win is $+7.4$; the MetaQA reversal, -9.2); a larger gap raises or barely moves the AUROC (monotone rise on four datasets, flat on 2WikiMHQA, peak at δ=4 on MuSiQue), so the -2 offset is conservative rather than inflationary, and a missing-$\to$0 or drop-the-call rule moves it no further (max 3.8). Recomputed on the full main-paper basis (`llr_imputation_sensitivity_full_2026-07-19.json`: all three seeds, the (qid, seed)-matched frames behind main Table 2, paired and query-clustered bootstraps, two-sided, BH-FDR; δ=2 reproduces the pinned emitted-mass score on every matched row), the starred set (archived prompts) is invariant across all seven variants: HotpotQA, 2WikiMHQA, and CWQ remain significant under both tests, and no other dataset enters. Although the terminal median LLR is exactly -2 on 11–65% of particles (imputation touches 25–79%), the sweep bounds the effect: the usable signal is the judge's Yes/No direction, not the imputed magnitude.
- **Scoring fallback.** If all three paraphrase scoring calls fail for a path, $\log \phi := 0$ (an uninformative $\phi = 1$), leaving the particle in the ensemble rather than silently dropping it. In the archived runs this fallback never fired: every answered query of every configuration in the main result tree, the cross-backbone check, and the superseded first Qwen run recorded all

$3K$ judge log-likelihood ratios and all K extractions (main configuration: 6,597 of 6,597 answered rows at 128 metered calls; all configurations: 28,791 of 28,791; the meter counts only judge calls that returned an LLR, so a row short of $4K$ would indicate an incomplete record rather than, by itself, a path on which all three calls failed; none occurs; census pinned in the code supplement). An answer key that normalizes to the empty string (an empty extraction, or one without ASCII letters or digits) occurs in six main-configuration rows at posterior mass below 10^{-4}; the emitted answer's own key, checked directly on each emitted answer rather than inferred from the mode of the aggregated answer distribution (the two can differ, since the emitted answer is the highest-weight particle's), is empty in none of them. In the cross-backbone check the empty key occurs in two rows, in one of which (a HotpotQA query) it is both the distribution's mode and the emitted answer's key, at mass .86: the extracted answer is written in Chinese characters, which the aggregation normalizer maps to the empty string.

- **Proposal fallback.** Neighbor weights are formed from the PPR vector, with a uniform fallback only when the total PPR mass over all neighbors of the current vertex is zero; visited vertices are then removed and the remaining weights renormalized. If no unvisited neighbor has positive mass at that stage, the deployed implementation stops extending the path (the particle keeps its path so far) rather than falling back to uniform; an earlier version of this item described a uniform fallback at that stage, which the code does not implement.
- **Answer equivalence.** Answer aggregation and Hit@1 both lowercase, drop articles, and collapse whitespace, but they are not the same normalizer: aggregation replaces every character outside ASCII letters, digits, and whitespace by a space, so $\rho(a)$ sums weights over strings such as `j k rowling`, whereas the Hit@1 scorer deletes ASCII punctuation without inserting a space and keeps non-ASCII letters (`jk rowling`). Re-scoring both methods' archived final answers with the aggregation normalizer, confidences held fixed, flips the Hit@1 label on 8 of the 6,600 rows for ours and 12 for the beam and moves a per-dataset AUROC by at most 0.29 (ours) and 0.62 (beam) points (cache replay, no new calls; pin `normalizer_gap_2026-09-19.json`); the registered scorer is unchanged. The scorer credits two or more posterior keys on 2.6% to 68.1% of answered rows per dataset (fragmentation as seen by the scorer, not entity identity); merging them would change the posterior mode on 53 of 6,597 answered rows. Per dataset: Table 8 (post-hoc, archived prompts).
- **Answer-path determinism (ablation).** Disabling the abstention gate or the $\log \hat{Z}_q$ diagnostic changes no answer: Hit@1 is identical on all six datasets (Δ=0.000 per dataset in the aggregated results); both components consume the sampler's terminal state without altering the traversal or the answer aggregation.
- **Dissection labels.** The beam-confidence dissection and its depth-adjusted repair control (main §5.1) label correctness by strict exact match computed analysis-locally, which can differ marginally from the aggregate Hit@1 labels used elsewhere; all other analyses use Hit@1. On 2WikiMHQA the beam's incorrect answers are more confident on average than its correct ones (.40 vs. .30), which produces its below-chance AUROC; forcing full depth-3 traversal closes only 18% of the gap on HotpotQA (clustered n.s.).
- **MDE, exactly.** The 80%-power minimum detectable effects quoted in main §5.4 (4.8–6.6 AUROC points) are one-sided at α=.05: $(z_{.95} + z_{.80}) \approx 2.49$ times the paired-bootstrap standard error of the per-dataset C1 AUROC delta (B=1000, matched (query, seed) pairs; SEs 1.9–2.7). Being paired-based, they understate a clustered-basis MDE (D7).
- **Conformal quantile, exactly.** On the calibration split the correct-answer ESS values are sorted ascending and $\tau_{\text{conformal}}$ is the value at zero-based index $\lfloor \alpha\,(n_c + 1) \rfloor - 1$ (clamped to ≥ 0), where n_c is the number of correct calibration queries. ESS ties do occur in practice (the clipped judge log-probabilities discretize the weights); the abstention rule is the strict inequality $\text{ESS} < \tau$, so a test value tying the threshold falls on the answer side; this can only lower the abstention rate, leaving the one-sided bound intact. If $n_c < \lceil 1/\alpha \rceil$ (= 10 at α=0.1), the target α is unresolvable at that calibration size under this rule, so the data-dependent threshold is conservatively disabled ($\tau_{\text{conformal}} := 0$) and only the ESS floor below remains active; this is the degenerate regime disclosed for MuSiQue (9/30 splits, Table 13). Because n_c is itself random (the number of correct queries in the calibration split), the bound is read conditionally: given n_c, the n_c correct calibration ESS values and

a correct test query's ESS are exchangeable under the stated assumption, so $\mathbb{P}(\text{ESS} < \tau \mid \text{correct}, n_c) \leq \lfloor \alpha(n_c+1) \rfloor / (n_c+1) \leq \alpha$ for $n_c \geq \lceil 1/\alpha \rceil$, and 0 below it (threshold disabled; the floor below never discards a correct answer), hence also marginally over n_c. ESS is the registered abstention statistic (label-free sampler dispersion); a plug-in threshold on $\rho(a_{\text{emit}})$ itself, without a distribution-free guarantee, is Table 28.

- **Abstention threshold floor.** The deployed rule is $\tau_{\text{abstain}} = \max(\tau_{\text{conformal}}, 1)$: a query abstains when terminal ESS drops below 1 even if the conformal quantile is lower. Under valid normalized weights ESS ≥ 1 always holds, so the floor is inert there; it can bind only when the walk yields no scored particles at all (e.g., seed-entity extraction or PPR failure; a scoring-call failure instead falls back to ϕ=1 and keeps its particle, per the fallback above). A correct answer implies at least one finite-weight particle (ESS ≥ 1), so the floor never discards a correct answer and the certified direction is unaffected. In the archived main run this no-particle path fired on 3 of the 6,600 query–seed rows (one CWQ query on all three seeds: PPR failure before the walk, zero LLM calls, counted as abstained).
- **Aggregation unit for the conformal check.** The 30-split held-out check (Table 13) operates at query level: per-query ESS averaged over the three seeds, correctness by seed-majority; seeds of the same query never straddle the calibration/test boundary; exchangeability is assumed at the query level. The deployed system, however, is a single 129-call run, so we also replay the identical protocol at that unit (POST-HOC): each LLM seed validated separately as its own single-run system, same splits and calibration rule, no aggregation. The picture is unchanged: held-out false abstention averages .054–.112 across the 18 dataset–seed pairs, above α=.1 on 2 of 18, by at most $+.012$.
- **Provider logprobs shapes (cross-backbone check).** Together returns token logprobs in the legacy Completions shape (`content` null, top-level `top_logprobs` maps); a reader expecting only the modern OpenAI shape records none, and every judge LLR silently falls back to 0 for both methods. Our first Qwen run hit exactly this: with all weights uniform, our confidence degenerated to the answer-frequency vote share, which still ranked correctness at 83.45/79.84 top-1 AUROC (HotpotQA/2WikiMHQA), while the beam's single-path score collapsed to a constant (AUROC exactly 50.0). The repaired judge-weighted rerun (quarantined scoring cache, fixed reader) gives 84.82/78.86, with the symmetric Qwen beam at 71.35/52.48; the main text cites only the repaired numbers. The accident doubles as an unplanned ablation: with the judge signal removed entirely, the K=32 ensemble's confidence moves about a point (down on HotpotQA, up on 2WikiMHQA) while the single-path beam score fails catastrophically. GPT-4.1-mini responses use the modern shape throughout; the primary results are unaffected.

D Pipeline Diagram and a Worked Example

Both items respond to external pre-submission feedback on presentation (a system block diagram, Figure 1, placed in the main text; a concrete with/without example). The pipeline has three outputs with separate roles: the confidence $\rho(a_{\text{emit}})$, which ranks answers (§5.1); the auxiliary evidence signal $\log \hat{Z}_q$ (§3.4); and ESS, used only by the abstention gate (§3.4). The example is drawn verbatim from the archived per-query artifacts; its selection criterion and the symmetric count are disclosed below.

D.1 Worked example: score saturation (2WikiMHQA, seed 42)

Setup. Archived seed-42 rows, 2WikiMHQA (300 queries). Selected post-hoc from the 4 queries where the beam is confident ($\geq .9$) and wrong while the sampler is right, as the clearest single-relation illustration; the symmetric count is reported at the end. On this query all ten final beam paths are distinct, so what follows is a score-saturation failure, not a beam-width artifact.

Question (qid `53e18b8c...`): "Who is Ailwyn Fellowes, 1st Baron Ailwyn's paternal grandfather?" Gold: William Henry Fellowes.

Beam (ToG-style). The top-scoring final path passes through the gold entity and terminates one hop past it,

AILWYN FELLOWES, 1ST BARON AILWYN → WILLIAM HENRY FELLOWES → EDWARD FELLOWES (LLR = 12.5),

and answer extraction on that path returns its terminal entity, Edward Fellowes, the father, not the asked-for grandfather. The confidence proxy saturates: $\sigma(\mathrm{LLR}) > .999$.

Sampler (ours). Across the $K{=}32$ independent walks the judge-weighted answer posterior concentrates on the gold entity: mass .998 on william henry fellowes vs. .002 on edward fellowes. The emitted answer is correct, with confidence $\rho(a_{\mathrm{emit}}) = .998$.

The mechanism is distributional, not anecdotal. On the same seed-42 frame the beam's $\sigma(\mathrm{LLR})$ saturates ($\geq .999$) on 53 of 300 queries, and accuracy within that maximally confident subset is .377 (essentially the .370 base rate): the beam's top confidence grade carries no information about correctness. The sampler's posterior saturates on 104 of 300 queries with subset accuracy .663 against a .383 base rate. This is the per-query mechanism behind the pooled 46.4-vs-77.3 AUROC cell (main Table 2). For symmetry: at the $\geq .9$ threshold the anecdote counts are balanced (4 confident-wrong-beam / right-sampler queries vs. 4 the other way): the advantage is the ranking, not cherry-picked stories.

E Risk–Coverage Curves: All Six Datasets

Figure 3 extends main-text Figure 2 to all six datasets under identical conventions (shared y-axis; ours scored by $\rho(a_{\mathrm{emit}})$, shaded band = per-seed min–max over 3 seeds). Two panels are omitted from the main text, MuSiQue and CWQ; the other four appear there. Where the beam leads: on MuSiQue it is ahead at low coverage; on WebQSP its higher raw accuracy carries its curve (.733 vs. .671 at 50%, main §5.2); on MetaQA-3hop it dominates at every coverage, the cardinality failure reported in main §5.1, disclosed in full. On CWQ ours leads on AURC (.439 vs. .465, main §5.2).

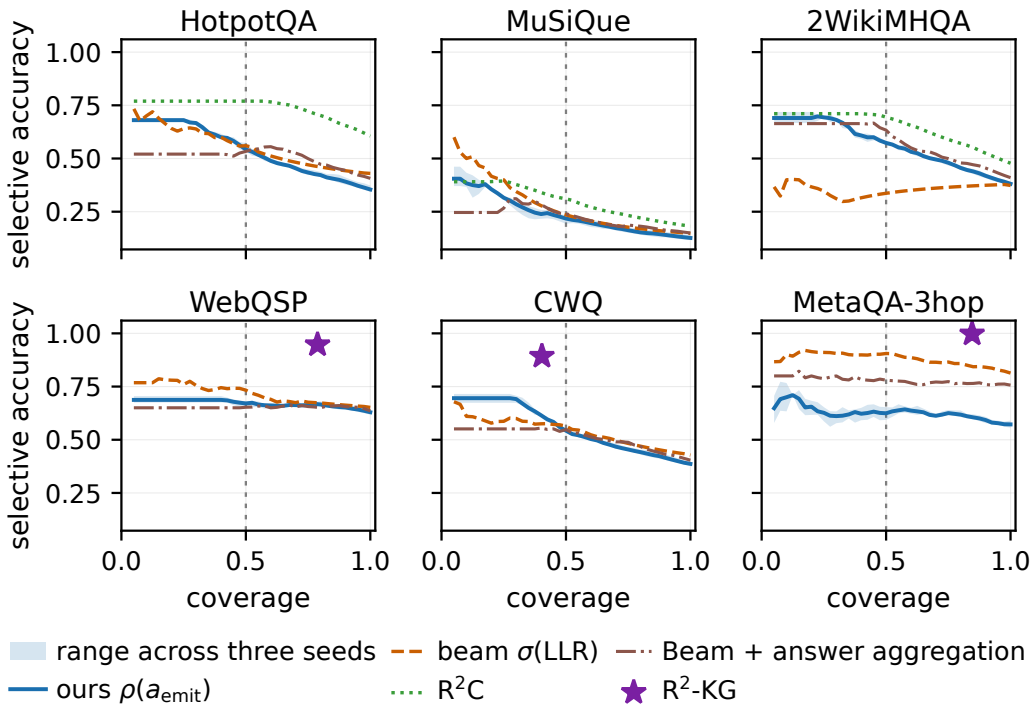


Figure 3: Selective accuracy vs. coverage, all six datasets. Conventions as in main-text Figure 2: each method ranks its own answers by its own confidence; R^2C runs on the Wikipedia side only (it reads the corpus); R^2-KG is a single fixed operating point (star); the dash-dot curve is the beam with answer aggregation (post-hoc; Appendix J.4; its mass saturates at 1.0 on 13–49% of rows per panel, prorated as ties). Vertical line: 50% coverage. Curves visualize the pre-registered AURC metric and are drawn with the prorated tie rule, whereas the reported AURC uses the block rule (Appendix J.1); crossovers are post-hoc.

F Emitted Answer and Confidence: Alignment (Post-Hoc)

The emitted answer is the highest-weight sampled particle's answer. This selection rule is distinct from the mode of the aggregated answer posterior and is not interpreted as a MAP estimator over path space, while the registered confidence proxy is the top-1 mass of the answer posterior, $\max_a \rho(a)$ (§3.2 of the main text); the two can designate different answers: the semantic gap that motivated the emitted-mass primary (D6). Table 5 quantifies this from the archived per-query records (cache replay, zero additional LLM calls; post-hoc). They coincide on 81.5% of answered queries (mismatches: 1,218 of 6,597). Divergence marks a dispersed posterior: on mismatched queries both the mean confidence and Hit@1 are far lower. On the answer side the choice is a wash: emitting $\arg\max_a \rho(a)$ instead would change Hit@1 on the mismatched queries from 207 to 224 correct ($+17$ of 6,597 answered, <0.3pp overall; better on 4/6 datasets, worse on 2WikiMHQA 28 $\rightarrow$ 17 and MetaQA-3hop 67 $\rightarrow$ 53), so the shipped design keeps the registered highest-weight-particle answer rather than adopt a post-hoc swap. On the confidence side, scoring the emitted answer's own mass $\rho(a_{\text{emit}})$ leaves the replayed answered-only per-seed-mean AUROC essentially unchanged (70.5 $\rightarrow$ 70.6; largest per-dataset moves MuSiQue $+1.0$, 2WikiMHQA -0.5), and it is the score whose semantics match the answer actually returned: per D6 it is the paper's primary, and the vs.-beam significance structure survives the switch (main Table 2). Basis note: this replay is an answered-only, per-seed-mean quantity; main Table 2 reports the pooled AUROC; the two bases must not be mixed (per-dataset cells differ, e.g., MuSiQue 79.4 here vs. 79.2 pooled). The replayed mode-mass column reproduces the registered variant (Table 3) within rounding (± 0.1).

		Hit@1 (%)		mean conf.		
Dataset	mism. %	align.	mism.	align.	mism.	AUROC†
HotpotQA	21.0	41.6	12.2	.86	.48	77.3 / 77.5
MuSiQue	35.1	18.0	2.8	.72	.41	78.4 / 79.4
2WikiMHQA	21.3	44.5	14.6	.83	.46	77.9 / 77.3
WebQSP	7.1	65.0	35.8	.93	.57	57.0 / 57.3
CWQ	16.7	43.1	16.8	.90	.51	73.4 / 73.3
MetaQA-3hop	18.3	61.0	40.6	.79	.42	58.8 / 58.9

Table 5: Emitted-answer vs. posterior-mode alignment per dataset (our default configuration, answered queries, 3 seeds pooled for counts). †per-seed AUROC averaged over seeds: mode mass / $\rho(a_{\text{emit}})$ as the confidence score; the $\rho(a_{\text{emit}})$ column is the shipped primary (D6). Post-hoc; raw: `answer_confidence_alignment_2026-07-17.json`.

Emitting the posterior mode, a what-if (post-hoc, archived prompts). The analyses above keep the emitted answer. Tables 6–7 score, on the frame of Table 2, a rule that would emit the posterior mode $\arg\max_a \rho(a)$ with confidence $\max_a \rho(a)$ (the largest stored mass; exact ties to the first stored key). The mode is known only as a normalized key, so the emitted answer is also scored on its key, and the rule contrast (T2) uses key scoring on both sides. The method's emitted answer and every number above are unchanged. The plan, including the readings and their wording, was frozen before any computation (shipped as `2026-09-25_answer_frequency_and_judge_bins_posthoc_anonymized.yaml`). Hit@1 of the mode rule minus the emitted answer as archived (points): HotpotQA -1.1, MuSiQue $+1.0$, 2WikiMHQA -6.3, WebQSP -0.5, CWQ $+0.2$, MetaQA-3hop -1.9; these differences, which the main text quotes, combine the change of rule with the change from the archived string to its key: against the emitted answer scored on its key, the rule change alone is HotpotQA $+1.0$, MuSiQue $+1.2$, 2WikiMHQA -1.0, WebQSP $+0.8$, CWQ $+0.9$, MetaQA-3hop -1.6. Against the beam (T1): advantage established on HotpotQA, 2WikiMHQA; not established on MuSiQue, WebQSP, CWQ, MetaQA-3hop. Near-tied mode rows (another key within 10^{-12} of the largest mass) whose near-tied keys carry different labels: HotpotQA 0/8, MuSiQue 1/22, 2WikiMHQA 0/8, WebQSP 0/2, CWQ 0/7, MetaQA-3hop 8/34.

