STATEFUL PRIVACY-UTILITY MEASUREMENT FOR GRAPHRAG EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) use GraphRAG to support structured reasoning by grounding answers in graph evidence. However, the same evidence that supports an authorized answer may also help reveal protected facts when combined with information from earlier interactions. Editing evidence to reduce this risk can change the generated answer, creating a history-dependent tradeoff between privacy and answer value. Existing work on release control, evidence transformation, and attribution leaves a gap in jointly comparing these effects for complete regenerated outputs under the same prior information. We introduce **Stateful Privacy Funnel (SPriF)**, an offline framework for state-conditioned comparison of complete answer–evidence releases. SPriF regenerates answers for each evidence choice and a matched evidence-removal baseline while fixing the question, history, and other current evidence. A frozen LLM auditor measures incremental privacy cost from complete outputs, while a frozen reader separately measures reference-answer log-likelihood gains under the corresponding evidence contexts. Within a fixed graph-admissible inventory, the conditional funnel identifies the minimum auditor-measured cost meeting a specified reader-value requirement. Experiments on 8,206 evidence transformations across three KGQA benchmarks show that evidence-only scoring can miss minimum-cost complete releases, while relevant history can make history-free choices fail the same reader-value requirement. Under empty history and no other current evidence, post-hoc analysis examines admissible original releases with positive privacy cost and reference-hit answers. In 73.3% of these cases, lower-cost alternatives also generate reference hits and provide positive reader value. To reduce the cost of scoring candidate releases, we evaluate partial measurement with learned or mean filling. Under a fixed 64-action measurement budget, full-inventory privacy-ranking Spearman reaches $\rho \approx 0.75$ with 39.2% fewer test-time direct-scoring calls. Code is available at `https://anonymous.4open.science/r/SPriF`.

1 INTRODUCTION

Large language models (LLMs) use retrieval-augmented generation (RAG) to ground answers in external knowledge. Within this paradigm, GraphRAG has attracted attention for using explicit relationships between entities to support structured reasoning (Li et al., 2025; Zhang et al., 2026). It provides triples, paths, and subgraphs as evidence for question answering. However, the same evidence that supports an authorized answer may also help reveal protected facts when combined with other visible information. Shared Strava routes, for example, could jointly expose protected locations even after nearby segments were hidden (Hassan et al., 2018). An authorized but privacy-curious GraphRAG user may similarly connect answers and their supporting facts across interactions. Privacy evaluation must therefore assess the incremental effect of a complete answer-evidence release in light of previously visible information, alongside its value for answering the current question.

Limitations of existing methods. Existing work addresses stateful disclosure control, evidence transformation, and response attribution. These approaches, however, evaluate different objects: disclosures across interactions, retained facts, or context contributions to an existing answer. In GraphRAG, editing evidence can change both the facts exposed to the user and the answer generated from them. Evaluating retained facts or attributing the original answer therefore does not fully characterize the privacy-utility tradeoff of the regenerated release. The remaining gap is how to

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.

compare the incremental privacy cost and reader value of candidate releases under the same prior information, and how to approximate this comparison efficiently. This gives rise to three challenges.

Challenge I: Measuring incremental privacy cost under prior information. Stateful release-control protocols assess disclosures across interactions but do not jointly score regenerated answers (Xie & Li, 2026; Xu et al., 2026). In GraphRAG, the joint privacy effect of an answer and its evidence depends on prior information. For example, a route's starting point may become a home-location clue when an earlier statement says the route begins near home. Evidence-only scoring omits the answer's contribution, while history-free scoring overlooks interactions with prior information. The challenge is to isolate a focal evidence choice's incremental privacy cost by comparing complete releases while fixing history, the question, and all other current evidence.