Dataset	Rule	Hit@1	AUROC	AURC	SA@50	Hit@1_{mm}	Mode − emit (key) [95% CI]	p_{clu}
HotpotQA	Emit	.354	77.4	.455	.542	.122		
n_{mm}=189	Emit (key)	.333	77.5	.479	.518	.116		
	Mode	.343	76.2	.475	.520	.164	-1.3 [-3.5, $+0.9$]	.247
	Beam	.430	65.6	.458	.560			
MuSiQue	Emit	.127	79.2	.758	.218	.028		
n_{mm}=316	Emit (key)	.124	79.7	.760	.216	.025		
	Mode	.137	75.7	.754	.220	.060	-4.0 [-8.3, $+0.1$]	.055
	Beam	.148	77.2	.744	.235			
2WikiMHQA	Emit	.381	77.3	.431	.573	.146		
n_{mm}=192	Emit (key)	.328	73.8	.527	.489	.135		
	Mode	.318	75.7	.529	.487	.089	$+1.9$ [-0.7, $+4.8$]	.151
	Beam	.373	46.4	.631	.337			
WebQSP	Emit	.629	57.3	.328	.671	.358		
n_{mm}=106	Emit (key)	.616	58.3	.335	.667	.349		
	Mode	.624	56.6	.335	.667	.462	-1.8 [-3.1, -0.5]	.003*
	Beam	.652	59.3	.284	.733			
CWQ	Emit	.387	73.3	.439	.543	.168		
n_{mm}=250	Emit (key)	.380	73.4	.445	.531	.168		
	Mode	.389	71.9	.444	.531	.220	-1.5 [-3.5, $+0.6$]	.147
	Beam	.428	65.9	.465	.564			
MetaQA-3hop	Emit	.572	58.8	.371	.624	.406		
n_{mm}=165	Emit (key)	.569	58.8	.374	.622	.406		
	Mode	.553	61.1	.376	.622	.321	$+2.3$ [-0.5, $+5.1$]	.105
	Beam	.813	68.0	.131	.905			

Table 6: Posterior-mode emission, a what-if (post-hoc, archived prompts), on the frame of Table 2, pooled over the three seeds (per-seed values in the pin). Emit = the emitted answer and its mass as archived (the paper's numbers); Emit (key) = the same answer scored on its normalized key; Mode = the answer with the largest posterior mass, with that mass as confidence (scored on its key); Beam = the Table 2 beam. AUROC $\times 100$; AURC and SA@50 (selective accuracy at 50% coverage) tie-safe. Hit@1_{mm} = Hit@1 on the n_{mm} rows whose mode key differs from the emitted key, the only subgroup analysed (n/a when n_{mm}=0). Last two columns (T2): AUROC of the mode rule minus the key-scored emitted answer, query-clustered bootstrap (B=2000, add-one, two-sided); *survives BH-FDR (q=.05) over the six datasets; n/a = undefined comparison. Pin: `review_posthoc_z1z2z3_2026-09-25.json`.

Dataset	Seed	n	Mode	Beam	Δ [95% CI]	p_{pair}	p_{clu}	Reading
HotpotQA	42	300	75.8	65.6	$+10.1$ [$+2.2$, $+18.7$]	.016		
	123	300	77.0	65.6	$+11.4$ [$+3.5$, $+20.0$]	.004		
	456	300	75.9	65.6	$+10.2$ [$+2.5$, $+18.1$]	.006		
	pooled	900	76.2	65.6	$+10.6$ [$+2.9$, $+18.2$]		.007	(i)
MuSiQue	42	300	75.8	76.6	-0.8 [-10.3, $+8.4$]	.828		
	123	300	76.1	77.5	-1.4 [-10.6, $+7.6$]	.766		
	456	300	75.1	77.5	-2.4 [-12.9, $+8.6$]	.688		
	pooled	900	75.7	77.2	-1.5 [-10.5, $+7.5$]		.758	(iv)
2WikiMHQA	42	300	74.1	46.6	$+27.5$ [$+19.2$, $+36.4$]	$<.002$		
	123	300	76.9	46.2	$+30.7$ [$+23.4$, $+38.9$]	$<.002$		
	456	300	76.2	46.4	$+29.8$ [$+22.1$, $+38.0$]	$<.002$		
	pooled	900	75.7	46.4	$+29.3$ [$+21.7$, $+37.1$]		.001	(i)
WebQSP	42	500	57.9	59.3	-1.4 [-8.9, $+5.6$]	.764		
	123	500	54.4	59.3	-4.9 [-12.6, $+2.0$]	.158		
	456	500	57.5	59.3	-1.8 [-9.0, $+5.3$]	.618		
	pooled	1500	56.6	59.3	-2.7 [-9.6, $+3.8$]		.441	(iv)
CWQ	42	500	73.1	65.9	$+7.1$ [$+1.2$, $+12.6$]	.022		
	123	500	72.8	65.9	$+6.9$ [$+1.8$, $+12.5$]	.008		
	456	500	69.9	65.9	$+4.0$ [-1.6, $+9.6$]	.152		
	pooled	1500	71.9	65.9	$+5.9$ [$+0.8$, $+11.3$]		.029	(iv)
MetaQA-3hop	42	300	62.4	68.0	-5.5 [-15.8, $+5.0$]	.274		
	123	300	62.8	68.0	-5.2 [-15.7, $+6.3$]	.346		
	456	300	58.1	68.0	-9.9 [-20.1, $+0.9$]	.066		
	pooled	900	61.1	68.0	-6.9 [-16.3, $+3.3$]		.169	(iv)

Table 7: Posterior-mode emission against the beam (T1; post-hoc, archived prompts): confidence AUROC ($\times 100$) of the mode rule and of the Table 2 beam, Δ = mode $-$ beam. Seed rows: query-level paired bootstrap (B=1000, two-sided, no add-one; an exact zero prints as its bound). Pooled rows: query-clustered bootstrap (B=2000, add-one, two-sided; it carries the reading). Readings (i) advantage established, (iii) reversal, (iv) not established, after BH-FDR (q=.05) over the six datasets; no equivalence test was registered. Pin: `review_posthoc_z1z2z3_2026-09-25.json`.

Dataset	n_{ans}	Unsc.	$\lvert G\rvert$=0	$\lvert G\rvert$=1	$\lvert G\rvert$=2	$\lvert G\rvert\geq 3$	Merge flips mode	Mass$_G$	Max	Max share	Gain (n)
HotpotQA	900	0	50.9	45.0	3.3	0.8	3 (0.3)	.816	.633	.783	.358 (30)
MuSiQue	900	0	75.1	22.3	2.4	0.1	3 (0.3)	.418	.374	.845	.189 (10)
2WikiMHQA	900	0	50.2	46.4	3.0	0.3	0 (0.0)	.550	.514	.829	.063 (16)
WebQSP	1,500	0	26.9	42.1	11.5	19.5	1 (0.1)	.848	.732	.856	.157 (395)
CWQ	1,497	0	39.1	49.1	5.7	6.1	1 (0.1)	.590	.483	.808	.236 (100)
MetaQA-3hop	900	0	10.6	21.3	9.8	58.3	45 (5.0)	.689	.535	.734	.212 (433)

Table 8: Answer-key fragmentation as seen by the scorer (post-hoc, archived prompts), pooled over the three seeds (per-seed values in the pin). G = the posterior answer keys the paper's Hit@1 scorer credits; contains-matching can credit different entities sharing the gold string and miss aliases, so $\lvert G\rvert\geq 2$ is fragmentation as seen by the scorer, not entity identity. Shares (%) are over the n_{ans} answered rows; Unsc. = answered rows without a fragmentation reading (empty answer or no distribution), included in n_{ans}. Merge flips mode = rows whose mode key is not in G while the summed credited mass exceeds the mode's mass (count and share). Over rows with $\lvert G\rvert\geq 2$: Mass$_G$ = mean summed credited mass, Max = mean largest credited mass, Max share = mean share of the credited mass held by the largest credited key (rows with positive credited mass; 0 fragmented rows with zero credited mass). Gain = mean credited mass left out of the emitted confidence over the n rows whose emitted key is in G with $\lvert G\rvert\geq 2$. Pin: `review_posthoc_z1z2z3_2026-09-25.json`.

G CONFIDENCE ROBUSTNESS AND FAILURE-MODE DIAGNOSTICS (POST-HOC)

High-confidence errors are not stratified by failure type (graph evidence, judge, extraction, alias, neighbourhood, path length); Appendix G.3 separates judge score from extraction only; Appendix D.1 gives one worked example and Appendix G.2 one dataset-level mechanism.

G.1 SENSITIVITY TO THE PARTICLE WEIGHT TRANSFORM

The sampler weights completed paths by $\exp(l^{(k)})$, where $l^{(k)} = \log \phi^{(k)}$ is the judge's median LLR (main §3.1); the exponent is a modeling choice, so we replay the archived particle states (single seed 42, cache-only, zero new LLM calls) and re-rank the same emitted answers under transformed weights: the temperature family $w^{(k)} \propto \exp(\beta l^{(k)})$, $\beta \in \{0.25, 0.5, 1, 2\}$; the bounded transform $w^{(k)} \propto \sigma(l^{(k)})$, which is the normalized p_{yes} variant exactly, since $p_{\text{yes}} = \sigma(l^{(k)}) = p(\text{Yes})/[p(\text{Yes})+p(\text{No})]$ (the Yes/No-renormalized score, not raw $p(\text{Yes})$) and the median over an odd number of paraphrases commutes with the monotone σ; and scale-free reciprocal-rank weights $w^{(k)} \propto 1/\text{rank}(l^{(k)})$. Emission is held fixed (the argmax particle is β-invariant for $\beta > 0$; fallback-vote emissions are counted in the pin and likewise held fixed), so only the confidence ranking moves. Replay fidelity is gated per row: across all 2,199 replayed rows there were zero archived-vs-replay mismatches, and β=1 reproduces the archived emitted-mass score within 10^{-6} on every answered row. Table 9 reports answered-only pooled AUROC on a single seed (42); this basis is narrower than main Table 2's (three seeds, matched frames), so cells must not be compared across the two tables.

Against the disclosure rule fixed before observation (a 2pt corridor around β=1; any exceedance is reported per cell, and a blanket robustness claim is then disallowed), the result is two-sided. Within the temperature family, four of six datasets move by at most 1.90 AUROC points across the full β range, and sharpening to β=2 costs at most 0.13 anywhere; but flattening to β=0.25 costs -2.90 on MuSiQue and -5.43 on MetaQA-3hop. Discarding magnitudes outright degrades further: $\sigma(\text{LLR})$ loses up to -3.68 (MuSiQue) and reciprocal-rank up to -6.77 (MetaQA-3hop), with the corridor exceeded on 4/6 datasets. We therefore make no blanket transform-robustness claim; the honest summary is directional: the ranking tolerates sharpening but not flattening, i.e., the magnitudes of the judge LLRs, not merely the particle order, carry the confidence signal. The registered per-token logprob clip (± 10) is not inert: 33.07%–74.92% of individual paraphrase LLRs sit at the clip boundary per dataset (Table 9, last column). A full-pipeline cache replay that refloors the cached raw token logprobs (`llr_clip_sensitivity_2026-07-19.json`, seed 42, answered rows; the deployed -10 reproduces every pinned emitted-mass score within 10^{-6}) sweeps the floor to -5 and -20 (the latter the logprobs-API minimum, i.e., as wide as the raw data allow) and moves the confidence AUROC by at most 1.94 points on any dataset (MetaQA-3hop; all others ≤ 1.34); the β sweep covers only a global rescale of the exponent, the clip replay the floor itself.

	$w \propto \exp(\beta l)$						clip-bnd.
Dataset	β=0.25	0.5	1	2	$\sigma(l)$	1/rank	(%)
HotpotQA	76.10	77.05	77.82	78.21	76.33	74.49	39.34
MuSiQue	74.19	76.71	77.09	77.96	73.41	71.99	33.07
2WikiMHQA	76.87	76.84	76.91	76.86	75.76	75.92	44.66
WebQSP	57.64	58.83	59.54	59.65	57.38	56.08	64.03
CWQ	74.64	74.83	74.94	74.81	73.42	73.58	64.84
MetaQA-3hop	56.36	59.25	61.79	63.68	60.68	55.02	74.92

Table 9: Confidence AUROC ($\times 100$) under weight transforms (post-hoc): answered-only pooled, seed 42, emission held fixed. The β=1 column is the shipped $\rho(a_{\text{emit}})$ on this replay basis (per-row parity with the archived scores within 10^{-6}); values are not comparable to main Table 2 (three seeds, matched frames). Last column: share of individual paraphrase LLRs at the ± 10 clip boundary. Raw: `llr_scale_sensitivity_2026-07-17.json`.

G.2 WHY WEBQSP: THE POSTERIOR SATURATES ON WRONG ANSWERS TOO

Main §5 leaves WebQSP's weak confidence (below the pre-registered 65 floor) with only the KG-conditional label, while MetaQA-3hop's reversal has a cardinality diagnosis. The archived posteriors

supply a candidate mechanism for WebQSP at zero cost (no replay; every AUROC in this subsection is on the three-seed pooled basis, parity-gated to the pinned values with Δ=0.0000 on all six datasets): saturation on incorrect answers. On WebQSP, 59.17% of incorrect answered rows place $\rho(a_{\text{emit}}) \geq .99$, versus 12.72%–25.65% on the three Wikipedia datasets, so the correct-vs-incorrect mean-mass gap collapses to .07 (vs. .27/.29/.35 on HotpotQA/2WikiMHQA/MuSiQue). With 1–2-hop questions and a collapsed candidate space (median 2 distinct answers in the terminal posterior; median terminal ESS 4.2 of 32 particles), the walkers reach consensus whether or not the consensus answer is right, and the emitted-answer mass retains almost no dynamic range to rank with. The neighboring cells triangulate the mechanism rather than contradict it: MetaQA-3hop saturates least on incorrect rows (23.90%) and is hurt instead by mass splitting over a wide answer space (median 13 distinct answers; the cardinality analysis of main §5.1), while CWQ saturates heavily on correct rows (70.69%) but far less on incorrect ones (39.37%), and its AUROC survives at 73.3. It is wrong-answer saturation, rather than saturation per se, that destroys the ranking. This is a descriptive diagnosis of the archived runs, not a causal claim; the paper's KG-conditional framing stands. Raw: `webqsp_freebase_diag_2026-07-17.json`.

Dataset	Ours per-seed	mean $\pm$ std	Beam per-seed	mean $\pm$ std
HotpotQA	77.8/75.6/79.0	77.48 $\pm$ 1.71	65.6/65.6/65.6	65.64 $\pm$ 0.00
MuSiQue	77.1/81.3/79.8	79.40 $\pm$ 2.13	76.6/77.5/77.5	77.22 $\pm$ 0.53
2WikiMHQA	76.9/76.3/78.7	77.33 $\pm$ 1.25	46.6/46.2/46.4	46.38 $\pm$ 0.21
WebQSP	59.5/57.0/55.5	57.33 $\pm$ 2.05	59.3/59.3/59.3	59.27 $\pm$ 0.00
CWQ	75.0/73.0/72.0	73.35 $\pm$ 1.54	65.9/65.9/65.9	65.93 $\pm$ 0.00
MetaQA-3hop	61.8/60.4/54.5	58.90 $\pm$ 3.87	68.0/68.0/68.0	67.99 $\pm$ 0.00

Table 10: Per-seed own-correctness AUROC (seeds 42/123/456), ours (emitted-answer mass $\rho(a_{\text{emit}})$, the paper's primary; supplement D6) vs. ToG-style beam (σ(top-path LLR)). Per-seed detail supplements Table 2 of the main paper; the registered top-1–mass variant is preserved in the D6 companion table. Mode-B beam cells (WebQSP/CWQ/MetaQA-3hop) are the protocol-frozen deduplicated rerun (main §4's width disclosure); its expansion is seed-invariant, hence std 0. The mean of the three per-seed values can differ from main Table 2's pooled value at the first decimal.

Dataset	Hit@1$_{ours}$	Hit@1$_{HR2}$	EM$_{ours}$	EM$_{HR2}$	p
HotpotQA	0.354	0.690	0.313	0.480	$<$.001
MuSiQue	0.127	0.353	0.108	0.238	$<$.001
2WikiMHQA	0.381	0.661	0.363	0.401	.038

Table 11: HippoRAG2 corpus-reading upper reference (Mode A, 3 seeds, 300 queries each; matched pairs; p = one-sided paired bootstrap for HippoRAG2 better on EM). The pre-registered secondary EM-parity criterion (C3) is not met on any dataset; the pre-registered cost-ratio falsification (FC3, $>30\times$) is not triggered (our total cost is 0.44$\times$ HippoRAG2's, on the registered basis: HippoRAG2's corpus indexing counted, our KG indexing pre-built and excluded: an asymmetric basis, so this answers only the registered falsification check and is not a cost-efficiency claim). Expected information asymmetry, disclosed per protocol: HippoRAG2 reads source passages; ours is KG-only.

Dataset	R^{2}C budget	Ours	R^{2}C	Δ	95% CI	one-sided p
HotpotQA	B=40	77.28	74.25	+3.03	[-1.31, 7.11]	.079
MuSiQue	B=40	78.34	76.76	+1.59	[-3.76, 6.61]	.294
2WikiMHQA	B=40	77.88	77.40	+0.48	[-3.40, 4.15]	.405

Table 12: Call-matched control detail (Δ = ours $-$ R^{2}C; 3-seed pooled matched pairs, tie-safe rank basis, registered top-1 mass; the B=10 comparison is Table 34). The B=40 run quadruples R^{2}C's perturbation budget to match our fixed call budget (128 scoring+extraction calls/query). Under the emitted-mass primary the deltas are $+3.19/+2.48/-0.10$ (p=.055/.184/.526; main §5.4). Pre-registered as the cost-matched control for the C1 comparison; the 3-seed 2WikiMHQA paired test cited in main §5.4 is post-hoc backfill.

Dataset	FA mean	min	max	splits $\leq \alpha$	degen.	recall
HotpotQA	0.089	0.000	0.265	20/30	0	0.040
MuSiQue	0.047	0.000	0.292	26/30	9	0.014
2WikiMHQA	0.083	0.000	0.244	22/30	0	0.095
WebQSP	0.089	0.049	0.181	21/30	0	0.038
CWQ	0.101	0.023	0.205	17/30	0	0.071
MetaQA-3hop	0.101	0.034	0.209	17/30	0	0.038

Table 13: Conformal abstention, held-out check over 30 random 30/70 calibration/test splits (target α=.1): false abstention P(abstain | correct) stays at or near α, while abstention recall P(abstain | incorrect) is only .01–.10: the rule certifies safety; it does not filter errors (disclosed in Limitations). Degenerate = splits with fewer than 10 correct calibration answers (τ floors to 0; MuSiQue). POST-HOC diagnostic of the pre-registered rule.

Gold answers	n	Ours Hit@1	Beam Hit@1	Ours AUROC	Beam AUROC
1	141	0.582	0.830	62.1	60.7
2–5	303	0.568	0.792	64.4	61.5
6–20	270	0.596	0.889	58.8	75.7
21+	186	0.538	0.726	47.8	75.2

Table 14: MetaQA-3hop by gold-answer cardinality: our emitted-mass confidence collapses as the legitimate answer set grows (below chance at 21+), while the beam's does not; this is the mechanism behind the MetaQA-3hop row of main Table 2. POST-HOC decomposition. Beam column = the deduplicated rerun (main §4).

Dataset	n	Ours $\rho(a_{\text{emit}})$	P(True)	Δ	95% CI	p
HotpotQA	900	77.44	73.24	+4.20	[0.26, 8.02]	.016
2WikiMHQA	900	77.30	58.99	+18.30	[14.29, 22.62]	$<$.001

Table 15: Verbalized P(True) self-check vs. our emitted-answer mass $\rho(a_{\text{emit}})$ (paper primary) on identical final answers (same-answer design: only the confidence signal differs). POST-HOC baseline (protocol of Kadavath et al. 2022).

Dataset	K	AUROC	Hit@1	$/query	K=32 advantage
HotpotQA	8	73.63	0.347	$0.0015	+4.85 (p=.035)
HotpotQA	16	73.18	0.377	$0.0030	+5.31 (p=.009)
HotpotQA	32	78.48	0.347	$0.0060	
2WikiMHQA	8	74.74	0.367	$0.0014	+2.76 (p=.116)
2WikiMHQA	16	77.85	0.363	$0.0029	-0.35 (p=.580)
2WikiMHQA	32	77.50	0.383	$0.0057	

Table 16: K-sensitivity sweep (seed 42, tie-safe rank basis): accuracy is flat in K; confidence quality rises with K on HotpotQA and saturates by K=16 on 2WikiMHQA; cost is linear in K. K=32 advantage = K=32 minus that row, paired bootstrap. POST-HOC (not pre-registered).

Dataset	Ours AUPRC (base)	Beam AUPRC (base)	Random AURC floor
HotpotQA	61.5 (35.4)	58.0 (43.0)	0.644 ± 0.016
MuSiQue	32.8 (12.7)	43.6 (14.8)	0.873 ± 0.011
2WikiMHQA	63.0 (38.1)	37.6 (37.3)	0.617 ± 0.016
WebQSP	67.4 (62.9)	72.2 (65.2)	0.370 ± 0.012
CWQ	61.6 (38.7)	56.1 (42.8)	0.613 ± 0.013
MetaQA-3hop	63.3 (57.2)	87.6 (81.3)	0.427 ± 0.017

Table 17: AUPRC with base rates in parentheses (a threshold-free complement to AUROC; we win the same three datasets as the AUROC verdict, HotpotQA, 2WikiMHQA, and CWQ, with base rates lower on HotpotQA and CWQ and near-equal, +0.8pp, on 2WikiMHQA), and the random-ranking AURC floor (mean $\pm$ std over 1,000 shuffles). Mode-B beam cells and base rates use the deduplicated rerun (D8). POST-HOC.

Dataset	Ours	Beam	Δ	95% CI	p
HotpotQA	0.177	0.250	-0.073	[-0.107, -0.042]	$<.001$
MuSiQue	0.146	0.174	-0.028	[-0.059, 0.003]	.043
2WikiMHQA	0.179	0.372	-0.193	[-0.227, -0.154]	$<.001$
WebQSP	0.249	0.214	+0.034	[0.009, 0.062]	.995
CWQ	0.193	0.256	-0.063	[-0.085, -0.040]	$<.001$
MetaQA-3hop	0.263	0.112	+0.151	[0.110, 0.193]	$>.999$

Table 18: Excess AURC (AURC minus the same-accuracy oracle ranking's AURC; lower is better): removing the base-error confound that raw AURC compounds aligns with the AUROC verdict of main Table 2: the three significant AUROC wins, two losses (WebQSP, MetaQA-3hop), and a nominal MuSiQue win (p=.043, block-end ties) that reverses under the prorated tie rule and is not counted (Appendix J.1). Mode-B beam rows use the deduplicated rerun (D8). Paired bootstrap over matched (query, seed) pairs, B=1000; one-sided p for ours lower. POST-HOC.

Dataset	Uniform	Weighted	Δ	p	Hit@1_u	Hit@1_w
HotpotQA	76.2	77.3	-1.03	.122	0.336	0.354
MuSiQue	75.7	78.3	-2.66*	.029	0.118	0.127
2WikiMHQA	79.3	77.9	+1.47	.947	0.308	0.381
WebQSP	55.2	57.0	-1.87*	.001	0.602	0.629
CWQ	73.3	73.4	-0.13	.415	0.394	0.387
MetaQA-3hop	51.0	58.7	-7.69*	$<.001$	0.531	0.572

Table 19: Vote-share ablation: uniform vote share (self-consistency-style answer frequency) vs. judge-weighted top-1 posterior mass over the same 32 particles (cache replay of the archived runs; zero additional LLM calls; replayed answers reproduce the archived values on all 6,600 pairs). The judge weights improve confidence AUROC on five of six datasets (significantly on three; on MetaQA-3hop the uniform vote collapses to chance) and answer selection on five of six; the exceptions (2WikiMHQA confidence, $+1.5$ for the uniform vote, n.s.; CWQ answer selection, .394 vs. .387) are visible in the table. Δ = uniform $-$ weighted; paired bootstrap, one-sided p for weighted better ($^*p<.05$). POST-HOC.

Dataset	Top-1 mass	$-H_w$	$-H_u$
HotpotQA	77.3	77.6	77.8
MuSiQue	78.3	79.1	76.4
2WikiMHQA	77.9	$78.8^{\uparrow}$	$80.5^{\uparrow}$
WebQSP	57.0	57.1	$55.2^{\downarrow}$
CWQ	73.4	73.4	73.5
MetaQA-3hop	58.7	$56.6^{\downarrow}$	$50.9^{\downarrow}$

Table 20: Alternative confidence signals on the same particles (POST-HOC): negative answer-distribution entropy, weighted ($-H_w$) and unweighted ($-H_u$; semantic-entropy-style with exact-match answer classes), vs. the pre-registered top-1 posterior mass. ↓/↑: signal significantly worse/better than top-1 mass (paired bootstrap 95% CI excludes zero). Signal choice moves AUROC by at most ± 2.7 except on MetaQA-3hop, where both entropies degrade (unweighted to chance, 50.9); top-1 mass is the only signal that degrades nowhere; on 2WikiMHQA the entropies are significantly better, disclosed.

Dataset	Hit@1	Oracle@32	Sel. eff.	Mean #answers
HotpotQA	0.354	0.519	0.68	5.4
MuSiQue	0.127	0.258	0.49	8.7
2WikiMHQA	0.381	0.591	0.64	6.5
WebQSP	0.629	0.749	0.84	3.7
CWQ	0.387	0.626	0.62	5.3
MetaQA-3hop	0.572	0.903	0.63	12.6

Table 21: Search-vs-selection decomposition (POST-HOC, cache replay): oracle recall@32 = fraction of queries where some particle's answer matches gold; selection efficiency = Hit@1 / oracle. The accuracy gap has both components: the sampler finds the correct answer far more often than the aggregate emits it (selection losses; on MetaQA-3hop the answer is present for 90% of queries), and the KG-only search ceiling itself is well below corpus-reading accuracy (search losses, e.g. .26 on MuSiQue).

Comparison	ΔAUROC	pooled p	clustered p	B
Beam: HotpotQA	+11.79	$<.001$	.002	2000
Beam: 2WikiMHQA	+30.91	$<.001$	.001	2000
Beam: CWQ	+7.38	$<.001$	.002	2000
R^2C: HotpotQA	+3.72	.039	.120	2000
R^2C: MuSiQue	+2.75	.157	.223	2000
R^2C: 2WikiMHQA	+0.67	.374	.398	2000

Table 22: Query-clustered bootstrap (each query's three seeds resampled jointly) vs. pooled matched-pair bootstrap, emitted-mass primary: our vs.-beam results survive clustering; the HotpotQA vs.-R^2C delta weakens to non-significance under clustering (p=.039 → .120; disclosed in the main paper's Limitations). Mode-B (CWQ) beam cells use the protocol-frozen deduplicated rerun (D8), matching main Table 2. Both p columns are one-sided in the observed direction, as computed (pooled: paired bootstrap, B=1000; clustered: add-one), the diagnostic basis of appendix D7; main Table 2 reports the vs.-beam family two-sided. POST-HOC robustness check.

Dataset	τ	Coverage	Ours acc.	Beam acc. @ cov.	Δ
HotpotQA	1.60	0.957	0.345	0.446	-0.101
MuSiQue	1.00	1.000	0.120	0.147	-0.027
2WikiMHQA	3.60	0.913	0.369	0.383	-0.015
WebQSP	1.33	0.926	0.622	0.611	+0.011
CWQ	1.13	0.964	0.376	0.394	-0.019
MetaQA-3hop	1.31	0.967	0.559	0.648	-0.090

Table 23: Matched-coverage control at the split-conformal operating point: τ is fit by the same qid-level 30/70 procedure as the main paper's coverage verdicts (α=.1, split seed 42), then both methods are evaluated qid-level on the full pool (seed-mean signal, seed-majority correctness): ours answers where terminal ESS $\geq \tau$; the beam is truncated to the same coverage by its own confidence ranking. At this high-coverage operating point ($\geq$91%) absolute accuracy dominates the comparison and the beam's raw-accuracy edge carries over (it leads on 5/6); our selective advantage sits in the lower-coverage range of the risk–coverage curves (main Figure 2) on the accuracy-tied dataset (2WikiMHQA). Reported in full. MuSiQue's τ=1 (coverage 1.0) is the calibration degeneracy disclosed in Limitations. POST-HOC. Beam ranking/accuracy = archived pre-repair run (main §4); matched-coverage results for the deduplicated rerun were not evaluated under this protocol.

Dataset	Top-1	Top-3	Top-5	Δ(3–1)	p
HotpotQA	77.3	75.7	69.4	-1.6	.901
MuSiQue	78.3	76.9	73.8	-1.4	.862
2WikiMHQA	77.9	79.4	77.1	+1.5	.119
WebQSP	57.0	53.4	51.9	-3.7	.999
CWQ	73.3	71.6	64.7	-1.8	.967
MetaQA-3hop	58.7	54.3	50.9	-4.4	.999

Table 24: Cumulative top-k posterior mass as the confidence signal (POST-HOC; cache replay, zero extra calls), the natural repair for the multi-answer collapse. It does not help: top-3 mass is worse than the pre-registered top-1 mass on 5/6 datasets (only 2WikiMHQA moves up, n.s.), and on MetaQA-3hop the decay steepens with k. By gold-answer cardinality on MetaQA-3hop (top-1$\rightarrow$top-3): 1: 64.2$\rightarrow$66.9; 2–5: 64.0$\rightarrow$59.2; 6–20: 57.8$\rightarrow$48.3; 21+: 46.7$\rightarrow$44.0; the 21+ stratum worsens, so the collapse is not a mass-splitting artifact that cumulative mass can undo; the judge weights carry no set-level signal there. Δ/p: paired bootstrap, top-3 vs. top-1, one-sided. AUROC pooled over the three seeds on our answered rows (main Table 2's matched frame also counts unanswered rows).

Dataset	Mass-on-beam-ans.	Beam σ(LLR)	Δ	p	Zero-mass
HotpotQA	67.7	65.6	+2.1	.217	16%
MuSiQue	68.2	77.2	-9.0	.995	27%
2WikiMHQA	77.1	46.4	+30.8	$<$.001	26%
WebQSP	59.3	56.5	+2.8	.092	20%
CWQ	73.6	66.9	+6.7	$<$.001	18%
MetaQA-3hop	52.8	73.2	-20.4	$>$.999	31%

Table 25: Grafting the particle posterior onto the beam's answers (POST-HOC; cache replay): the beam's own final answer is scored by the posterior mass our sampler assigns to that answer string, replacing σ(LLR) as the beam's confidence (labels = the beam's own correctness). The graft significantly repairs the beam's broken signal on 2WikiMHQA (+30.8) and helps on CWQ (+6.7, $p<.001$; WebQSP +2.8, n.s.), but hurts where the beam's own signal is already strong (MuSiQue, MetaQA-3hop), and 16–31% of beam answers receive zero mass (absent from our particle pool): a drop-in confidence layer for a fixed answerer works exactly where the base signal is broken, not universally. Δ/p: paired bootstrap, lookup vs. own, one-sided. A beam answer absent from our particle pool is scored 0 and kept in the frame (the zero-mass column), never dropped. Appendix J.5 repeats the graft on the repaired Mode-B beam with the answer held fixed and five estimators; its Mode-A differences equal these by construction and are tested two-sided with query clusters there. The beam replayed here is the archived pre-repair run (Mode-B width collapse, main §4); main Table 2's Mode-B beam cells report the deduplicated rerun.

Dataset	σ(LLR)	BP path-wt	BP ent.-mass	Ours	Δ(ours$-$BP)	p
HotpotQA	65.6	64.1	66.1	77.3	+11.2	$<$.001
MuSiQue	77.5	75.6	72.6	78.3	+5.8	.049
2WikiMHQA	46.4	44.2	44.7	77.9	+33.2	$<$.001
WebQSP	56.5	50.4	53.6	57.0	+3.4	.047
CWQ	66.9	56.4	59.8	73.3	+13.4	$<$.001
MetaQA-3hop	73.2	50.7	51.3	58.7	+7.4	.001

Table 26: The beam-path-posterior control (POST-HOC; cache-replay rerun of the beam with path-level logging, zero new spend): Eq. (2)'s softmax-and-aggregate applied to the beam's own ten final-hop path LLRs, i.e., top-1 path weight and terminal-entity mass (answer proxy; the beam extracts no per-path answers), each scored against the beam's own correctness. Aggregating the beam's paths does not repair its confidence: 2WikiMHQA stays below chance (46.4 $\rightarrow$ 44.7), no dataset improves significantly, and MetaQA-3hop degrades sharply (73.2 $\rightarrow$ 51.3, its ten same-depth paths carrying no spread). Our top-1 mass (registered basis; Ours column) beats the beam-path posterior on 6/6 ($p \leq .049$, one-sided paired bootstrap; all six survive BH-FDR within this post-hoc six-test family; each method on its own answers/labels; AUROC pooled over the three seeds on the replayed run's rows, the Ours column on the matched subset, which trims only CWQ, 1,497/1,500 rows, hence its 73.3 here vs. 73.4 in the D6 companion table); on the Mode-A side, where the archived beam kept its intended width, this shows that aggregating the beam's paths under Eq. (2) alone does not close the gap; the answer-aggregation variant below shows that answer-level aggregation recovers most of it on 2WikiMHQA, so the gain is not isolated to independent sampling; the pre-repair Mode-B cells are diagnostic only. A stronger variant grants the beam answer aggregation, its per-path answers extracted and aggregated into an answer posterior (seed 42, one extraction call per path), and still trails the emitted-mass primary on all six datasets (Δ $+0.9$ to $+11.6$; four significant, $p\leq.018$; 2WikiMHQA recovers to 76.0 vs. our 76.9, n.s.; on MetaQA-3hop the beam's ten final paths are near-identical (mean 1.2 distinct), so that cell is a degenerate control; raw: `beam_perpath_answer_6ds_2026-07-17.json`). Rerun determinism vs. the archived runs: 6,599/6,600 identical final LLRs; 6,594/6,600 identical answers (the residual pairs differ only at exact LLR ties; five of the six σ(LLR) cells are unchanged at one decimal, and the flips move MuSiQue to 77.5, vs. 77.2 archived, the value main Table 2 reports). The beam whose paths are aggregated here is the archived pre-repair run; on Mode B its collapsed width (main §4) is itself the degeneracy these cells expose; main Table 2's Mode-B beam cells report the deduplicated rerun.

Dataset	Raw	−depth	Conf−partial	LLR−partial	Rank	Ours
HotpotQA	65.6	63.3	61.8	58.0	55.9	77.3
MuSiQue	77.2	73.1	70.0	63.0	60.5	78.3
2WikiMHQA	46.4	52.2	44.4	48.1	47.2	77.9
WebQSP	56.5	54.7	53.8	55.8	54.4	57.0
CWQ	66.9	65.0	60.3	57.0	56.3	73.4
MetaQA-3hop	73.2	50.0	71.8	72.2	71.4	58.7

Table 27: Depth-adjusted beam confidence control (POST-HOC; zero additional LLM calls): correctness AUROC of the archived beam confidence after four label-free depth adjustments: sign-flipped depth alone, confidence and LLR depth-partialling (score minus its depth-stratum mean), and within-depth percentile rank. If the beam's below-chance confidence were a mere depth offset, these would repair it; instead every adjustment stays at or near chance on 2WikiMHQA (44.4–52.2 vs. ours 77.9) and all four hurt on HotpotQA (55.9–63.3 vs. raw 65.6): conditional on depth, the beam score carries almost no correctness signal, so the repair requires re-scoring paths, not re-weighting stops. Raw column reproduces the archived beam AUROC exactly (sanity 6/6). AUROC tie-safe canonical, pooled 3 seeds; Ours column is the registered top-1 mass (main Table 2 reports the emitted-mass primary). Beam = archived pre-repair run (main §4); main Table 2's Mode-B beam cells report the deduplicated rerun. This table complements main §5.1's forced full-depth-3 rerun (early stopping disabled, same scoring), which likewise fails to repair the signal: it closes 18% of the HotpotQA gap (clustered p=.125, n.s.) and scores 43.4 on 2WikiMHQA, still below chance (`fixed_hb_gap_canonical_2026-07-18.json`).

Dataset	r=.2: cov (risk) wins	r=.3: cov (risk) wins	r=.4: cov (risk) wins
HotpotQA	0.09/0.07 (0.32/0.33) 10	0.25/0.15 (0.33/0.36) 19	0.38/0.34 (0.40/0.39) 17
MuSiQue	0.00/0.02 (NA/0.38) 0	0.00/0.03 (0.69/0.39) 2	0.04/0.06 (0.67/0.37) 11
2WikiMHQA	0.14/0.01 (0.26/0.83) 20	0.27/0.01 (0.32/0.78) 28	0.45/0.02 (0.39/0.74) 30
WebQSP	0.03/0.18 (0.33/0.29) 2	0.36/0.42 (0.33/0.32) 11	0.99/0.87 (0.36/0.38) 14
CWQ	0.04/0.03 (0.32/0.36) 5	0.24/0.08 (0.31/0.39) 25	0.38/0.33 (0.38/0.43) 17
MetaQA-3hop	0.08/0.50 (0.25/0.20) 0	0.27/0.79 (0.33/0.29) 0	0.85/0.99 (0.39/0.36) 4

Table 28: Fixed-target selective-risk operating points (POST-HOC; zero additional LLM calls): a threshold is calibrated on a 30% qid-level split to meet a target selective risk r, then evaluated on the held-out 70%, the identical plug-in recipe for both methods (no distribution-free guarantee claimed; 30 random splits, seeds 42–71; tie-safe thresholds; ours scored by the emitted-mass primary $\rho(a_{\text{emit}})$). Cells: realized coverage ours/beam (realized risk ours/beam) and the number of splits, of 30, where ours yields strictly higher coverage. Where the posterior's ranking is better (HotpotQA, 2WikiMHQA, CWQ), it converts to more answered queries at the same target; on 2WikiMHQA the beam's below-chance confidence admits almost no feasible operating point (30/30 at r=.4); where the beam's ranking is better (MetaQA-3hop) the advantage reverses, and at r=.1 ours is essentially infeasible on all six (mean coverage $\leq$.034; the beam likewise, except MetaQA-3hop at .22). Plug-in thresholds overshoot at tight targets (ours realizes .25–.33 at r=.2), identically on both sides. Against R^{2}C the same recipe is dominated by its higher raw accuracy (it reads the corpus; main §6), so this system-level comparison is meaningful only within the same KG-only interface. Mode-B beam cells replay the archived pre-repair run (main §4); the deduplicated rerun's higher full-coverage accuracy does not determine its fixed-target operating points, which were not evaluated here. NA marks a target with no feasible operating point (realized coverage $\approx$0).

Dataset	Hit@1 Δ (pp)	95% CI (clust.)	p (clust.)	EM ours	EM beam	EM Δ (pp)	p (clust.)
HotpotQA	-7.6	[-11.6, -3.6]	.002	0.313	0.387	-7.3	.001
MuSiQue	-2.1	[-5.1, +0.8]	.166	0.108	0.131	-2.3	.119
2WikiMHQA	+0.8	[-3.7, +5.0]	.728	0.363	0.353	+1.0	.648
WebQSP	-2.3	[-4.9, +0.5]	.095	0.419	0.448	-2.9	.059
CWQ	-4.1	[-7.2, -1.1]	.002	0.331	0.362	-3.1	.026
MetaQA-3hop	-24.1	[-29.6, -18.8]	.001	0.379	0.427	-4.8	.128

Table 29: Round-18 robustness companion (POST-HOC; cache replay, zero extra LLM calls): query-clustered bootstrap for the main accuracy comparison, strict exact-match (EM) re-scoring, and the 2WikiMHQA selective-accuracy interval. The registered pair-level bootstrap of main Table 1 remains primary; here the bootstrap resamples queries (a resampled query carries all three of its seed pairs; B=2000, seed 42, percentile CIs, two-sided add-one p; clusters = 300–500 queries per dataset), addressing dependence among a query's seed runs. The three large Hit@1 deficits (HotpotQA, CWQ, MetaQA-3hop) remain significant; MuSiQue and WebQSP widen to statistical ties (p=.166/.095); the registered reading (no accuracy gains for us) is unchanged. Strict EM (normalized exact match against any gold alias, the C3 protocol) agrees in sign with contains-matching on 6/6; under clustering HotpotQA and CWQ stay significant (the paired EM bootstrap has all five deficits at $p\leq.026$), and the MetaQA-3hop deficit shrinks from -24.1 to -4.8 pp (clustered p=.128): the set-valued golds make the two protocols diverge most there. Frames: the printed matched frames of main Table 1 (Mode-B beam = deduplicated rerun). 2WikiMHQA 50%-coverage selective accuracy (main Figure 2, .573 vs. .337): Δ=+.237, query-clustered 95% CI $[+.155, +.317]$, p=.001.