Challenge II: Comparing admissible releases under a reader-value requirement. Privacy reduction alone can favor removing evidence needed to answer the question. Content-reduction and truth-preserving methods preserve query results or factual truth (Hu & Sintos, 2026; Pilán et al., 2025), but neither establishes the value of a regenerated answer. Graph-admissible facts may be irrelevant to the question, and even sufficient context need not yield a correct generation (Joren et al., 2025). Reference-answer likelihood provides a graded comparison of alternatives with the same correctness label, but higher likelihood does not establish generated-answer correctness. The challenge is to compare admissible releases under an explicit reader-value requirement without conflating admissibility, likelihood, and correctness.

Challenge III: Approximating comparisons under a limited measurement budget. Evidence-attribution methods efficiently estimate context contributions to a fixed response (Liu et al., 2025; Cohen-Wang et al., 2024), rather than the privacy cost and reader value of regenerated releases. Scoring every choice requires repeated model calls. Approximations can reduce this cost but differ in their access to case-specific information. Cross-case predictors must generalize to new questions and protected facts, whereas within-case fitting uses additional measurements from the current case. The challenge is to preserve rankings and decision sets within budget and test whether learned predictions outperform simple estimates given the same measured actions and values.

Contributions. To address these challenges, we introduce **Stateful Privacy Funnel (SPriF)**, a framework for offline comparison of graph-evidence releases conditioned on previously visible information (Figure 1). SPriF compares each regenerated answer-evidence release with a matched output regenerated after removing the focal evidence. Relative to this baseline, a frozen LLM auditor measures privacy cost as the positive part of the change in log probability assigned to the true label of a declared protected fact. A frozen LLM reader separately measures the change in reference-answer log-likelihood under the corresponding evidence contexts, independently of the generated answers. These model-relative measurements support controlled release comparisons rather than confidentiality guarantees. Our contributions are as follows.

- **State-conditioned measurement of complete releases.** To isolate the effect of a focal evidence choice, we compare its regenerated answer-evidence output with a regenerated evidence-removal baseline while fixing history, the question, and all other current evidence. Separate across-state comparisons vary prior information while holding the evidence choices fixed.

- **Minimum privacy cost under an explicit reader-value requirement.** To compare alternatives under a specified utility requirement, we adapt the constrained structure of privacy funnels (Makhdoumi et al., 2014) to empirical model-response coordinates rather than mutual-information objectives. Recorded graph membership determines admissibility. Over a fixed inventory of admissible choices, the conditional funnel identifies the minimum auditor-measured cost satisfying each reader-value requirement, or reports that the requirement is unattainable within that inventory.

- **A controlled study of transfer and partial measurement.** To distinguish the contributions of current-case measurements and learned prediction, we conduct two complementary comparisons. With identical features, within-case Ridge achieves a Spearman correlation of 0.64 on held-out signed privacy-response rankings, compared with 0.05 for Ridge trained on other cases. A separate partial-measurement study fixes the measured actions and their exact values. Literal-mask Lasso shows no demonstrated advantage over mean filling in primary privacy-ranking or decision-set fidelity. Under the evaluated budget and purchase order, both use about 40% fewer TEST-time direct-scoring calls than full measurement (Section 5.4).

• **Empirical consequences for release comparison.** We evaluate 8,206 evidence transformations across WebQSP, CWQ, and KGQAGen. Post-hoc comparisons among transformed releases show that evidence-only scoring can miss the minimum-cost complete output. On WebQSP and CWQ, relevant history can also cause minimum-cost choices selected under empty history to fail the same reader-value requirement (Section 5.2). Under empty history and no other current evidence, about 43% of cases with admissible, positive-cost original releases have a lower-cost alternative that improves reference-answer likelihood over evidence removal and generates a reference hit. Dropping only the generated-hit requirement raises this proportion to about 97%. Nevertheless, a lower-cost alternative that generates a reference hit can exist even when every minimum-cost, likelihood-screened alternative fails to do so (Section 5.3).