Dataset	PPR proposal	Uniform proposal	Δ	95% CI	p (1-sided)	n pairs
HotpotQA	77.3	77.0	+0.3	$[-2.1, +2.6]$	.413	900
MuSiQue	78.3	75.0	+3.3	$[-0.6, +6.9]$	.046	900
2WikiMHQA	77.9	79.5	−1.6	$[-3.9, +0.6]$	.932	900
WebQSP	57.0	54.8	+2.3	$[-0.3, +4.6]$	.040	1,500
CWQ	73.4	73.8	−0.4	$[-2.7, +1.9]$	.639	1,500
MetaQA-3hop	58.7	55.6	+3.0	$[-1.1, +6.7]$	.067	900

Table 30: Pre-registered PPR-proposal ablation (A4), full per-dataset table (registered top-1–mass AUROC, tie-safe canonical basis, paired bootstrap over matched (query, seed) pairs, B=1000, 95% percentile CIs; one-sided p in the observed direction, per D7: a two-sided reading at most doubles these, moving both nominal gains across .05): the PPR-guided proposal vs. the ablated uniform proposal (vanilla bootstrap sampling), each scored on its own answers. Gains reach nominal one-sided significance on 2/6 only (MuSiQue +3.3, WebQSP +2.3); the remaining four are null, two directionally negative (2WikiMHQA, CWQ). This is the full table behind main §5.5's deflationary claim: the proposal adds at most a small confidence gain and is not established as the source of the headline result. Because the proposal is the path prior p_0, the uniform column also changes the target $p_0\phi$: the two columns compare two path priors on their own answers, not two samplers of one posterior. PPR column = the registered top-1 basis (D6 companion table).

G.3 Particle-answer correctness by terminal judge score (post-hoc, archived prompts)

The judge scores a completed path without seeing the answer later extracted from it (§3.1). This subsection separates the two by an exact replay of the archived run from its content-keyed LLM cache: every call was served from the cache (851,013 hits, no miss, no network attempt), and a replayed row had to reproduce the archived final answer, emitted-answer mass, top log weight and $\log Z$ lower bound within 10^{-6} (0 of 6,600 rows tripped a check and are excluded; the CWQ query without particles, a PPR failure archived without an answer, is skipped on its three seeds). The plan was frozen before the replay (shipped as `2026-09-25_answer_frequency_and_judge_bins_posthoc_anonymized.yaml`). For every particle k: s_k = its terminal median judge LLR; credited answer = the Hit@1 of its extracted answer (0 if empty); on path = some vertex of its path would be credited by the same scorer; extraction error = an uncredited answer on an on-path particle. The bins of s_k are fixed from the

scorer's atoms (Appendix C): B0 no surviving LLR (placeholder 0); B1 $s \leq -5$; B2 $-5 < s < -2$; B3 $s = -2$, the imputed atom (tolerance 10^{-12}); B4 $-2 < s \leq 0$; B5 $0 < s \leq 5$; B6 $s > 5$; the cuts at ± 5 are design choices. Table 31 gives the per-dataset rates and the AUROC of s_k; Tables 32–33 give the bins for all particles and for the emitted particles (direct emissions; a row answered through the weighted-vote fallback has no emitted particle). On Mode B a bare MID vertex cannot be credited by a string match, so the on-path rates there are lower bounds. The per-particle rows file `review_posthoc_z4_particles_2026-09-25.jsonl.gz` holds every particle's score and credit; `--from-rows` re-derives every cell from it with zero LLM calls and checks it against the registered frame and the pin; the code supplement ships the plan only in anonymized form, so there `verify_z4_from_rows_release.py` runs the same checks on the plan digest the pin records. No answer-aware judge, one that scores the question, the path and a candidate answer together, was run.

Dataset	Particles	Scored	Rows d/f/a	Cred.	Err.	Err.$_t$	AUROC of s [95% CI]
HotpotQA	28,800	28,800	900/0/0	32.8	25.4	34.3	51.0 [48.8, 53.1]
MuSiQue	28,800	28,800	900/0/0	11.0	30.0	31.4	51.2 [47.8, 54.2]
2WikiMHQA	28,800	28,800	900/0/0	34.4	44.4	46.0	59.0 [56.8, 61.2]
WebQSP	48,000	48,000	1,500/0/0	61.3	14.6	11.5	50.4 [48.6, 52.0]
CWQ	47,904	47,904	1,497/0/0	37.5	29.2	30.8	53.1 [51.2, 55.1]
MetaQA-3hop	28,800	28,800	900/0/0	55.2	6.0	7.3	56.9 [55.4, 58.6]

Table 31: Particle-answer correctness and the terminal judge score, per dataset (post-hoc, archived prompts; pooled over the three seeds). Scored = particles with at least one surviving judge LLR; Rows d/f/a = replayed rows answered by a particle directly / through the weighted-vote fallback / abstained; Cred. = % of particles with a credited answer; Err. and Err.$_t$ = extraction-error rates (%) over all bins, among particles with a credited vertex on the path and with a credited last vertex; AUROC ($\times 100$) of the median judge LLR s for a credited answer over the scored particles, with a query-clustered percentile interval (B=2000; a cluster is a query with all its particles on all seeds). No p-values. Pin: `review_posthoc_z4_judge_bins_2026-09-25.json`.

Dataset	Bin of s	n	Share	Cred.	On path	Err.	Err.$_{\mathrm{t}}$	Cred.$_{\mathrm{off}}$	Empty	MID
HotpotQA	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	11,456	39.8	33.5	7.3	21.7	26.4	30.0	0.0	0.0
	B2 $(-5,-2)$	423	1.5	42.3	22.0	26.9	14.3	33.6	0.0	0.0
	B3 $s=-2$	15,808	54.9	30.4	2.0	37.2	57.4	29.7	0.0	0.0
	B4 $(-2,0]$	249	0.9	53.4	27.3	20.6	33.3	43.6	0.0	0.0
	B5 $(0,5]$	404	1.4	53.5	37.1	23.3	14.3	39.8	0.0	0.0
	B6 $s>5$	460	1.6	59.6	52.2	25.4	27.5	43.2	0.0	0.0
MuSiQue	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	9,383	32.6	10.7	1.3	35.7	50.0	10.0	0.0	0.0
	B2 $(-5,-2)$	136	0.5	25.0	8.8	16.7	n/a	19.4	0.0	0.0
	B3 $s=-2$	19,010	66.0	10.7	0.1	36.4	22.2	10.7	0.0	0.0
	B4 $(-2,0]$	99	0.3	32.3	10.1	20.0	33.3	27.0	0.0	0.0
	B5 $(0,5]$	93	0.3	22.6	12.9	8.3	0.0	12.3	0.0	0.0
	B6 $s>5$	79	0.3	31.6	22.8	11.1	0.0	14.8	0.0	0.0
2WikiMHQA	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	13,434	46.6	25.7	7.4	47.1	52.4	23.5	0.0	0.0
	B2 $(-5,-2)$	264	0.9	36.0	13.3	31.4	25.0	31.0	0.0	0.0
	B3 $s=-2$	14,761	51.3	42.1	4.1	42.7	51.3	41.5	0.0	0.0
	B4 $(-2,0]$	149	0.5	33.6	19.5	24.1	0.0	23.3	0.0	0.0
	B5 $(0,5]$	117	0.4	47.9	13.7	12.5	0.0	41.6	0.0	0.0
	B6 $s>5$	75	0.3	58.7	29.3	36.4	0.0	56.6	0.0	0.0
WebQSP	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	32,990	68.7	62.4	14.6	16.3	14.3	58.7	0.0	50.4
	B2 $(-5,-2)$	984	2.1	74.7	57.2	10.8	5.6	55.3	0.0	52.3
	B3 $s=-2$	12,044	25.1	55.1	1.2	9.9	33.3	54.7	0.0	35.1
	B4 $(-2,0]$	612	1.3	72.5	64.4	13.5	6.4	47.2	0.0	52.3
	B5 $(0,5]$	718	1.5	73.7	68.1	11.9	4.5	42.8	0.0	53.2
	B6 $s>5$	652	1.4	79.9	75.8	7.9	1.2	41.8	0.0	56.7
CWQ	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	33,735	70.4	37.0	10.2	31.1	32.9	33.4	0.0	43.4
	B2 $(-5,-2)$	966	2.0	53.3	34.5	19.2	13.8	38.9	0.0	42.7
	B3 $s=-2$	11,128	23.2	33.9	1.7	37.1	62.5	33.3	0.0	30.7
	B4 $(-2,0]$	551	1.2	58.6	40.7	24.6	28.6	47.1	0.0	49.7
	B5 $(0,5]$	847	1.8	54.4	39.0	22.1	23.1	39.5	0.0	45.2
	B6 $s>5$	677	1.4	62.3	52.9	25.4	16.7	48.6	0.0	49.3
MetaQA-3hop	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	24,185	84.0	55.1	25.8	6.8	9.5	41.9	0.0	0.0
	B2 $(-5,-2)$	720	2.5	85.0	72.6	0.8	0.0	47.2	0.0	0.0
	B3 $s=-2$	3,281	11.4	44.7	3.2	10.4	n/a	43.2	0.0	0.0
	B4 $(-2,0]$	273	0.9	86.4	75.8	0.0	0.0	43.9	0.0	0.0
	B5 $(0,5]$	266	0.9	81.2	77.1	1.0	0.0	21.3	0.0	0.0
	B6 $s>5$	75	0.3	57.3	54.7	0.0	0.0	5.9	0.0	0.0

Table 32: Judge-score bins, all particles (post-hoc, archived prompts), pooled over the three seeds (per-seed values in the pin). All columns after n are % of the bin's particles unless stated: Share = of the dataset's particles; Cred. = credited answer; On path = a credited vertex on the path; Err. = uncredited answer among the on-path particles (extraction error); Err.$_{\mathrm{t}}$ = the same among particles whose last vertex is credited; Cred.$_{\mathrm{off}}$ = credited answer among particles without a credited vertex; Empty = empty answer; MID = a bare Freebase MID on the path (not creditable by a string match). n/a = empty denominator. Descriptive; no monotonicity is claimed. Pin: `review_posthoc_z4_judge_bins_2026-09-25.json`.

Dataset	Bin of s	n	Share	Cred.	On path	Err.	Err.$_{\mathrm{t}}$	Cred.$_{\mathrm{off}}$	Empty	MID
HotpotQA	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B2 $(-5,-2)$	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B3 $s=-2$	573	63.7	26.5	2.4	21.4	50.0	25.2	0.0	0.0
	B4 $(-2,0]$	49	5.4	30.6	22.4	9.1	n/a	13.2	0.0	0.0
	B5 $(0,5]$	112	12.4	44.6	37.5	14.3	16.7	20.0	0.0	0.0
	B6 $s>5$	166	18.4	61.4	49.4	12.2	18.8	35.7	0.0	0.0
MuSiQue	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B2 $(-5,-2)$	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B3 $s=-2$	761	84.6	9.6	0.0	n/a	n/a	9.6	0.0	0.0
	B4 $(-2,0]$	42	4.7	31.0	14.3	33.3	50.0	25.0	0.0	0.0
	B5 $(0,5]$	44	4.9	18.2	13.6	0.0	0.0	5.3	0.0	0.0
	B6 $s>5$	53	5.9	37.7	24.5	7.7	0.0	20.0	0.0	0.0
2WikiMHQA	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	2	0.2	0.0	0.0	n/a	n/a	0.0	0.0	0.0
	B2 $(-5,-2)$	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B3 $s=-2$	749	83.2	37.2	5.2	43.6	100.0	36.2	0.0	0.0
	B4 $(-2,0]$	50	5.6	34.0	24.0	25.0	0.0	21.1	0.0	0.0
	B5 $(0,5]$	58	6.4	43.1	13.8	12.5	0.0	36.0	0.0	0.0
	B6 $s>5$	41	4.6	53.7	29.3	16.7	0.0	41.4	0.0	0.0
WebQSP	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	23	1.5	56.5	21.7	20.0	0.0	50.0	0.0	65.2
	B2 $(-5,-2)$	15	1.0	73.3	40.0	16.7	0.0	66.7	0.0	46.7
	B3 $s=-2$	828	55.2	56.2	1.0	25.0	n/a	56.0	0.0	30.7
	B4 $(-2,0]$	107	7.1	64.5	63.6	19.1	16.7	35.9	0.0	56.1
	B5 $(0,5]$	211	14.1	69.7	71.1	17.3	8.0	37.7	0.0	53.1
	B6 $s>5$	316	21.1	75.6	74.1	10.7	0.0	36.6	0.0	50.9
CWQ	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	34	2.3	11.8	5.9	50.0	n/a	9.4	0.0	55.9
	B2 $(-5,-2)$	14	0.9	50.0	42.9	33.3	n/a	37.5	0.0	50.0
	B3 $s=-2$	903	60.3	28.0	2.3	38.1	100.0	27.2	0.0	30.0
	B4 $(-2,0]$	100	6.7	60.0	38.0	10.5	0.0	41.9	0.0	48.0
	B5 $(0,5]$	205	13.7	52.7	40.0	18.3	50.0	33.3	0.0	50.7
	B6 $s>5$	241	16.1	61.4	53.5	20.9	13.3	41.1	0.0	46.9
MetaQA-3hop	B0 none	0	0.0	n/a	n/a	n/a	n/a	n/a	n/a	n/a
	B1 $s\leq -5$	73	8.1	69.9	52.1	5.3	n/a	42.9	0.0	0.0
	B2 $(-5,-2)$	47	5.2	83.0	61.7	0.0	0.0	55.6	0.0	0.0
	B3 $s=-2$	526	58.4	43.7	4.2	13.6	n/a	41.9	0.0	0.0
	B4 $(-2,0]$	101	11.2	78.2	66.3	0.0	0.0	35.3	0.0	0.0
	B5 $(0,5]$	123	13.7	74.8	69.9	2.3	0.0	21.6	0.0	0.0
	B6 $s>5$	30	3.3	80.0	76.7	0.0	0.0	14.3	0.0	0.0

Table 33: Judge-score bins, emitted particles only (post-hoc, archived prompts), pooled over the three seeds (per-seed values in the pin). All columns after n are % of the bin's particles unless stated: Share = of the dataset's particles; Cred. = credited answer; On path = a credited vertex on the path; Err. = uncredited answer among the on-path particles (extraction error); Err.$_{\mathrm{t}}$ = the same among particles whose last vertex is credited; Cred.$_{\mathrm{off}}$ = credited answer among particles without a credited vertex; Empty = empty answer; MID = a bare Freebase MID on the path (not creditable by a string match). n/a = empty denominator. Descriptive; no monotonicity is claimed. Pin: `review_posthoc_z4_judge_bins_2026-09-25.json`.

H Material Deferred from the Main Text

The following material was moved here from the main text for space; its wording is unchanged.

H.1 ALGORITHM SUMMARY

Algorithm 1 Particle-KGRAG (one query q; 129 LLM calls)

Require: KG $\mathcal{G}$; query q; $K{=}32$; depth cap $T{=}6$; conformal threshold τ (§3.4)

1: $\boldsymbol{\pi}_q \leftarrow \mathrm{PPR}(\mathcal{G}, \text{seeds of } q)$; $q^{(1)}, q^{(2)}, q^{(3)} \leftarrow$ paraphrases
2: **for** $k = 1, \ldots, K$ (independent) **do**
3: $s^{(k)}_{1:T} \leftarrow$ walk $\leq T$ steps from a seed, each step $\propto \boldsymbol{\pi}_q$ over unvisited neighbors; $a^{(k)} \leftarrow$ answer from $s^{(k)}_{1:T}$
4: $\log \phi^{(k)} \leftarrow \mathrm{median}_i \ \mathrm{LLR}\big(s^{(k)}_{1:T}, q^{(i)}\big)$ (§3.1)
5: **end for**
6: $\tilde{w} \leftarrow \mathrm{softmax}_k(\log \phi^{(k)})$; $\rho(a) \leftarrow \sum_{k:\, a^{(k)}=a} \tilde{w}^{(k)}$ Eq. (2)
7: $\log \hat{Z}_q \leftarrow \mathrm{logsumexp}_k(\log \phi^{(k)}) - \log K$ (§3.4)
8: $a_{\mathrm{emit}} \leftarrow a^{(k^\star)}$, $k^\star = \arg\max_k \tilde{w}^{(k)}$; conf. $\leftarrow \rho(a_{\mathrm{emit}})$
9: $\mathrm{ESS} \leftarrow \big(\sum_k \tilde{w}^{(k)}\big)^2 / \sum_k \big(\tilde{w}^{(k)}\big)^2$; abstain if $\mathrm{ESS} < \tau$

Ensure: a_{emit}; confidence $\rho(a_{\mathrm{emit}})$; abstain flag; evidence $\log \hat{Z}_q$

H.2 FURTHER RELATED WORK

Sampling-based uncertainty. Sampling-based LLM-UQ (answer-frequency self-consistency and its confidence-weighted refinement (Wang et al., 2023; Taubenfeld et al., 2025), semantic entropy (Kuhn et al., 2023), verbalized confidence (Tian et al., 2023)) shares our sample-then-aggregate shape; top-1 posterior mass generalizes answer-frequency confidence from iid text samples to judge-weighted, KG-grounded sequential paths; the uniform-weight special case is exactly the self-consistency vote share. RTSoG (Long et al., 2025b) and ReKG-MCTS (Song et al., 2025) (training-free tree search over KG paths) steer with reward or rollout signals, not a posterior, and quantify no uncertainty.

Particle methods for LM inference. Our sampler is the no-resampling special case of twisted sequential Monte Carlo (SMC) (Zhao et al., 2024) (§3.1), whose learned twist functions sit outside our training-free scope, hence the plainer name self-normalized importance sampling, matching the code (§3). Particle-based inference-time scaling is emerging (Puri et al., 2025; Giannone et al., 2026); the closest (Emmenegger et al., 2026) combines particles, a normalizing-constant estimate, and conformal abstention; the difference from ours lies in the sample space (LM outputs rather than KG paths from a PPR-guided walk), the judge potential, and the direction of the controlled risk, not in the independence of the completed samples, which our K paths share. Conformal calibration with SNIS weights and ESS has Bayesian-conformal precedents (Fong & Holmes, 2021; Lou & Luo, 2026), which we transplant to LLM-judged graph retrieval for false abstention.

H.3 ADDITIONAL METHOD NOTES

Why "SNIS," not "twisted SMC." Our sampler has intermediate twist $\psi_t \equiv 1$ and resampling never occurs: with a null twist the running log-weights stay at 0, so intermediate ESS equals K and the resampling trigger ($\mathrm{ESS}/K < 0.5$) is structurally unreachable; the terminal potential induces the weight variance, and that terminal ESS is the abstention statistic of §3.4. The pre-registration's "twisted SMC" label is disclosed as a mismatch in §6; every theoretical statement below is a plain importance-sampling result.

H.4 ADDITIONAL EXPERIMENTAL-SETUP DETAILS

Pre-registration. The full protocol was pre-specified and locally frozen before the main run (a tagged repository commit; what the freeze does and does not certify: Appendix B): sample sizes, seeds, metrics, decision rules (e.g., C1: mean top-1 AUROC $\geq$ 65 with per-dataset $\Delta \geq 0$ vs. R^2C (Mode-A) or a 65 floor (Mode-B) on $\geq$4/6; the beam-parity rule), and falsification criteria. Deviations are disclosed in the appendix. Split-conformal calibration uses a 30/70 query-level split

($\alpha = 0.1$). Code (method, baselines, analysis), configurations, and pinned analysis outputs ship as an anonymized code supplement.