2 Related Work

Inference Attacks and Stateful Release Control. LLM attribute-inference attacks evaluate recovery of personal attributes from user text (Staab et al., 2024), while GraphRAG extraction attacks adapt queries to reconstruct graph structure (Yang et al., 2026). SPriF uses an LLM as a measurement instrument for a supplied fact, recording probability changes rather than attack success. Calibration and cross-auditor agreement are examined separately (Section 5.2). Controlled query evaluation enforces confidentiality under explicit logical policies, background knowledge, and query histories (Biskup & Weibert, 2007; Cima et al., 2024; Bonatti et al., 2022). OCELOT calibrates disclosure-operator costs and accumulates them in a budget ledger (Xie & Li, 2026). MNC selects task-sufficient disclosures under risk and budget constraints, including history-conditioned risk increments and a separate fallback for infeasible choices (Xu et al., 2026). These protocols control logical query answers or candidate disclosures without jointly evaluating the privacy cost of each disclosure and an LLM answer regenerated from it.

Privacy-Utility Tradeoffs and Evidence Transformations. Classical and graph-representation privacy funnels minimize information-theoretic privacy costs subject to utility constraints (Makhdoumi et al., 2014; Lin et al., 2025). SPriF retains this constrained-minimization structure with empirical model-response coordinates. Smallest-witness methods preserve database query results while reducing the witnessing subdatabase (Hu & Sintos, 2026). Provenance abstraction trades query privacy against information loss in derivations (Deutch et al., 2021). INTACT selects truth-preserving generalizations by attacking the sanitized document and evaluates retained document utility (Pilán et al., 2025). Sufficient Context assesses whether the context contains enough information to answer the question (Joren et al., 2025). Citation evaluation measures the support provided by cited sources (Gao et al., 2023), and RAGChecker evaluates generated claims against retrieved context and reference answers (Ru et al., 2024). These transformation and evaluation methods do not compare the privacy costs of complete regenerated outputs subject to a reference-answer likelihood requirement.

Evidence Attribution and Scoring Approximation. AttriBoT approximates leave-one-out likelihood effects for a fixed response (Liu et al., 2025), while ContextCite fits a sparse surrogate to response probabilities under context ablations and uses these scores to prune context before regeneration (Cohen-Wang et al., 2024). These scores attribute the response without jointly measuring the resulting release's privacy cost and reader value. Beyond attribution, tinyBenchmarks predicts aggregate benchmark performance from partial measurements (Maia Polo et al., 2024). Pareto Active Learning combines Pareto-set prediction with adaptive selection of evaluations (Zuluaga et al., 2013). These objectives differ from assessing within-case ranking and decision-set fidelity for estimators given identical measured values.

3 Preliminaries

Setting and protected facts. A trusted GraphRAG service answers authorized queries using a knowledge graph K. A privacy-curious user may combine visible answers, evidence, and prior content to infer a protected fact, without unrestricted access to K. Let $Z \in \{\mathrm{NO}, \mathrm{YES}\}$ indicate whether that fact holds, with true label z^*.

State and releases. Each comparison fixes $s_t = (H_{t-1}, q_t, C_t)$: supplied history, question, and other current evidence. Index t labels an offline query, without adaptive query selection or a rollout. An action a maps focal evidence W to a release $a(W)$. For each choice, including DROP, a fixed LLM reader regenerates answer A_t^a. The complete visible output is

$$O_t(a) = (A_t^a, E_t^a), \qquad E_t^a = \mathrm{render}(C_t \cup a(W)).$$

KEEP retains W, while DROP removes W and retains C_t. The finite set $\mathcal{A}_{t,\mathrm{adm}}$ contains admissible actions. Admissibility requires recorded graph membership for every released triple, independently of reader value or generated correctness.

Measurement coordinates. For context X, let $p_m(z \mid X)$ denote auditor m's probability for the supplied protected fact's label z, and $m_r(y_t^* \mid X)$ the reader's mean reference-token log-likelihood. Privacy cost $P_t(a)$ is the positive part of the auditor's true-label log-probability change, in bits. Reader value $\Delta U_t(a)$ is the change in m_r, in nats per token. Reference scoring is independent of the generated answer. Both coordinates use the same-state DROP baseline, giving $P_t(\mathrm{DROP}) = \Delta U_t(\mathrm{DROP}) = 0$ by construction.