Reimplementation disclosure. We compare against controlled reimplementations (all baselines share our exact subgraphs, backbone, and scoring), not published numbers. Published ToG headline results rely on GPT-4-class backbones, live full-KG SPARQL, oracle topic entities, and lenient matching; a third-party rerun required per-dataset hand-tuning (Jo et al., 2025).

H.5 Pre-registered confidence criteria and the held-out abstention check

Pre-registered criteria. The registered confidence criterion (C5: top-1–mass AUROC $\geq$ 65 on $\geq$4/6 datasets) passes at exactly 4/6 (appendix, D6); the emitted-mass primary clears the same floor on the same four. The registered evidence-signal criterion (C6: $\log \hat{Z}_q$ AUROC $\geq$ 60 on $\geq$3/6) also passes at 4/6 (per-dataset values: appendix); per its registered framing, $\log \hat{Z}_q$ is a modest, corpus-conditional, diagnostic-only signal. A registered overlap analysis (C4) shows the posterior is not merely re-ranking the beam: Particle-KGRAG is correct where the beam is wrong on 30–85 (query, seed) rows per dataset (of 900–1,500), i.e., 19–50 distinct queries on at least one seed and 8–26 on a majority of seeds; Cohen's κ between the two methods' correctness labels is .61–.73 on five datasets and .27 on MetaQA-3hop (overlap, not orthogonality, as pre-registered; recomputed on the canonical matched frames, pin `c4_overlap_recompute_2026-09-19.json`). The abstention rule's registered direction also holds out of sample (post-hoc): over 30 random calibration splits, held-out $\mathbb{P}(\text{abstain} \mid \text{correct})$ averages .047–.101 against α=.1 (above only on CWQ/MetaQA, both .101; MuSiQue degeneracy: §6).

H.6 Cost and latency: a fixed call budget

Particle-KGRAG's LLM budget is fixed by construction (129 calls/query, §3): estimated cost \$0.0057–0.0066/query (GPT-4.1-mini) vs. the beam's \$0.0048–0.0078 under its deduplicated accounting; the beam is cheaper on four of six datasets (2–26%), we on two (6–23%); and, unlike the beam's difficulty-dependent call count, ours is constant. R^2-KG's dual-agent iteration costs \$0.0153–0.0312/query (2.3–4.7$\times$ ours), rising with query complexity. R^{2}C at default B=10 is comparable to ours; the call-matched B=40 variant ($\sim$\$0.024/query) controls for budget in §5.4. Two caveats. First, the ToG-2-style baseline is the cheapest and fastest method in the suite (one final LLM call plus dense retrieval; \$0.0007 and $\sim$2s/query): low cost per se is not our claim; a fixed budget with usable confidence is. Second, our sequential client averages 102s/query on HotpotQA (vs. 46s beam, 64s R^{2}C B=10); the terminal sweep consists of independent calls that a parallel client could issue concurrently; we did not measure such a client, so no latency advantage is claimed.

H.7 A stronger abstainer on Mode B (R^2-KG)

R^2-KG's dual-agent abstention is a stronger selective predictor on the Mode-B side: abstaining on 21%/59%/15% (WebQSP/CWQ/MetaQA), it reaches Hit@1 .746/.843 (WebQSP/MetaQA), above ours (.629/.572); on CWQ the same aggressiveness drives it to .360, below ours (.387). On set-valued gold its multi-answer output is genuinely better (set-F1 .722 vs. our .237 on MetaQA; leakage screen benign, appendix). It runs at a fixed, non-tunable operating point (star, Figure 2).

H.8 Particle count and cross-backbone check (post-hoc)

More particles buy confidence, not accuracy (post-hoc). A K sweep (K=8/16/32, seed 42, HotpotQA + 2WikiMHQA): Hit@1 is flat in K (.35–.38), while top-1–mass AUROC peaks at K=32 on HotpotQA (+5.3 over K=16, p=.009) and saturates by K=16 on 2WikiMHQA, at cost linear in K (appendix). A cross-backbone check agrees (post-hoc): with Qwen2.5-7B, Hit@1 falls to .20/.25 yet top-1 AUROC is 84.8/78.9, above the primary backbone's and the matched Qwen beam's (71.3/52.5).

H.9 THE PRE-REGISTERED C1 COMPARISON WITH R^2C

Dataset	Particle-KGRAG	R^2C	Δ	p
HotpotQA	77.3	73.7	$+3.6$	.050
MuSiQue	78.3	76.5	$+1.9$	.262
2WikiMHQA	77.9	76.6	$+1.2$	.270

Table 34: Pre-registered C1: confidence AUROC, Particle-KGRAG registered top-1 mass vs. R^2C perturbation confidence (Wikipedia benchmarks; 900 matched pairs per dataset). The emitted-mass primary preserves the direction on all three (appendix, D6).

I PARAPHRASE-MEDIAN COMPONENT ABLATION (POST-HOC)

This post-hoc ablation asks whether the three-paraphrase median in the terminal judge (96 of the 129 calls per query) is necessary for the confidence signal. We replay the archived main run from the LLM cache (6 datasets $\times$ 3 seeds, 6,597 answered query–seed rows, 851,013 cache hits, zero misses, zero new calls), recover the three per-particle paraphrase LLRs, and recompute the particle weights with (i) the deployed median, (ii) each single paraphrase's LLR, and (iii) their mean; the archived paraphrases are re-used, not re-sampled. Under convention A the archived emitted answer is held fixed and its correctness label is recomputed from the archived per-query gold, so only the confidence $\rho_v(a_{\text{emit}})$ changes; under convention B each variant re-emits its own argmax-particle answer under the deployed rule and is scored on that answer. The plan, the decision rule, and the output path were declared before the replay ran; the deployed-median variant reproduces the archived run under five parity gates (determinism, emitted mass, median structure, emission rule, label parity with the archived gold; zero violations). The replay was run once more after an external review added the fifth gate (label parity); every number reported here is unchanged, and both outputs ship with the code.

Result: the declared necessity rule (all three single-paraphrase 95% CIs of ΔAUROC(median $-$ paraphrase) above zero, convention A) holds on 0 of 18 dataset–seed cells. Under convention A the paired-bootstrap CI of ΔAUROC contains zero in 51 of 54 (dataset, seed, paraphrase) cells; it lies above zero (median ahead) in two cells (MetaQA-3hop, seed 42, paraphrase 1; MetaQA-3hop, seed 123, paraphrase 2) and below zero in one cell (2WikiMHQA, seed 456, paraphrase 1). Under convention B it contains zero in 53 of 54 cells; it lies above zero (median ahead) in one cell (CWQ, seed 42, paraphrase 2); Table 36 reports seed means only, so that cell is not visible there. Per-dataset mean ΔAUROC (convention A) spans -1.0 to $+3.7$ points (Table 35), and re-emitting from a single paraphrase moves Hit@1 by at most 0.031 in any (dataset, seed, paraphrase) cell (Table 36 shows seed means). Replacing the median by the mean of the three LLRs gives a CI containing zero in 17 of 18 cells (convention A) (above zero in 0, below in 1). The pre-registered paraphrase-variance diagnostic ϵ (mean per-particle std of the three LLRs; registered as a mini-pilot gate and deferred in the Stage-2 registration, not previously reported) lies between 1.37 and 1.63 across datasets. These are unadjusted non-detections (54 cellwise intervals per convention, no multiplicity correction) with no pre-declared equivalence margin: a single-paraphrase judge is shown to lose confidence in the above-zero cells and nowhere shown to preserve it. On these runs the per-query budget could therefore fall from 129 to 65 calls; we report this as a negative result for the component, not as evidence for the cheaper setting: the three-paraphrase design was fixed before any of these numbers were seen, and a single-paraphrase deployment would also forgo the variance diagnostic.

Dataset	median (deployed)	paraphrase 1	paraphrase 2	paraphrase 3	mean of 3	CI>0 (1/2/3)	ϵ
HotpotQA	77.5 ± 1.7	77.6 ± 1.6	77.5 ± 2.1	76.2 ± 1.4	77.2 ± 1.1	0/0/0	1.61
MuSiQue	79.4 ± 2.1	77.1 ± 2.0	79.6 ± 1.7	79.6 ± 3.2	78.9 ± 2.4	0/0/0	1.47
2WikiMHQA	77.3 ± 1.3	78.2 ± 1.9	77.2 ± 2.4	76.9 ± 0.9	76.5 ± 1.9	0/0/0	1.63
WebQSP	57.3 ± 2.1	57.4 ± 2.1	56.9 ± 2.2	57.3 ± 3.0	57.9 ± 2.5	0/0/0	1.56
CWQ	73.3 ± 1.5	72.8 ± 1.3	73.6 ± 2.0	74.3 ± 1.1	74.0 ± 1.7	0/0/0	1.62
MetaQA-3hop	58.9 ± 3.9	56.7 ± 4.3	55.2 ± 3.1	56.8 ± 2.6	57.7 ± 2.3	1/1/0	1.37

Table 35: Paraphrase-median component ablation, convention A (POST-HOC; cache replay, zero additional LLM calls): confidence AUROC of the emitted-answer mass $\rho_v(a_{\text{emit}})$ when the per-particle judge score is the deployed median of the three paraphrase LLRs, a single paraphrase's LLR, or their mean; the emitted answer and its label are held fixed at the archived run, so only the confidence changes. Mean $\pm$ sample std over the three seeds (42/123/456), answered rows only; this is a seed mean, whereas the center of Table 2 is the pooled AUROC, so the deployed-median column can differ from it at the first decimal. CI>0: number of seeds (of 3) on which the query-level paired-bootstrap 95% CI of ΔAUROC(median $-$ paraphrase i) lies above zero, for i=1/2/3. ϵ: the pre-registered paraphrase-variance diagnostic (mean over particles of the sample std of the three paraphrase LLRs, averaged over seeds); every particle carried all three LLRs. Plan and decision rule declared before the replay ran (`2026-09-18_paraphrase_median_ablation_posthoc.yaml`); the replay output is `paraphrase_median_ablation_2026-09-18_v2.json`.

	confidence AUROC (re-emitted answer)					Hit@1 (re-emitted answer)				
Dataset	median	para. 1	para. 2	para. 3	mean of 3	median	para. 1	para. 2	para. 3	mean of 3
HotpotQA	77.5 ± 1.7	76.1 ± 1.6	77.3 ± 3.0	76.7 ± 1.4	76.7 ± 1.0	0.354	0.364	0.359	0.357	0.361
MuSiQue	79.4 ± 2.1	78.6 ± 1.0	80.9 ± 2.5	79.0 ± 0.9	76.9 ± 1.7	0.127	0.123	0.130	0.124	0.130
2WikiMHQA	77.3 ± 1.3	78.6 ± 2.0	77.4 ± 0.9	76.5 ± 0.5	75.8 ± 1.3	0.381	0.378	0.372	0.381	0.383
WebQSP	57.3 ± 2.1	57.5 ± 1.6	57.6 ± 1.5	57.4 ± 2.1	57.8 ± 2.1	0.629	0.631	0.628	0.633	0.630
CWQ	73.3 ± 1.5	72.7 ± 1.3	72.0 ± 0.6	74.3 ± 1.2	73.3 ± 1.3	0.387	0.387	0.398	0.391	0.391
MetaQA-3hop	58.9 ± 3.9	60.5 ± 2.4	58.0 ± 2.8	57.2 ± 2.7	56.0 ± 1.9	0.572	0.571	0.554	0.592	0.570

Table 36: Paraphrase-median component ablation, convention B (POST-HOC; cache replay): each variant re-emits its own argmax-particle answer under the deployed rule, and is scored on its own answer (confidence AUROC, mean $\pm$ std over three seeds) and by its own Hit@1 (mean over seeds); answered rows only, abstention not re-applied. The deployed median column reproduces the archived run under five parity gates (determinism, emitted mass, median structure, emission rule, label parity with the archived gold; zero violations).

J Additional Post-Hoc Robustness Checks

J.1 Ranked-metric tie convention

Our confidence scores are continuous, but the beam's σ(top-path LLR) is coarse and heavily tied, so how a ranked metric charges a tie group matters asymmetrically. Two order-invariant conventions appear in this work and they are not the same estimator. *Block end* (the convention behind the reported AURC and excess AURC): a tie group enters coverage as one block and every position inside it is charged the cumulative risk at the block's end. *Prorated* (the convention the risk–coverage curves of Figure 2 are drawn with, and the one behind every selective-accuracy number we report): the boundary group contributes its remaining quota at the group's mean accuracy; averaged over coverage this is the expectation of the registered raw prefix definition under uniformly random tie-breaking. The two agree whenever all scores are distinct, and diverge in proportion to how coarse a method's confidence is.

Table 37 recomputes both metrics under both conventions from the same archived matched frames (cache replay, no new calls). Our own values move by less than 0.001 everywhere, the beam's by up to 0.040. One AURC verdict is convention-dependent (HotpotQA) and one excess-AURC verdict is (MuSiQue). On MuSiQue the excess-AURC advantage reported in §5.2 (nominal, p=.043 under block end) reverses under the prorated convention (0.146 vs. 0.134) and is not significant there

(paired p=0.760, query-clustered p=0.662); we therefore do not count it. The three excess-AURC wins that carry the claim hold under both conventions and under both bootstraps, as do the five stable AURC verdicts; under prorated ties their one-sided p-values (paired, query-clustered) are HotpotQA .001, .044, whose clustered interval includes zero; 2WikiMHQA $<$ .001, $<$ .001; CWQ $<$ .001, .004. A random ranking's expected AURC is a method's own error rate, so each method has its own floor: ours stays below its floor on all six (closest MetaQA-3hop and WebQSP), whereas the beam is above its own on 2WikiMHQA under both conventions (0.631 and 0.650 against a floor of 0.627), consistent with its sub-chance AUROC there. Existing p-values do not transfer between conventions, so every prorated comparison in the table was re-tested from scratch (paired bootstrap B=1000; query-clustered B=2000, add-one p; one-sided, ours lower). The registered estimator is unchanged and remains primary; this is a robustness check, not a re-registration.

	AURC				Excess AURC			
	block end		prorated		block end		prorated	
Dataset	Ours	Beam	Ours	Beam	Ours	Beam	Ours	Beam
HotpotQA	0.455	0.458	0.455	0.439	0.177	0.250	0.177	0.231
MuSiQue	0.758	0.744	0.758	0.704	0.146	0.174	0.146	0.134
2WikiMHQA	0.431	0.631	0.431	0.650	0.179	0.372	0.179	0.390
WebQSP	0.328	0.284	0.328	0.281	0.249	0.214	0.249	0.212
CWQ	0.439	0.465	0.439	0.455	0.193	0.256	0.193	0.246
MetaQA-3hop	0.371	0.131	0.371	0.133	0.263	0.112	0.263	0.114

Table 37: AURC and excess AURC under both tie conventions (POST-HOC; cache replay of the archived matched frames, zero additional LLM calls; lower is better). *Block end* = the reported convention; *prorated* = the convention the risk–coverage figure is drawn with. Our values are nearly convention-invariant because the emitted-answer posterior mass is near-continuous; the beam's σ(top-path LLR) is heavily tied and moves much further. The p-values quoted for this table in the text are one-sided (ours lower), re-tested under each convention. Pin: `aurc_tie_convention_2026-09-19.json`.

J.2 What the ESS gate costs in answered accuracy

The registered conformal analysis reports false abstention, $\mathbb{P}(\text{abstain} \mid \text{correct})$, and Table 13 already shows that on five of six datasets the gate discards correct answers at a higher rate than incorrect ones. It does not report the quantity a deployer would ask for: whether the answered set left behind is more accurate than no gating at all. We add that comparison by replaying the archived per-query runs through the same calibration protocol (same α, calibration ratio, τ floor and 30 splits; no new LLM calls), first reproducing the pinned false-abstention values as a positive control (18/18 (dataset, seed) systems).

Gating lowers answered-set accuracy on all 18 dataset–seed systems, by 0.1 to 2.2 points (mean over the 30 splits). This is a descriptive result over correlated systems, not 18 independent tests, and it does not affect the false-abstention bound, which is one-sided and holds regardless. It does mean the gate should not be read as an error filter: at this operating point it removes correct answers slightly faster than it removes wrong ones, so its value is bounding how often a correct answer is discarded, not raising the accuracy of what remains. The useful confidence ranking (§5.1) and this ESS gate are separate mechanisms and should be judged separately. Pin: `ess_gate_accuracy_cost_2026-09-19.json`.

J.3 Beam confidence under matched judge-call conditions

Two conditions of the judge call differed between the ToG-style beam and our method; we test whether they account for the confidence gap. First, the path text: our prompts write every edge with a label slot, `e -[r]-> e'`, whereas the beam's write entities only, `e -> e'`; the template around the path is byte-identical (as-sent form: the judge item of the implementation details). Second, the floor: our judge log-probabilities are clipped at -10, whereas the beam's calls pass no clip and inherit the wrapper default of -20, the API's own minimum, which can then bind only where the token-absence imputation pushes a value below it. Relation labels exist in our prompts only on the

Mode B side (every Mode A edge carries the constant placeholder `_undirected_`; census in the implementation details), so the control covers WebQSP, CWQ, and MetaQA-3hop.

We re-scored the beam's archived final paths (38,235 paths over the 3,900 query–seed pairs of the canonical deduplicated run) under four conditions: entity-only text at -20 (the archived condition, kept as a parity anchor), entity-only text at -10, relation-annotated text at -20, and relation-annotated text at -10 (both conditions matched to ours). The plan, including the decision rule below, was written and committed before the re-scoring ran. Nothing is re-traversed and no answer changes, so Hit@1 is unchanged by construction; only the confidence, σ(top-path LLR) as archived, can move. The relation map is built from the same per-question triples in both directions, so every one of the 87,291 path edges resolves to a label, which is more label information than our own Mode B prompts carry. Gates: the anchor reproduced the archived LLR within 10^{-9} on 38,235 of 38,235 paths (100.00%), the scoring error rate was 0.00%, and the 12,669 new judge calls cost $0.48 (the entity-only calls were cache hits).

Table 38 gives every cell. Moving the floor alone changes the beam's AUROC by -0.5 to $+0.3$ points across the nine dataset–seed cells (the -20 floor binds on 2.5% of its scored paths, the -10 floor on 18.5%); relation text alone changes it by -2.0 to $+1.9$; both together by -2.2 to $+0.8$. Against ours, the paired difference under the matched condition is -11.6 to $+11.3$ points (archived condition, same rows: -13.5 to $+9.1$). By the rule declared before the run (the 95% paired-bootstrap interval of the difference excludes zero; we require it on every seed of a dataset), the registered confidence advantage survives matching on one of the three Mode B datasets (CWQ); under the archived condition the same per-seed rule holds on one (CWQ). In detail: WebQSP: the interval includes zero on all three seeds under both conditions; CWQ: the interval lies above zero on all three seeds under both conditions; MetaQA-3hop: the interval includes zero on two and lies below zero on one of the three seeds under both conditions. The registered comparison remains primary; these numbers are a post-hoc control and replace nothing. Table 39 adds the pre-declared secondary ranked metrics, AURC and excess AURC under both tie conventions of Appendix J.1: across the four conditions the beam's values move by at most 0.038 on any dataset. Pin: `beam_condition_matching_2026-09-19_v2.json` (per-query rows: `beam_condition_matching_rows_2026-09-19.jsonl`).

				Beam AUROC by judge-call condition				Ours − beam (both)	
Dataset	Seed	n	Ours	archived	clip -10	relation	both	Δ [95% CI]	p
WebQSP	42	500	59.5	59.3	59.6	61.2	60.1	-0.5 [-7.8, $+6.8$]	.546
	123	500	57.0	59.3	59.6	61.2	60.1	-3.1 [-10.2, $+4.0$]	.813
	456	500	55.5	59.3	59.6	61.2	60.1	-4.6 [-11.6, $+2.4$]	.891
	mean		57.3 ± 2.1	59.3	59.6	61.2	60.1	-2.7	
CWQ	42	500	75.0	65.9	65.4	64.0	63.7	$+11.3$ [$+5.9$, $+16.6$]	$<.001$
	123	500	73.0	65.9	65.4	64.0	63.7	$+9.3$ [$+3.9$, $+14.9$]	$<.001$
	456	500	72.0	65.9	65.4	64.0	63.7	$+8.3$ [$+2.7$, $+13.5$]	.002
	mean		73.3 ± 1.5	65.9	65.4	64.0	63.7	$+9.6$	
MetaQA-3hop	42	300	61.8	68.0	67.8	66.2	66.2	-4.4 [-14.4, $+6.0$]	.795
	123	300	60.4	68.0	67.8	66.2	66.2	-5.8 [-15.6, $+5.0$]	.859
	456	300	54.5	68.0	67.8	66.2	66.2	-11.6 [-21.8, -0.7]	.981
	mean		58.9 ± 3.9	68.0	67.8	66.2	66.2	-7.3	

Table 38: Confidence AUROC (%) of the ToG-style beam on its archived final paths under four judge-call conditions (POST-HOC; Mode B; rows and labels archived, no re-traversal, no answer change), our emitted-mass AUROC on the same rows, and the paired difference ours − beam under the matched condition (query-level paired bootstrap, B=1000, 95% interval; one-sided p = share of replicates with a difference ≤ 0). The beam is deterministic given the fixed query set (temperature 0), so its four columns repeat across seeds exactly as in the registered run; the difference varies through our seeds. *archived* = entity-only path text, floor -20 (parity anchor); *clip* -10 = entity-only text, floor -10; *relation* = relation-annotated text, floor -20; *both* = relation-annotated text, floor -10 (matched to ours). Mean rows: mean $\pm$ sample standard deviation over the three seeds (a seed-invariant value is printed once). Pin: `beam_condition_matching_2026-09-19_v2.json`.

		AURC		Excess AURC	
Dataset	Ranking	block end	prorated	block end	prorated
WebQSP	beam, archived	0.284	0.281	0.214	0.212
	beam, clip -10	0.282	0.279	0.212	0.210
	beam, relation	0.259	0.251	0.189	0.181
	beam, both	0.283	0.276	0.214	0.207
	ours (mean $\pm$ sd)	0.328 ± 0.008	0.328 ± 0.008	0.248 ± 0.010	0.248 ± 0.010
CWQ	beam, archived	0.465	0.455	0.256	0.246
	beam, clip -10	0.481	0.472	0.272	0.263
	beam, relation	0.497	0.478	0.288	0.269
	beam, both	0.503	0.483	0.293	0.274
	ours (mean $\pm$ sd)	0.440 ± 0.009	0.440 ± 0.009	0.193 ± 0.012	0.193 ± 0.012
MetaQA-3hop	beam, archived	0.131	0.132	0.112	0.113
	beam, clip -10	0.120	0.119	0.101	0.101
	beam, relation	0.109	0.107	0.090	0.088
	beam, both	0.110	0.108	0.091	0.089
	ours (mean $\pm$ sd)	0.371 ± 0.012	0.371 ± 0.012	0.262 ± 0.021	0.261 ± 0.021

Table 39: Secondary ranked metrics of the matched-condition control (POST-HOC; lower is better): AURC and excess AURC of the beam under the four judge-call conditions of Table 38 (seed-invariant, so one row each) and of our emitted-mass confidence on the same rows (mean $\pm$ sample standard deviation over the three seeds), under both tie conventions of Appendix J.1. Same rows and labels as Table 38. Pin: `beam_condition_matching_2026-09-19_v2.json`.

J.4 The answer-aggregation beam on three seeds and the repaired Mode-B beam

The answer-aggregation control of the main text (§5.3) originally ran on one seed and, on Mode B, on the archived pre-repair beam. Its plan and the fixed-answer plan (Appendix J.5) call it the full-recipe control; we name it for what it grants the beam (answer-level aggregation, not our judge or walk depth). This appendix extends the identical protocol to the deduplication-repaired Mode-B beam and to all three of our seeds, under a plan that fixed the hypothesis, the four readings, the equivalence margins, the attribution licence, the gates and the verdict wording before any new extraction call (shipped as `2026-09-19_beam_perpath_answer_repaired_posthoc_anonymized.yaml`; its execution log discloses the zero-call dry run, a cache-only Mode-A rehearsal, and the review fixes made after that rehearsal). Protocol: one answer is extracted per archived final beam path with the same extractor, prompt, model and extraction seed as ours (one extraction per unique question–path pair, reused across seeds); each answer's mass is the softmax over the query's archived path LLRs summed over normalized answer keys; the beam's confidence is the top answer's mass and its label is that answer's Hit@1. Ours is the archived emitted-answer mass and label on the same (query, seed) rows. Nothing is re-traversed and no archived answer, label or confidence changes.

Gates (all passed). The beam side is seed-invariant on every query of every dataset (identical path lists and LLRs across seeds), so it has one realization per query and the three seeds vary only our side. Mode A reads the path-logging replay of the canonical beam (agreement with the canonical run on final answer and top LLR: 99.89%, 99.67%, 99.78%); its 8,197 extraction calls were all served from the single-seed control's cache and reproduced that control's three cells (300/300, 300/300, 300/300 queries), so the cache had not drifted. Mode B reads the canonical repaired records directly. Mean distinct final paths per query: HotpotQA 9.1, MuSiQue 8.8, 2WikiMHQA 9.4, WebQSP 10.0, CWQ 9.6, MetaQA-3hop 9.8 (no query with all paths identical; the pre-repair MetaQA-3hop cell of the single-seed control was degenerate). Both arms' archived AUROCs and the paired and clustered p of the registered comparison reproduced the canonical pins within 10^{-6} before any new number was computed. Extraction: 20,942 calls over the six datasets, 9,325 of them new (the rest cache replays), zero errors, zero all-path failures, zero empty top keys; the script's cache-inclusive token estimate for all calls is \$0.80 (863 s at eight workers). Our abstained rows (3 in total) enter as score 0 / label 0, as in the single-seed control. A shuffled-mass control (the beam's top mass permuted across rows with a fixed seed) stays within ten points of chance on every dataset (AUROC 51.0, 48.9, 48.2,

50.0, 50.3, 52.7). Mass-saturated beam rows (top mass = 1): HotpotQA 363/900, MuSiQue 207/900, 2WikiMHQA 375/900, WebQSP 732/1500, CWQ 588/1500, MetaQA-3hop 120/900.

Margins. Reading (ii) uses a percentile-bootstrap TOST at a per-dataset margin fixed at registration: 2.49 (the paper's 80%-power one-sided MDE multiple) times the clustered SD of the *registered* comparison on the same frame, a proxy for the new comparison's SD: HotpotQA 9.8, MuSiQue 11.3, 2WikiMHQA 10.0, WebQSP 8.5, CWQ 6.5, MetaQA-3hop 12.3 AUROC points. They are wide, and proportional to each dataset's registered noise, so a (ii) reading is a statement about this design's resolution, not evidence of practical equivalence, and the 90% clustered interval is printed beside it. An attribution sentence is licensed only for a dataset with a (ii) reading, among the three registered FDR survivors, whose margin is smaller than its registered gap, with neither a (i) nor a (iii) reading, and whose guaranteed share survives the registered gap's own sampling uncertainty (the gap's clustered 95% CI lower bound minus the margin must be positive: 2WikiMHQA $+13.2$; HotpotQA and CWQ negative, so the sentence is forbidden there whatever the outcome).

Results (Tables 40–42; every reading is within this post-hoc six-comparison family). Primary: advantage established on HotpotQA, MuSiQue, CWQ; bounded within the design's resolution on 2WikiMHQA, WebQSP; not established (MetaQA-3hop) (attribution licensed on 2WikiMHQA). Our confidence is ahead on all six pooled point estimates ($+1.3$ to $+9.3$); the single-seed control's direction is reproduced on every seed. On 2WikiMHQA the pooled difference is $+1.3$ (95% clustered CI $[-3.8, +6.1]$, 90% $[-3.2, +5.5]$), the TOST rejects at the 10.0-point margin, and the bound is tighter than the registered gap of 30.9 points, so the answer-aggregation layer accounts for at least 20.9 of those 30.9 points and independent sampling is not isolated as the cause of that part; on the registered gap's CI lower bound the share is at least 13.2 points. Both numbers are arithmetic on estimates and are reported as such. Beam-label variant: advantage established on HotpotQA, MuSiQue, CWQ, MetaQA-3hop; not established (2WikiMHQA, WebQSP) (no attribution licensed). Registered-basis variant: advantage established on HotpotQA, CWQ; bounded within the design's resolution on 2WikiMHQA, WebQSP; not established (MuSiQue, MetaQA-3hop) (attribution licensed on 2WikiMHQA). Hit@1 of the aggregated top answer / the archived beam answer / ours: HotpotQA 0.407 / 0.430 / 0.354; MuSiQue 0.150 / 0.147 / 0.127; 2WikiMHQA 0.410 / 0.373 / 0.381; WebQSP 0.636 / 0.652 / 0.629; CWQ 0.404 / 0.428 / 0.387; MetaQA-3hop 0.757 / 0.813 / 0.572. The registered comparison (Table 2) remains primary; this control replaces nothing, and the single-seed cell stays cited as the prior cell. Pin: `beam_perpath_answer_repaired_2026-09-19.json`; per-row file `beam_perpath_answer_repaired_rows_2026-09-19.jsonl` (every path's LLR and extracted answer, both arms' gold and scores), from which `--from-rows` rebuilds every table cell offline with zero LLM calls (run in the supplement's smoke check).

Dataset	Seed	n	Ours	Beam	Δ [95% CI]	p_{pair}	p_{clu}	p_{TOST}	Reading
HotpotQA	42	300	77.8	68.1	$+9.7$ [$+3.6$, $+15.7$]	.004			
	123	300	75.6	68.1	$+7.5$ [$+1.1$, $+13.1$]	.026			
	456	300	79.0	68.1	$+10.9$ [$+5.0$, $+16.9$]	$<.002$			
	pooled	900	77.4	68.1	$+9.3$ [$+4.0$, $+14.9$]	$<.002$	.001	.437	(i)
MuSiQue	42	300	77.1	70.9	$+6.1$ [-2.4, $+15.1$]	.152			
	123	300	81.3	70.9	$+10.3$ [$+3.0$, $+18.3$]	.006			
	456	300	79.8	70.9	$+8.9$ [$+0.9$, $+17.7$]	.030			
	pooled	900	79.2	70.9	$+8.3$ [$+1.3$, $+15.6$]	$<.002$	.023	.211	(i)
2WikiMHQA	42	300	76.9	76.0	$+0.9$ [-5.0, $+6.9$]	.760			
	123	300	76.3	76.0	$+0.3$ [-5.8, $+6.1$]	.876			
	456	300	78.7	76.0	$+2.7$ [-3.1, $+8.3$]	.332			
	pooled	900	77.3	76.0	$+1.3$ [-3.8, $+6.1$]	.472	.621	.001	(ii)‡ M=10.0
WebQSP	42	500	59.5	53.8	$+5.7$ [$+0.6$, $+10.7$]	.028			
	123	500	57.0	53.8	$+3.2$ [-2.1, $+8.0$]	.212			
	456	500	55.5	53.8	$+1.7$ [-3.7, $+6.6$]	.532			
	pooled	1500	57.3	53.8	$+3.5$ [-1.0, $+7.9$]	.016	.140	.012	(ii) M=8.5
CWQ	42	500	75.0	67.6	$+7.4$ [$+2.8$, $+11.8$]	.002			
	123	500	73.0	67.6	$+5.5$ [$+1.1$, $+9.8$]	.020			
	456	500	72.0	67.6	$+4.4$ [-0.3, $+8.6$]	.058			
	pooled	1500	73.3	67.6	$+5.7$ [$+1.9$, $+9.4$]	$<.002$	.005	.348	(i)
MetaQA-3hop	42	300	61.8	54.0	$+7.8$ [-1.1, $+16.5$]	.094			
	123	300	60.4	54.0	$+6.4$ [-2.5, $+15.6$]	.172			
	456	300	54.5	54.0	$+0.5$ [-8.9, $+9.9$]	.906			
	pooled	900	58.8	54.0	$+4.9$ [-3.4, $+12.7$]	.070	.252	.031	(iv)

Table 40: Repaired-beam answer-aggregation control, primary: our emitted-answer mass against the beam's top-answer mass, each arm labelled by its own answer. Confidence AUROC ($\times 100$) per seed and pooled over the three seeds; Beam = the beam with answer aggregation; Δ = ours $-$ beam. Seed rows: query-level paired bootstrap (B=1000, 95% percentile interval, two-sided p_{pair} = twice the smaller tail, no add-one; an exact zero prints as its bound). Pooled rows: query-clustered bootstrap over (query, seed) triples (B=2000, add-one tails, 95% interval; p_{clu} carries the reading), the pooled paired p_{pair} (anti-conservative by construction, since the beam arm's three seed rows per query are exact triplicates; printed for continuity, used for no reading), and p_{TOST} at the pre-registered margin M (2.49 $\times$ the clustered SD of the registered comparison on the same frame). Readings (i)–(iv) as declared in the plan, each after BH-FDR (q=.05) over the six datasets. The beam arm is seed-invariant (its archived paths and LLRs are identical across seeds), so seed-to-seed variation is ours alone. ‡attribution sentence licensed by the plan's conditions (a)–(e) and stated in the text. Pin: `beam_perpath_answer_repaired_2026-09-19.json`.

Dataset	Seed	n	Ours	Beam	Δ [95% CI]	p_{pair}	p_{clu}	p_{TOST}	Reading
HotpotQA	42	300	77.8	65.3	$+12.6$ [$+5.9$, $+18.8$]	$<.002$			
	123	300	75.6	65.3	$+10.4$ [$+4.4$, $+16.1$]	.004			
	456	300	79.0	65.3	$+13.7$ [$+7.6$, $+19.7$]	$<.002$			
	pooled	900	77.4	65.3	$+12.2$ [$+6.7$, $+18.0$]	$<.002$	.001	.801	(i)
MuSiQue	42	300	77.1	70.6	$+6.5$ [-2.0, $+15.0$]	.122			
	123	300	81.3	70.6	$+10.7$ [$+3.3$, $+18.2$]	.002			
	456	300	79.8	70.6	$+9.3$ [$+0.9$, $+17.6$]	.028			
	pooled	900	79.2	70.6	$+8.7$ [$+1.9$, $+15.8$]	$<.002$	.020	.214	(i)
2WikiMHQA	42	300	76.9	74.1	$+2.8$ [-3.7, $+9.7$]	.380			
	123	300	76.3	74.1	$+2.2$ [-4.1, $+8.8$]	.504			
	456	300	78.7	74.1	$+4.6$ [-1.6, $+11.0$]	.130			
	pooled	900	77.3	74.1	$+3.2$ [-2.5, $+8.7$]	.082	.289	.009	(iv)
WebQSP	42	500	59.5	52.7	$+6.9$ [$+1.7$, $+12.1$]	.012			
	123	500	57.0	52.7	$+4.3$ [-0.8, $+9.4$]	.088			
	456	500	55.5	52.7	$+2.8$ [-2.3, $+8.2$]	.280			
	pooled	1500	57.3	52.7	$+4.7$ [-0.1, $+9.3$]	.002	.055	.055	(iv)
CWQ	42	500	75.0	64.5	$+10.5$ [$+5.7$, $+15.2$]	$<.002$			
	123	500	73.0	64.5	$+8.5$ [$+3.9$, $+13.5$]	$<.002$			
	456	500	72.0	64.5	$+7.5$ [$+2.2$, $+11.9$]	.004			
	pooled	1500	73.3	64.5	$+8.8$ [$+4.5$, $+12.9$]	$<.002$	.001	.853	(i)
MetaQA-3hop	42	300	61.8	46.2	$+15.6$ [$+4.8$, $+24.9$]	.004			
	123	300	60.4	46.2	$+14.2$ [$+4.3$, $+24.2$]	.008			
	456	300	54.5	46.2	$+8.3$ [-2.7, $+18.9$]	.116			
	pooled	900	58.8	46.2	$+12.6$ [$+3.0$, $+21.3$]	$<.002$	.010	.526	(i)

Table 41: Repaired-beam answer-aggregation control, beam-label variant: the beam arm keeps its answer-aggregation mass but is labelled by its archived final answer. Confidence AUROC ($\times 100$) per seed and pooled over the three seeds; Beam = the beam with answer aggregation; Δ = ours $-$ beam. Seed rows: query-level paired bootstrap (B=1000, 95% percentile interval, two-sided p_{pair} = twice the smaller tail, no add-one; an exact zero prints as its bound). Pooled rows: query-clustered bootstrap over (query, seed) triples (B=2000, add-one tails, 95% interval; p_{clu} carries the reading), the pooled paired p_{pair} (anti-conservative by construction, since the beam arm's three seed rows per query are exact triplicates; printed for continuity, used for no reading), and p_{TOST} at the pre-registered margin M (2.49 $\times$ the clustered SD of the registered comparison on the same frame). Readings (i)–(iv) as declared in the plan, each after BH-FDR (q=.05) over the six datasets. The beam arm is seed-invariant (its archived paths and LLRs are identical across seeds), so seed-to-seed variation is ours alone. ‡attribution sentence licensed by the plan's conditions (a)–(e) and stated in the text. Pin: `beam_perpath_answer_repaired_2026-09-19.json`.

Dataset	Seed	n	Ours	Beam	Δ [95% CI]	p_{pair}	p_{clu}	p_{TOST}	Reading
HotpotQA	42	300	78.5	68.1	$+10.4$ [$+4.7$, $+16.2$]	.002			
	123	300	74.8	68.1	$+6.7$ [-0.2, $+12.7$]	.060			
	456	300	78.7	68.1	$+10.6$ [$+4.3$, $+16.5$]	$<.002$			
	pooled	900	77.3	68.1	$+9.2$ [$+3.8$, $+14.7$]	$<.002$	.001	.421	(i)
MuSiQue	42	300	76.6	70.9	$+5.7$ [-2.5, $+14.4$]	.182			
	123	300	81.5	70.9	$+10.6$ [$+3.1$, $+19.1$]	.006			
	456	300	77.1	70.9	$+6.2$ [-3.0, $+16.0$]	.176			
	pooled	900	78.3	70.9	$+7.4$ [-0.1, $+15.1$]	$<.002$	.053	.159	(iv)
2WikiMHQA	42	300	77.5	76.0	$+1.5$ [-4.3, $+7.6$]	.602			
	123	300	77.3	76.0	$+1.2$ [-4.7, $+7.1$]	.690			
	456	300	78.9	76.0	$+2.9$ [-2.6, $+8.3$]	.298			
	pooled	900	77.9	76.0	$+1.9$ [-3.3, $+6.7$]	.254	.472	.001	(ii)‡ M=10.0
WebQSP	42	500	59.4	53.8	$+5.6$ [$+0.5$, $+10.5$]	.032			
	123	500	56.6	53.8	$+2.8$ [-2.4, $+7.6$]	.270			
	456	500	55.1	53.8	$+1.3$ [-4.0, $+6.3$]	.626			
	pooled	1500	57.0	53.8	$+3.2$ [-1.3, $+7.6$]	.024	.179	.008	(ii) M=8.5
CWQ	42	500	75.6	67.6	$+8.1$ [$+3.4$, $+12.3$]	$<.002$			
	123	500	73.3	67.6	$+5.7$ [$+1.3$, $+10.0$]	.008			
	456	500	71.6	67.6	$+4.0$ [-0.6, $+8.2$]	.098			
	pooled	1500	73.4	67.6	$+5.9$ [$+2.0$, $+9.5$]	$<.002$	.004	.359	(i)
MetaQA-3hop	42	300	61.4	54.0	$+7.4$ [-1.6, $+15.9$]	.108			
	123	300	60.4	54.0	$+6.4$ [-2.7, $+16.0$]	.184			
	456	300	54.5	54.0	$+0.5$ [-9.2, $+9.5$]	.928			
	pooled	900	58.7	54.0	$+4.7$ [-3.4, $+12.5$]	.070	.255	.029	(iv)

Table 42: Repaired-beam answer-aggregation control, registered-basis variant: our registered top-1 mass replaces the emitted mass. Confidence AUROC ($\times 100$) per seed and pooled over the three seeds; Beam = the beam with answer aggregation; Δ = ours $-$ beam. Seed rows: query-level paired bootstrap (B=1000, 95% percentile interval, two-sided p_{pair} = twice the smaller tail, no add-one; an exact zero prints as its bound). Pooled rows: query-clustered bootstrap over (query, seed) triples (B=2000, add-one tails, 95% interval; p_{clu} carries the reading), the pooled paired p_{pair} (anti-conservative by construction, since the beam arm's three seed rows per query are exact triplicates; printed for continuity, used for no reading), and p_{TOST} at the pre-registered margin M (2.49 $\times$ the clustered SD of the registered comparison on the same frame). Readings (i)–(iv) as declared in the plan, each after BH-FDR (q=.05) over the six datasets. The beam arm is seed-invariant (its archived paths and LLRs are identical across seeds), so seed-to-seed variation is ours alone. ‡attribution sentence licensed by the plan's conditions (a)–(e) and stated in the text. Pin: `beam_perpath_answer_repaired_2026-09-19.json`.