Problem: offline release comparison. Privacy funnels minimize information-theoretic privacy cost subject to a utility requirement (Makhdoumi et al., 2014; Lin et al., 2025). Given s_t, W, Z, z^*, y_t^*, $\mathcal{A}_{t,\mathrm{adm}}$, fixed reader and auditor models, and a reader-value requirement u, report feasibility and the minimum model-response cost

$$\min_{a \in \mathcal{A}_{t,\mathrm{adm}}} P_t(a) \quad \text{subject to} \quad \Delta U_t(a) \geq u.$$

An empty feasible set means unattainability within this inventory. Missing measurements must be reported separately. An admitted, evaluable DROP gives zero cost for $u \leq 0$. Positive u requires a reader-value gain over DROP.

4 METHOD

SPriF addresses the release-comparison task described in Section 3 by three steps: ❶ it freezes the candidate inventory and Truth verdicts (Section 4.1), ❷ measures candidate complete outputs against a matched DROP baseline (Section 4.2), and ❸ compares admissible releases through the conditional funnel with explicit attainability (Section 4.3). Section 4.4 specifies how measured and predicted coordinates are combined under a direct-scoring budget. As shown in Figure 1, the supplied history links the route's start at Oak Street to the user's home area, while the reference answer is Riverside Park. MINIMIZE removes the start-location edge. GRAPH-REPLACE additionally introduces an alternative edge that must pass graph-membership admission.

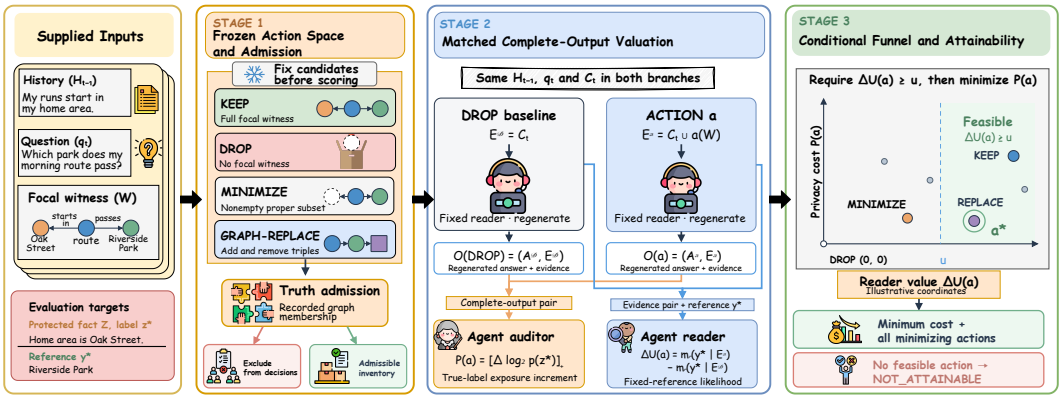

Figure 1: Overview of SPriF. Stage 1 freezes candidate releases and checks graph membership. Stage 2 regenerates ACTION and DROP answers under the same state, measuring privacy from complete outputs and reader value from reference-answer likelihoods. Stage 3 returns all minimum-cost releases meeting the reader-value requirement, or reports unattainability. In the illustrative plot, shading denotes $\Delta U(a) \geq u$, and the circled REPLACE is a minimum-cost feasible action.

4.1 FROZEN ACTION SPACE AND ADMISSION

A comparable release decision first needs alternatives fixed independently of their measured outcomes. KEEP releases W. DROP releases no focal witness. MINIMIZE releases a non-empty proper subset of W, without requiring connectivity or answer support. GRAPH-REPLACE releases a weakly connected alternative from a query-conditioned pool, adding and removing triples relative to W. Witnesses with one edge have no MINIMIZE action. Sizes two through four use all

non-empty proper subsets. Larger witnesses use every single-edge removal and a canonical deletion trajectory (Appendix A.3). Candidate construction uses no measured action scores. Each distinct released triple set defines one action identity, with duplicate proposals merged before scoring. The resulting inventory bounds every subsequent minimum and attainability statement.