The beam scored by unweighted answer frequency (post-hoc, archived prompts). Tables 43–44 rescore the beam of this control on the same rows with every final path counted once instead of weighted by the softmax of its LLR. Primary: the beam's answer is the most frequent answer key over its final paths (ties to the key whose first path ranks higher by LLR) and its confidence is that key's share of the paths; secondary: the share of the paths whose key equals the key of the beam's own archived final answer. This is in the spirit of the beam-weighted estimator of Fadeeva et al. (2025) with uniform weights and key equality as the similarity, not their estimator. Ours, the frames, the tests and the margins are this control's; the margins are inherited from its registered comparison, not the new scores' own resolution. The plan, including the readings and their wording, was frozen before any computation and no LLM call is made (shipped as `2026-09-25_answer_frequency_and_judge_bins_posthoc_anonymized.yaml`). Primary: advantage established on HotpotQA, MuSiQue, MetaQA-3hop; bounded within the inherited margin on 2WikiMHQA, WebQSP; not established on CWQ. Secondary: advantage established on HotpotQA, MuSiQue, CWQ, MetaQA-3hop; bounded within the inherited margin on 2WikiMHQA; not established on WebQSP. A shuffled-score control (the primary score permuted across each dataset's rows, fixed seed) stays within ten points of chance (AUROC HotpotQA 50.1, MuSiQue 49.3, 2WikiMHQA 47.0, WebQSP 50.0, CWQ 50.2, MetaQA-3hop 51.9). Hit@1 of the

most frequent answer / the softmax-top answer / the archived beam answer: HotpotQA .347 / .407 / .430; MuSiQue .127 / .150 / .147; 2WikiMHQA .357 / .410 / .373; WebQSP .594 / .636 / .652; CWQ .354 / .404 / .428; MetaQA-3hop .717 / .757 / .813. Tied-mode rows whose tied keys carry different labels: HotpotQA 3/54, MuSiQue 15/126, 2WikiMHQA 21/108, WebQSP 24/69, CWQ 33/114, MetaQA-3hop 84/168; rows with primary confidence 1: HotpotQA 360/900, MuSiQue 207/900, 2WikiMHQA 372/900, WebQSP 720/1500, CWQ 576/1500, MetaQA-3hop 120/900.

Dataset	Seed	n	Ours	Beam	Δ [95% CI]	p_{pair}	p_{clu}	p_{TOST}	Reading
HotpotQA	42	300	77.8	69.6	$+8.2$ [$+1.9$, $+14.5$]	.018			
	123	300	75.6	69.6	$+6.0$ [-0.3, $+12.2$]	.058			
	456	300	79.0	69.6	$+9.4$ [$+3.7$, $+15.6$]	.002			
	pooled	900	77.4	69.6	$+7.9$ [$+2.5$, $+13.4$]	$<.002$	.005	.271	(i)
MuSiQue	42	300	77.1	67.3	$+9.8$ [$+0.3$, $+19.4$]	.046			
	123	300	81.3	67.3	$+14.0$ [$+5.1$, $+23.9$]	.004			
	456	300	79.8	67.3	$+12.6$ [$+3.4$, $+22.7$]	.010			
	pooled	900	79.2	67.3	$+12.0$ [$+3.6$, $+20.4$]	$<.002$	.006	.577	(i)
2WikiMHQA	42	300	76.9	81.6	-4.7 [-10.4, $+1.0$]	.106			
	123	300	76.3	81.6	-5.3 [-10.9, $+0.4$]	.072			
	456	300	78.7	81.6	-2.9 [-8.3, $+2.5$]	.298			
	pooled	900	77.3	81.6	-4.3 [-9.1, $+0.3$]	.008	.066	.012	(ii) M=10.0
WebQSP	42	500	59.5	55.7	$+3.8$ [-1.2, $+8.9$]	.142			
	123	500	57.0	55.7	$+1.3$ [-3.6, $+5.8$]	.624			
	456	500	55.5	55.7	-0.2 [-5.6, $+4.9$]	.944			
	pooled	1500	57.3	55.7	$+1.6$ [-3.0, $+6.0$]	.264	.487	.001	(ii) M=8.5
CWQ	42	500	75.0	69.6	$+5.4$ [$+0.3$, $+10.3$]	.040			
	123	500	73.0	69.6	$+3.5$ [-1.6, $+8.3$]	.154			
	456	500	72.0	69.6	$+2.4$ [-2.6, $+7.0$]	.390			
	pooled	1500	73.3	69.6	$+3.7$ [-0.5, $+8.2$]	.006	.097	.096	(iv)
MetaQA-3hop	42	300	61.8	47.1	$+14.7$ [$+7.0$, $+22.2$]	$<.002$			
	123	300	60.4	47.1	$+13.3$ [$+5.5$, $+21.4$]	.002			
	456	300	54.5	47.1	$+7.4$ [-0.8, $+15.4$]	.070			
	pooled	900	58.8	47.1	$+11.8$ [$+4.9$, $+18.7$]	$<.002$	.002	.435	(i)

Table 43: The beam scored by unweighted answer frequency (post-hoc, archived prompts), primary: the beam's confidence is the share of its final paths whose answer key is the most frequent one; its label is the Hit@1 of that answer. Ours = our archived emitted-answer mass and label, as in Table 40. Confidence AUROC ($\times 100$); Δ = ours $-$ beam. Seed rows: query-level paired bootstrap (B=1000, 95% percentile interval, two-sided p_{pair}, no add-one; an exact zero prints as its bound). Pooled rows: query-clustered bootstrap (B=2000, add-one; p_{clu} carries the reading), the pooled paired p_{pair} (anti-conservative by construction, used for no reading) and p_{TOST} at Appendix J.4's pinned margin M for the same frame (inherited from that comparison, not this variant's own resolution). Readings (i)–(iv) after BH-FDR (q=.05) over the six datasets; (i)+(ii): the TOST also rejects. No attribution sentence is licensed. Pin: `review_posthoc_z1z2z3_2026-09-25.json`.

Dataset	Seed	n	Ours	Beam	Δ [95% CI]	p_{pair}	p_{clu}	p_{TOST}	Reading
HotpotQA	42	300	77.8	62.7	+15.1 [+8.1, +21.4]	<.002			
	123	300	75.6	62.7	+12.9 [+6.4, +19.4]	<.002			
	456	300	79.0	62.7	+16.3 [+9.8, +22.5]	<.002			
	pooled	900	77.4	62.7	+14.7 [+9.2, +20.2]	<.002	.001	.960	(i)
MuSiQue	42	300	77.1	71.1	+6.0 [−2.0, +14.6]	.132			
	123	300	81.3	71.1	+10.2 [+2.9, +17.8]	.008			
	456	300	79.8	71.1	+8.8 [+0.2, +17.0]	.048			
	pooled	900	79.2	71.1	+8.2 [+1.3, +14.8]	.002	.018	.193	(i)
2WikiMHQA	42	300	76.9	76.4	+0.5 [−5.3, +6.1]	.838			
	123	300	76.3	76.4	−0.1 [−6.3, +5.6]	.994			
	456	300	78.7	76.4	+2.3 [−3.4, +8.0]	.392			
	pooled	900	77.3	76.4	+0.9 [−4.5, +5.7]	.544	.738	<.001	(ii) M=10.0
WebQSP	42	500	59.5	52.6	+6.9 [+1.9, +11.7]	.004			
	123	500	57.0	52.6	+4.3 [−0.5, +9.6]	.090			
	456	500	55.5	52.6	+2.8 [−2.1, +8.1]	.302			
	pooled	1500	57.3	52.6	+4.7 [+0.2, +9.3]	.002	.044	.051	(iv)
CWQ	42	500	75.0	63.5	+11.6 [+6.7, +16.4]	<.002			
	123	500	73.0	63.5	+9.6 [+5.0, +14.8]	<.002			
	456	500	72.0	63.5	+8.5 [+3.8, +13.4]	<.002			
	pooled	1500	73.3	63.5	+9.9 [+5.8, +14.1]	<.002	.001	.946	(i)
MetaQA-3hop	42	300	61.8	39.6	+22.2 [+12.8, +30.4]	<.002			
	123	300	60.4	39.6	+20.8 [+12.5, +29.3]	<.002			
	456	300	54.5	39.6	+14.9 [+5.9, +23.6]	<.002			
	pooled	900	58.8	39.6	+19.3 [+11.2, +26.7]	<.002	.001	.961	(i)

Table 44: The beam scored by unweighted answer frequency (post-hoc, archived prompts), secondary: the beam's confidence is the share of its final paths that agree with its own archived final answer; its label is the Hit@1 of that answer. Ours = our archived emitted-answer mass and label, as in Table 40. Confidence AUROC ($\times 100$); Δ = ours − beam. Seed rows: query-level paired bootstrap (B=1000, 95% percentile interval, two-sided p_{pair}, no add-one; an exact zero prints as its bound). Pooled rows: query-clustered bootstrap (B=2000, add-one; p_{clu} carries the reading), the pooled paired p_{pair} (anti-conservative by construction, used for no reading) and p_{TOST} at Appendix J.4's pinned margin M for the same frame (inherited from that comparison, not this variant's own resolution). Readings (i)–(iv) after BH-FDR (q=.05) over the six datasets; (i)+(ii): the TOST also rejects. No attribution sentence is licensed. Pin: `review_posthoc_z1z2z3_2026-09-25.json`.

J.5 Fixed-answer grafting on the repaired beam: five estimators at the beam's own answer

The grafting control of Table 25 scores the beam's archived answer by our posterior's mass at that answer, on the archived pre-repair Mode-B beam; the answer-aggregation control of Appendix J.4 lets the beam change its answer. This subsection holds the answer and its label fixed (the canonical beam record's final answer and its Hit@1; Mode B = the deduplication-repaired beam) and reads five confidence estimators at that answer on the same (query, seed) rows, under a plan that fixed the estimators, the contrasts, the three readings, the zero-mass rule, the gates and the wording before any computation (shipped as `2026-09-20_beam_answer_fixed_confidence_posthoc_anonymized.yaml`). E0 is the beam's archived σ(LLR); E1 the beam's own path-level answer aggregation (softmax over its archived path LLRs, mass per normalized answer key) read at the archived answer; E2 our archived particle posterior read at that answer (keys compared as stored, one normalizer pass on the answer); E3, the uniform vote share of that answer among all our particles, was NOT run: its replay gate requires the cache guard to see LLM traffic for every replayed query (no traffic is read as a bypassed guard) and stopped on CWQ, where the archived run itself answers 1 query (3 rows of Table 19's replay) with no particle and no LLM call, a case the rule cannot tell from a bypassed guard (HotpotQA, MuSiQue, 2WikiMHQA, WebQSP had replayed with parity 100.0% and no cache miss); the plan allows no re-attempt for E3, and the analysis reports four estimators and five contrasts, as its plan provides; E1′ our own terminal judge (the launcher's paraphrase-median yes/no LLR ex-

actly as the archived run used it: same model, API seed, as-sent prompt template (archived form), empty evidence and logprob floor of -10, the archived paraphrases read from the cache) applied to the beam's own archived path strings, softmax over the medians and the same per-path answers as E1, read at that answer; this needed new judge calls (62,813 distinct cache keys of the census's 62,813: every unique query–path pair times three paraphrases, 0 already cached), counted by a zero-call census before any call and made on 2026-09-20 under a cap of $8.00 (estimate $2.16; 0 retried calls); the set of keys written equals the census's set and every census key was cached afterwards. Before them, 576 replayed particle paths of 18 rows were re-scored through the same code path and reproduced the archived run's medians to 10^{-9}, and all 20,942 archived beam path LLRs were reproduced from the cache by the same scorer called once on the original question under the beam's floor of -20 (the scorer's default, which the beam inherited), so E1′−E1 changes the judging alone: three paraphrases, the median and our floor against one call under the beam's floor. The floor by itself would change 3,037 of those beam LLRs (largest change 10.0). The beam's calls were made in 2026-07 and E1′'s now; the served model behind the same model string is not controlled between the two dates. The scorer substitutes $\log 0.5$ for a missing yes/no logprob without raising (an LLR of exactly 0 or ± 2): 11,162 of the beam's LLRs and 25,573 of the judge's carry that signature. E1 and E2 differ in the path set (ten beam-search paths against $K{=}32$ sampled paths), the path count, the path text (the beam's paths carry no relation labels) and the judging (the beam's LLR is the same scorer called once on the original question; ours is its paraphrase median), so E2−E1 is the particle posterior against the beam's own aggregation as a whole; it isolates none of the four. E2−E1′ and E1′−E1 decompose it exactly: the first changes the path set, count and text at the same judge, the second changes only the judging on the beam's own paths. Frame: the 6,600 rows of Appendix J.4; gates reproduce Table 2's beam column (E0) and Table 25's Mode-A cells (E2, zero-mass rates) to 10^{-6}; on 5 Mode-A rows the path-logging replay had ended on a different final answer than the canonical run (a disclosed cache-drift artefact) and the canonical answer is used. A row whose answer is absent from an estimator's support is scored 0 and kept; no exclusion variant is computed.

Results (Tables 45–48; every reading within its own post-hoc six-comparison family, query-clustered two-sided p with BH-FDR $q{=}.05$; the pooled paired p is anti-conservative where an arm is seed-invariant and is kept in the pin only). Primary contrast E2−E1, the particle posterior against the beam's own aggregation at a fixed answer and label: not established on HotpotQA, MuSiQue, 2WikiMHQA, WebQSP, CWQ, MetaQA-3hop (Δ -9.4 to $+2.6$). E2−E1′ (the sampled path set, count and text at the same judge): not established on HotpotQA, MuSiQue, 2WikiMHQA, WebQSP, CWQ, MetaQA-3hop. E1′−E1 (the paraphrase-median judging on the beam's own paths): not established on HotpotQA, MuSiQue, 2WikiMHQA, WebQSP, CWQ, MetaQA-3hop. E2−E0 (the graft of Table 25 on the repaired beam): advantage on 2WikiMHQA; reversal on MetaQA-3hop; not established on HotpotQA, MuSiQue, WebQSP, CWQ. E1−E0 (the beam's own aggregation against its σ(LLR) at its own answer): advantage on 2WikiMHQA; reversal on MetaQA-3hop; not established on HotpotQA, MuSiQue, WebQSP, CWQ. E2−E3: not run. Zero-mass rows (the answer absent from the estimator's support, Table 45): E1 and E1′ (same support) on 12–40% of rows, E2 on 16–54%; inside E2's zero set the beam is correct on 0.070–0.873 of rows against 0.176–0.743 outside it (per dataset). These are the fixed sentences of the plan. The two beam-side controls ask different questions (Appendix J.4 lets the beam change its answer under answer aggregation; this one holds its answer fixed), so their readings stand side by side and neither revises the other: the analysis neither attributes the registered 2WikiMHQA gap nor revises any sentence of Appendix J.4. It prints 48 p-values (30 clustered, 18 per-seed paired), added to the count of Appendix D9.

Dataset	n	E0	E1	E1′	E2	E3	0_{E1}	0_{E2}	0_{E3}	Hit@1$\|_{E2=0}$	Hit@1$\|_{E2>0}$	agree
HotpotQA	900	65.6	67.1	66.0	67.7	–	12%	16%	–	0.304	0.455	55%
MuSiQue	900	77.2	77.6	76.5	68.2	–	35%	27%	–	0.070	0.176	36%
2WikiMHQA	900	46.4	77.8	77.0	77.1	–	16%	26%	–	0.143	0.453	48%
WebQSP	1500	59.3	53.4	52.6	56.0	–	17%	23%	–	0.601	0.667	58%
CWQ	1500	65.9	65.9	65.6	68.0	–	15%	18%	–	0.279	0.460	54%
MetaQA-3hop	900	68.0	41.0	40.9	39.7	–	40%	54%	–	0.873	0.743	25%

Table 45: Fixed-answer grafting on the repaired beam: the five estimators read at the beam's own archived answer, scored against that answer's correctness (AUROC $\times 100$, pooled over three seeds). E0 = the beam's archived σ(LLR); E1 = the beam's own path-level answer-aggregation mass at that answer (Appendix J.4); E1′ = our paraphrase-median judge applied to the beam's own archived paths (new judge calls; softmax over the medians, the same per-path answers as E1, so its support and zero set are E1's); E2 = our particle posterior's mass at that answer (the graft of Table 25, here on the repaired Mode-B beam); E3 = the uniform vote share of that answer among all our particles (E2 under uniform weights). 0_E: rows whose answer is absent from that estimator's support, scored 0 and kept (E0 has none by construction); Hit@1 of the beam inside and outside E2's zero set; agree: rows whose emitted answer equals the beam's. E0, E1 and E1′ are seed-invariant; E2 and E3 vary with our seed. The Mode-A E2 and E0 cells equal Table 25's by construction (same rows, answers and labels); the tests differ (two-sided clustered here, one-sided paired there). Pin: `beam_answer_fixed_confidence_2026-09-20.json`. E3 was not run (its replay stopped under its plan's rules; see the text).

Dataset	E2−E1	E2−E1′	E1′−E1
HotpotQA	+0.6 [−4.3, +5.5] .791	+1.8 [−3.0, +6.5] .483	−1.2 [−2.8, +0.5] .171
MuSiQue	−9.4 [−17.5, −1.6] .015	−8.3 [−16.3, −0.6] .033	−1.1 [−3.9, +1.5] .388
2WikiMHQA	−0.7 [−4.9, +3.6] .802	+0.1 [−3.8, +4.1] .929	−0.8 [−2.2, +0.8] .316
WebQSP	+2.6 [−0.4, +5.6] .095	+3.4 [+0.5, +6.4] .016	−0.8 [−1.9, +0.3] .173
CWQ	+2.2 [−1.0, +5.3] .214	+2.5 [−0.7, +5.5] .118	−0.3 [−1.9, +1.3] .688
MetaQA-3hop	−1.3 [−6.7, +3.9] .625	−1.2 [−6.9, +4.3] .639	−0.1 [−3.2, +3.0] 1.000

Table 46: Fixed-answer grafting, the primary contrast and its exact decomposition: Δ AUROC [clustered 95% CI] and the query-clustered two-sided p (B=2000, add-one), each column its own BH-FDR family (q=.05) over the six datasets. E2−E1 is the pre-registered primary contrast: the particle posterior against the beam's own aggregation at a fixed answer and label (path set, path count, path text and judging all differ; nothing is isolated). E2−E1′ (the sampled path set, its count and its relation-labelled text, at the same judge) and E1′−E1 (the paraphrase-median judging on the beam's own paths: three paraphrases, the median and our logprob floor of -10 against the beam's one call on the original question and its floor of -20, nothing else) sum to E2−E1 exactly before rounding (AUROC differences telescope; the printed cells are rounded) and are read in their own families, not as a second chance at the primary. §advantage established (reading (i)); ¶reversal (reading (iii)); otherwise not established. No equivalence reading and no margin are defined for this analysis. Per-seed rows of the primary contrast: Table 48.