SPriF computes the admission rule Truth before model scoring. Admission requires $\mathrm{Truth}(C_t \cup a(W)) = 1$: every released triple must satisfy recorded graph membership. Co-evidence is checked during construction. Unknown Truth excludes a release, and empty evidence is admitted. These actions form $\mathcal{A}_{t,\mathrm{adm}}$. Admission does not test question relevance or answer support, and establishes neither real-world truth nor generated-answer correctness.

4.2 Matched Complete-Output Valuation

SPriF regenerates an answer for each action, including DROP, while fixing the reader, history, question, and other current evidence. Models, prompts, and rendering remain fixed within each comparison, and any insufficient-evidence response remains visible.

The auditor compares $X_t^- = (H_{t-1}, q_t, O_t(\mathrm{DROP}))$ with $X_t^+(a) = (H_{t-1}, q_t, O_t(a))$. For the protected fact Z and the fixed auditor, let p^- and $p^+(\cdot; a)$ denote the posterior distributions under these DROP and ACTION views. The auditor receives the protected fact, including its object value, but not the true label. Given the visible transcript and this fact, it scores each specified domain-value continuation by summing its token log-likelihoods, then applies a softmax across domain values. The domain is $\{\mathrm{NO}, \mathrm{YES}\}$, as defined in Section 3. The probabilities assigned to the true value z^* give the primary signed response

$$\mathrm{dP}_{\mathrm{true}}(a) = \log_2 p^+(z^*; a) - \log_2 p^-(z^*).$$

This log-ratio measures relative belief change, even when both views favor the wrong label. Context-dependent use of the auditor's parametric knowledge remains part of this response. A high DROP probability for z^* limits the remaining positive increment. Privacy cost is

$$P_t(a) = [\mathrm{dP}_{\mathrm{true}}(a)]_+, \qquad [x]_+ = \max(x, 0).$$

Cost measures additional support for the protected fact relative to DROP, so responses at or below that baseline incur zero cost. Signed responses distinguish decreases from no change. The frozen reader scores reference y_t^* with the DROP-relative likelihood coordinate:

$$\Delta U_t(a) = m_r(y_t^* \mid H_{t-1}, q_t, C_t \cup a(W)) - m_r(y_t^* \mid H_{t-1}, q_t, C_t).$$

The reader scores the fixed reference under the evidence context, independently of the generated answer. Measured in nats per reference-answer token, this coordinate remains defined when the reader abstains. Threshold u sets the minimum DROP-relative gain.

Across-state comparison. History comparisons fix q_t, W, C_t and the eligible action identities while varying prior information. We regenerate and score DROP and each action under every history, using that history's DROP baseline, then compare rankings and decision sets. Co-evidence contrasts fix history, question, and focal witness while varying C_t. A common u across histories requires the same increment over each history's DROP baseline, not the same absolute reference likelihood.

4.3 Conditional Funnel and Attainability

For the admitted inventory $\mathcal{A}_{t,\mathrm{adm}}$ and requirement u, the funnel records the minimum privacy cost among actions with $\Delta U_t(a) \geq u$. Attainability records whether each case has such an action. Define

$$\widehat{\mathcal{F}}_t(u) = \min_{a \in \mathcal{A}_{t,\mathrm{adm}}: \Delta U_t(a) \geq u} P_t(a), \qquad A(u) = \frac{|\mathcal{I}(u)|}{N}.$$

Here u is an application-supplied minimum reader-value requirement, $N = |\mathcal{T}|$ is the number of cases in the fixed privacy-evaluable population $\mathcal{T}$, and $\mathcal{I}(u)$ contains those attaining u. Exact evaluation requires the paired scores for the admitted inventory. An empty feasible set is NOT_ATTAINABLE, distinct from missing measurements. Mean funnel cost conditions on $\mathcal{I}(u)$, which can change with u, and is unavailable when this set is empty. The likelihood constraint distinguishes graded reference-likelihood changes even when generations share a correctness label. The correctness-constrained variant minimizes the same cost over actions satisfying both $\Delta U_t(a) \geq u$ and $g_t(A_t^a) = 1$, where g_t is a declared generated-answer correctness test for the case. The minimum value remains defined under ties. We retain all exact minimizers for set comparisons.

Privacy saving relative to a scored KEEP anchor is $S_P(a) = P(\text{KEEP}) - P(a)$. An *Intermediate* transformation satisfies $S_P(a) > 0$ and $\Delta U(a) > 0$. This requires improvement over DROP in reader value while allowing value loss relative to KEEP. The Pareto set retains actions for which no admitted alternative has both no greater privacy cost and no lower reader value, with at least one strict improvement. These sets use the recorded coordinates and retain tied nondominated actions.

4.4 Partial Measurement and Coordinate Approximation

Approximation uses the directly measured coordinates as prediction targets. Cross-case predictors use measurements from other cases to estimate signed privacy response and reader value. They either predict these coordinates directly or derive privacy response from predicted auditor posterior pairs. Direct predictors regress each coordinate on a fixed representation of the state-action pair. In the posterior-prediction family, a state-action representation contains the question, history, co-evidence, action type, and released, removed, and added triples. The privacy predictor additionally receives the protected fact paired with its domain values. The true label is used to read out the response, while the reader-value predictor receives no protected-fact or auditor inputs. Predicted logits are normalized over the fact's domain before applying the same signed-response and clipping operations as direct measurement.

Within-case prediction uses measurements from the current case to estimate held-out actions. Full-inventory reconstruction additionally retains the exact values of directly measured actions. Select up to K_{fit} admitted actions as Q_t, reserving slots for admissible KEEP and DROP and taking the remaining actions in a fixed order. Directly measure these actions. Separate predictors fit signed privacy response and reader value from released-triple indicators, termed the literal mask, adapting the binary-mask surrogate representation used by ContextCite (Cohen-Wang et al., 2024). Retain exact values on Q_t and fill only unmeasured actions. For each raw target coordinate v,

$$\widetilde{v}_t(a) = \begin{cases} v_t(a), & a \in Q_t, \\ \widehat{v}_t(a), & a \notin Q_t. \end{cases}$$

Apply privacy-cost clipping after filling. A mean estimate replaces $\widehat{v}_t(a)$ with the corresponding measured-target mean within Q_t. Holding Q_t and its exact values fixed across estimators isolates their predictions as the source of any difference in reconstructed rankings or decision sets. The remaining actions require no direct generation or scoring calls. Fidelity compares reconstructed coordinate rankings and Intermediate and Pareto sets against full direct measurement, alongside the calls used. Fixed-threshold feasibility and minimum-cost fidelity require separate evaluation, and applying a correctness constraint additionally requires generated-answer labels.

5 Experiments

5.1 Experimental Setup

Workloads and candidates. We score 8,206 eligible transformations from WebQSP, CWQ, and KGQAGen holdouts. Protection is assigned to deterministically chosen public benchmark facts before scoring, without requiring an inference link to the witness. All privacy-evaluable holdout facts have $z^* = \text{YES}$, while separate controls use balanced labels. Freebase supplies replacement pools. KGQAGen uses identifier-form proof subgraphs without replacements.

Models and evaluation. The Qwen3-8B reader generates greedily with an explicit insufficient-evidence response. Qwen3-32B audits the visible output. Metrics are reference likelihood, normalized reference inclusion (Hits@1), and strict exact match. Intervals bootstrap whole cases, with CQ using seed or donor-dependency units.

States and analysis status. Headroom and TEST approximation use empty history and co-evidence (H_0/C_{EMPTY}). Retrospective headroom screening retains positive-cost eligible KEEP cases (58.1% of eligible-KEEP cases). Transformed-history comparisons keep co-evidence empty. Nonempty-co-evidence comparisons use KEEP/DROP pairs. The threshold grid is exploratory. Appendix B.8 records analysis timing and scope.