Dataset	E2−E0	E1−E0
HotpotQA	$+2.1$ [-6.9, $+10.7$] .654	$+1.5$ [-6.7, $+10.0$] .755
MuSiQue	-9.0 [-20.1, $+2.8$] .118	$+0.4$ [-7.3, $+9.0$] .927
2WikiMHQA	$+30.8^{\S}$ [$+21.0$, $+40.6$] .001	$+31.4^{\S}$ [$+23.4$, $+39.7$] .001
WebQSP	-3.3 [-10.9, $+4.4$] .418	-5.9 [-13.6, $+1.9$] .133
CWQ	$+2.1$ [-3.7, $+7.6$] .469	-0.1 [-5.5, $+5.5$] .998
MetaQA-3hop	$-28.3^{\P}$ [-38.8, -17.0] .001	$-27.0^{\P}$ [-36.7, -16.7] .001

Table 47: Fixed-answer grafting, the remaining contrasts: Δ AUROC [clustered 95% CI] and the query-clustered two-sided p (B=2000, add-one), each column its own BH-FDR family (q=.05) over the six datasets. E2−E0 is the graft of Table 25 on the repaired beam; E1−E0 the beam's own aggregation against its σ(LLR) at its own answer. E2−E3 was not run (see the text). §advantage established (reading (i)); ¶reversal (reading (iii)); otherwise not established. No equivalence reading and no margin are defined for this analysis.

Dataset	Seed	n	E2	E1	Δ [95% CI]	p	Reading
HotpotQA	42	300	67.3	67.4	-0.1 [-5.3, $+5.1$]	.988	
	123	300	67.7	67.0	$+0.8$ [-4.2, $+5.7$]	.780	
	456	300	68.2	67.0	$+1.2$ [-4.2, $+6.7$]	.690	
	pooled	900	67.7	67.1	$+0.6$ [-4.3, $+5.5$]		(iv)
MuSiQue	42	300	68.3	77.2	-8.9 [-17.2, -1.2]	.022	
	123	300	69.9	77.7	-7.8 [-16.0, $+0.2$]	.054	
	456	300	66.5	78.0	-11.5 [-20.2, -3.0]	.004	
	pooled	900	68.2	77.6	-9.4 [-17.5, -1.6]		(iv)
2WikiMHQA	42	300	76.2	77.8	-1.6 [-6.3, $+3.0$]	.518	
	123	300	77.7	77.8	-0.1 [-4.4, $+4.6$]	.942	
	456	300	77.6	77.8	-0.2 [-4.6, $+4.4$]	.986	
	pooled	900	77.1	77.8	-0.7 [-4.9, $+3.6$]		(iv)
WebQSP	42	500	56.2	53.4	$+2.8$ [-0.4, $+6.0$]	.096	
	123	500	55.4	53.4	$+2.1$ [-0.9, $+5.1$]	.204	
	456	500	56.3	53.4	$+2.9$ [-0.3, $+6.1$]	.074	
	pooled	1500	56.0	53.4	$+2.6$ [-0.4, $+5.6$]		(iv)
CWQ	42	500	69.1	65.9	$+3.2$ [-0.2, $+6.6$]	.064	
	123	500	66.4	65.9	$+0.6$ [-2.9, $+4.1$]	.726	
	456	500	68.6	65.9	$+2.7$ [-0.4, $+5.8$]	.102	
	pooled	1500	68.0	65.9	$+2.2$ [-1.0, $+5.3$]		(iv)
MetaQA-3hop	42	300	40.5	41.0	-0.5 [-6.4, $+5.1$]	.856	
	123	300	37.9	41.0	-3.2 [-9.3, $+2.8$]	.288	
	456	300	40.7	41.0	-0.3 [-6.2, $+5.2$]	.896	
	pooled	900	39.7	41.0	-1.3 [-6.7, $+3.9$]		(iv)

Table 48: Fixed-answer grafting, primary contrast E2−E1 per seed. Seed rows: query-level paired bootstrap (B=1000, 95% percentile interval, two-sided p = twice the smaller tail, no add-one; an exact zero prints as its bound). Pooled rows: the clustered 95% CI and the reading; their p is printed once, in Table 46. E1 is seed-invariant, so seed-to-seed variation is ours alone.

K Pre-Registered Rerun with the Prompts Rendered as Written

On the six datasets, the per-dataset pattern of Table 2 under the rendered prompts (each dataset's mark and the sign of its Δ) is the same as the archived pattern recomputed on the same datasets (Table 49). The rerun is a supplementary validation registered before any rendered call was billed; the archived run of the main text stays primary and none of its numbers is replaced. The plan fixed the datasets and their order, the seeds, the gates, the readouts, the stopping rules and the wording of every sentence that reports the outcome (its reading rules); the plan and its execution log ship anonymized in the code supplement.

The defect and the one change. Every archived judge and extraction call sent its template unrendered (Appendix C, judge item): the templates name the slots {query} and {path_text}, the wrapper fills query, evidence and path, and on the resulting key error it sends the literal template, braces included, followed by the filled fields. The rerun renders every template that names a slot and lets a rendering error propagate. Every other setting is the archived one and is checked against the archived rows: model, temperature, API seeds, token limits, log-probability floors, imputation rule, K, paraphrases, beam width and depth, stop check, deduplication, samples, run seeds and each dataset's beam configuration. As sent under the rerun (newlines shown as /), the judge prompt is `Question: ⟨question⟩ / Reasoning path: ⟨path text⟩ / / Does this path support the answer? Answer yes or no: / / Answer with exactly one word: Yes or No. / Answer:` and the extraction prompt is `Question: ⟨question⟩ / Reasoning path: ⟨path text⟩ / / Provide a SHORT direct answer (entity name, year, number, or short phrase — NO explanation, NO additional words):`. The beam's stop-check template names no slot and keeps its archived form. The rendered judge requests the top 20 log-probabilities (archived: 5); the primary LLR reads the top 5 entries of the returned list with the archived rule, which is what the archived request returns, and the top-20 read is secondary S1. Rendered calls use their own cache namespace, so none is served an archived response. Three consequences of rendering as written were declared before any call. (a) The extraction template has no evidence slot: our per-particle extraction always passed empty evidence and loses nothing, but the beam's final extraction passed its joined final-beam paths as evidence, which the archived fallback appended and the rendered prompt drops, so under the rerun the beam's final answer is extracted from the question and its top path only, the input form of our per-particle extraction (secondary S2 restores the archived form). (b) The rendered extraction prompt ends with its instruction line, where the archived fallback ended with an "Answer:" cue (both methods). (c) The stop check keeps its archived form, but its responses are new samples, and its evidence changes wherever the rendered judge changes the beam's traversal.

Gates. Before any billed call, G1/G2 found both templates rendered with no slot left and the samples equal to the archived ones on all 2,200 queries (question, gold list, seed entities and archived paraphrases, frozen by a content digest that every rendered row carries); G3 replayed the archived pipeline from the archived cache alone over all 13,200 archived (query, configuration, seed) rows: 13,194 reproduce exactly, and the other 6 are the predeclared Mode-A beam rows that differ, consistent with a cache race in the archived run (none unexplained); G5 ran the registered readout on the archived rows and reproduced the archived pins of Tables 1–2 (all 153 checks). G4, a billed smoke of the first two sampled queries of each dataset (seed 42, both configurations), returned 24/24 valid rows with no slot left in any sent prompt. Each gate record binds the code it ran by hash, and any code change reruns the gates and the smoke.

Execution. The sweep ran as one session with 80 concurrent workers for 153 minutes and stopped by itself after its final pass. The smoke and the sweep made 1,122,916 billed calls (930,475 judge, 192,441 extraction; 111,541,051 prompt and 1,838,270 completion tokens), none failing after its last attempt (4 failed attempts were retried); the spend ledger records USD 47.56 (reported usage, plus 4 attempts without a usage line counted at their upper bound), under the registered cap of USD 100. Rows per dataset: Table 54.

Results. Confidence (R1, Tables 49–50): over the six datasets, ours scores 66.4 and the beam 58.7 under the rendered prompts, against 70.6 and 63.7 on the archived rows of the same datasets (the main text's conventions: ours, the mean of the per-seed AUROCs; the beam, the mean of the matched pooled AUROCs; ours on the matched pooled basis: 66.5 rendered, 70.6 archived). Marked datasets by the rule of Table 2, under the rendered prompts: significant: HotpotQA, 2WikiMHQA, CWQ; reversed: MetaQA-3hop. On the archived rows: significant: HotpotQA, 2WikiMHQA, CWQ; reversed: MetaQA-3hop. The archived pins' one-sided mechanics, printed as a secondary line, mark HotpotQA, 2WikiMHQA, CWQ under the rendered prompts and HotpotQA, 2WikiMHQA, CWQ on the archived rows. Accuracy (R2, Table 51): by the rule of Table 1, ours wins, ties and loses on 0/2/4 datasets under the rendered prompts and 0/1/5 on the archived rows; the rendered beam's answers come from a top-path-only extraction (consequence (a); the Beam (S2) column restores the archived extraction form), and no accuracy gain is claimed from this run.

Secondary readouts. D1, the imputation census of the rendered run: of the 930,475 distinct rendered judge prompts (925,276 path judgments of both methods, 5,199 beam stop checks), the first-token top 5 lacks "Yes" on 3.1%, "No" on <0.1% and both on 0% (top 20: <0.1%, 0%, 0%); the spend meter, over 930,475 billed judge calls, counts 28,979 without "Yes" and 6 without "No" in the top 5. The archived run's census under the same rule is in Appendix C. S1 (Table 52): our rendered rows replayed from the rendered cache alone (6,600 rows, 1,688,832 cache reads, no miss, no network attempt; the replay at the primary top-5 read first reproduced every stored row), with the judge LLR read from the top 20; marked by the rule of Table 2: significant: HotpotQA, 2WikiMHQA, CWQ; reversed: MetaQA-3hop. S1's first run stopped before replaying any row: the network refusal it installs before loading its inputs also blocked the dataset library's fallback to its local copy of the inputs. The loader was fixed after the primary readout, with the replay and analysis code unchanged, so S1's record carries the script's own mark that it postdates the readout. S2 (Tables 51 and 53): the beam's final answer re-extracted on the rendered traversal with the archived as-sent extraction form as a whole (the literal template, the appended question, evidence and top path, and the "Answer:" cue), so it measures the joint effect of that form and does not isolate the evidence; 2,200 distinct requests (seeds sharing a traversal share one), 0 beam rows with an empty final beam excluded, USD 0.23. With these answers and the beam's confidence unchanged, the marks are significant: HotpotQA, 2WikiMHQA, CWQ; reversed: MetaQA-3hop; Hit@1 verdicts (ours, tie, beam): 0/3/3.

Every other analysis in these appendices, including the imputation census and its seven-variant replay (Appendix C) and the controls of Appendices F–J, uses the archived rows and prompts.

Dataset	Prompts	n	Particle-KGRAG	Beam	Δ [95% CI]	p_{pair}	p_{clu}	inv.
HotpotQA	archived	900	77.4 ± 1.7	65.6 ± 0.0	$+11.8^{*}$ [+7.5, +16.8]	.002	.004	0
	rendered	900	74.4 ± 1.5	61.1 ± 0.0	$+13.2^{*}$ [+8.5, +17.8]	.002	.001	0
MuSiQue	archived	900	79.2 ± 2.1	77.2 ± 0.5	+2.0 [−3.7, +7.7]	.472	.673	0
	rendered	900	72.3 ± 1.0	66.7 ± 0.0	+5.7 [−0.2, +12.0]	.064	.254	0
2WikiMHQA	archived	900	77.3 ± 1.3	46.4 ± 0.2	$+30.9^{*}$ [+26.0, +35.7]	.002	.001	0
	rendered	900	72.3 ± 1.5	36.8 ± 0.0	$+35.5^{*}$ [+30.3, +41.2]	.002	.001	0
WebQSP	archived	1,500	57.3 ± 2.1	59.3 ± 0.0	−1.9 [−6.0, +1.8]	.318	.579	0
	rendered	1,500	56.5 ± 0.4	58.4 ± 0.0	−1.8 [−5.5, +2.3]	.402	.590	0
CWQ	archived	1,500	73.3 ± 1.5	65.9 ± 0.0	$+7.4^{*}$ [+4.3, +10.6]	.002	.005	0
	rendered	1,500	70.1 ± 1.9	61.6 ± 0.0	$+8.5^{*}$ [+5.2, +12.1]	.002	.001	0
MetaQA-3hop	archived	900	58.8 ± 3.9	68.0 ± 0.0	$-9.2^{\dagger}$ [−15.0, −3.2]	.002	.063	0
	rendered	900	53.2 ± 2.0	67.7 ± 0.0	$-14.5^{\dagger}$ [−19.9, −9.0]	.002	.001	0

Table 49: Rendered-prompt rerun, confidence (R1): the registered comparison of Table 2 recomputed by the functions that produced it, on the rendered rows and on the archived rows of the same datasets. n: matched (query, seed) pairs; AUROC ($\times 100$) pooled over them, $\pm$ the std of the three per-seed AUROCs; Δ = ours − beam with the paired-bootstrap 95% percentile interval (B=1000). Two-sided p = twice the smaller add-one tail: p_{pair} paired, p_{clu} query-clustered (B=2000; inv. = resamples with one label class, dropped). Marks as in Table 2, BH-FDR over the read-out datasets within each block: *both rejections with $\Delta > 0$; †the paired rejection with $\Delta < 0$.

Dataset	Prompts	top-1 mass	Beam	Δ [95% CI]	p_{pair}	p_{clu}
HotpotQA	archived	77.3	65.6	+11.6 [+7.3, +16.7]	.002	.004
	rendered	74.3	61.1	+13.2 [+8.6, +17.8]	.002	.001
MuSiQue	archived	78.3	77.2	+1.1 [−4.7, +7.1]	.719	.826
	rendered	72.3	66.7	+5.7 [−0.2, +12.1]	.060	.256
2WikiMHQA	archived	77.9	46.4	+31.5 [+26.8, +36.3]	.002	.001
	rendered	72.2	36.8	+35.4 [+30.1, +41.1]	.002	.001
WebQSP	archived	57.0	59.3	−2.2 [−6.4, +1.6]	.246	.524
	rendered	56.6	58.4	−1.8 [−5.4, +2.5]	.418	.603
CWQ	archived	73.4	65.9	+7.5 [+4.4, +10.6]	.002	.004
	rendered	69.8	61.6	+8.1 [+4.8, +11.8]	.002	.002
MetaQA-3hop	archived	58.7	68.0	−9.3 [−15.1, −3.1]	.002	.064
	rendered	53.1	67.7	−14.6 [−19.9, −9.1]	.002	.001

Table 50: Rendered-prompt rerun, the registered variant (Table 3): our score is the top-1 posterior mass instead of $\rho(a_{\text{emit}})$, on the same pairs and draws as Table 49; no marks are computed for this variant.

Dataset	Prompts	n	Particle-KGRAG	Beam	Δ (pp) [95% CI]	Verdict	Beam (S2)
HotpotQA	archived	900	.354	.430	-7.6^{***} [−10.2, −5.0]	beam	
	rendered	900	.380	.443	-6.3^{***} [−8.9, −3.7]	beam	.440 (beam)
MuSiQue	archived	900	.127	.148	-2.1^{*} [−4.1, −0.2]	beam	
	rendered	900	.152	.180	-2.8^{**} [−4.8, −0.8]	beam	.203 (beam)
2WikiMHQA	archived	900	.381	.373	+0.8 [−2.0, +3.4]	tie	
	rendered	900	.393	.467	-7.3^{***} [−10.2, −4.6]	beam	.417 (tie)
WebQSP	archived	1,500	.629	.652	-2.3^{**} [−4.2, −0.5]	beam	
	rendered	1,500	.629	.624	+0.5 [−1.1, +2.1]	tie	.646 (tie)
CWQ	archived	1,500	.387	.428	-4.1^{***} [−6.0, −2.2]	beam	
	rendered	1,500	.440	.432	+0.8 [−0.9, +2.7]	tie	.442 (tie)
MetaQA-3hop	archived	900	.572	.813	-24.1^{***} [−27.4, −20.8]	beam	
	rendered	900	.664	.720	-5.6^{***} [−8.7, −2.6]	beam	.737 (beam)

Table 51: Rendered-prompt rerun, accuracy (R2): Hit@1 by the rule of Table 1 on the rendered rows and on the archived rows of the same datasets (n matched pairs; one-sided paired bootstrap in the observed direction, B=1000: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; verdict by the 95% interval). Under the rendered prompts the beam's final answer is extracted from the question and its top path only (declared consequence (a)); Beam (S2) is its Hit@1 with the archived extraction form restored on the rendered traversal, with the verdict against ours. No accuracy gain is claimed from this run.

Dataset	Particle-KGRAG (top 20)	Beam	Δ [95% CI]	p_{pair}	p_{clu}
HotpotQA	72.7	61.1	$+11.6^{*}$ [+6.7, +16.1]	.002	.005
MuSiQue	76.4	66.7	+9.8 [+3.7, +15.9]	.004	.054
2WikiMHQA	68.5	36.8	$+31.7^{*}$ [+26.5, +37.2]	.002	.001
WebQSP	57.1	58.4	−1.2 [−4.9, +2.8]	.565	.701
CWQ	70.4	61.6	$+8.8^{*}$ [+5.4, +12.3]	.002	.001
MetaQA-3hop	53.2	67.7	$-14.5^{\dagger}$ [−19.9, −9.0]	.002	.001

Table 52: Rendered-prompt rerun, secondary S1: our judge LLR read from the top 20 returned log-probabilities (the primary reads the top 5), by cache replay of the rendered run; the beam side is the rendered beam unchanged. Columns and marks as in Table 49 (rendered rows).

Dataset	Particle-KGRAG	Beam (S2)	Δ [95% CI]	p_{pair}	p_{clu}
HotpotQA	74.4	60.5	$+13.9^{*}$ [+9.2, +18.8]	.002	.001
MuSiQue	72.3	65.5	+6.8 [+0.8, +13.1]	.032	.163
2WikiMHQA	72.3	40.5	$+31.7^{*}$ [+26.4, +37.5]	.002	.001
WebQSP	56.5	58.6	−2.1 [−5.7, +2.1]	.330	.550
CWQ	70.1	61.8	$+8.3^{*}$ [+4.9, +11.6]	.002	.001
MetaQA-3hop	53.2	72.2	$-19.0^{\dagger}$ [−24.6, −13.3]	.002	.001

Table 53: Rendered-prompt rerun, secondary S2: the beam's final answers re-extracted with the archived as-sent extraction form on the rendered traversal (labels follow the new answers; the beam's confidence is unchanged); ours is the rendered primary. Columns and marks as in Table 49 (rendered rows).

Dataset	Rows	Valid	Error	Invalid	Missing	LLM-free	No conf.	Read out
HotpotQA	1,800	1,800	0	0	0	0	0	yes
MuSiQue	1,800	1,800	0	0	0	0	0	yes
2WikiMHQA	1,800	1,800	0	0	0	0	0	yes
WebQSP	3,000	3,000	0	0	0	0	0	yes
CWQ	3,000	3,000	0	0	0	3	0	yes
MetaQA-3hop	1,800	1,800	0	0	0	0	0	yes

Table 54: Rendered-prompt rerun, rows per dataset (both configurations, three seeds). LLM-free: valid rows of ours that abstain before any call (accepted only where the archived row does the same); No conf.: valid beam rows without a confidence. A dataset is read out iff no row is missing and error plus invalid rows stay below 1% of its rows (the registered rule); the readout lists every excluded row with its status.