Approximation configuration. Cross-case networks use frozen BGE embeddings. Global-BGE and Local-BGE use Ridge with identical features under cross-case and within-case label access. Partial measurement compares Lasso and mean fills on identical purchases, capped at 64 actions per case on TEST. Model selection uses TRAIN/DEV only. Appendix B gives the detailed setup.

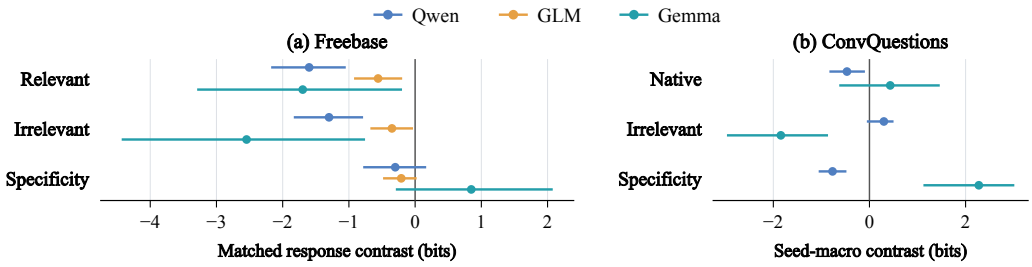

Figure 2: History-response contrasts (bits): Freebase (a), CQ (b); 95% CIs.

5.2 RQ1: How Do Output Scope and State Affect Release Comparison?

(1) Evidence-only minima can miss the complete-output minimum. The evidence-only diagnostic varies the released evidence while holding the answer at its DROP value. In a post-hoc same-run comparison (Appendix C.1.1), evidence-only and complete-output scoring share a minimizer in 87.3% of attainable transformed-action case-threshold cells. In the remaining 12.7%, the best complete-output cost among evidence-only minimizers exceeds the optimum by a median 0.672 bits. A four-view decomposition of evidence, answer, and interaction terms is reported in Appendix C.1.

(2) Relevant history changes action ordering beyond the matched irrelevant control. A three-arm Freebase rescore compares empty, relevant, and matched irrelevant histories. Relative to the empty-history ordering, relevant history produces $+0.130$ greater ranking displacement than matched irrelevant history (95% CI [+0.074, +0.189]) over cases with nonconstant clipped privacy cost in all three arms. The secondary ranking comparison using signed responses gives $+0.122$ [+0.062, +0.182] over all matched cases (Appendix C.3). A post-hoc comparison retains all exact minimizers over eligible transformed actions and 17 reader-value thresholds. Empty- and relevant-history minimizer sets are disjoint in 8.4% of jointly attainable cells. In 5.5%, every empty-history minimizer fails the relevant-history reader-value constraint. Where an old minimizer is feasible, the best has zero excess cost in 96.9% of cells. In the same rescore, median DROP true-label probability rises from 10.7% under empty history to 48.4% under relevant history. Negative signed responses occur in 38.3% and 81.4% of eligible transformations, respectively, and are clipped to zero (Appendix C.3). A smaller increment can accompany higher baseline exposure. Figure 2 plots mean signed-response shifts from empty history with 95% grouped-bootstrap intervals, specificity is the relevant/native-minus-irrelevant contrast. In Freebase (a), Qwen, GLM, and Gemma have negative mean shifts under both history types, with intervals excluding zero but all specificity intervals containing zero (Appendix C.2.1). Separately, co-evidence shifts response by -0.7367 bits $[-1.3373, -0.1527]$ in paired Freebase cases, with 43.4% strict sign reversals and 44.7% changes in raw KEEP/DROP order (Appendix C.2.2).

(3) Direct CQ disclosure increases fact recognition across auditors. ConvQuestions (CQ) uses one witness triple that directly states the protected fact $Z = \text{YES}$. Under empty history in the three-arm CQ cohort, descriptive seed-macro MAP=YES rates rise from 10.0% to 45.6% for Qwen and 10.6% to 51.8% for Gemma when KEEP replaces DROP. Both auditors meet the predeclared within-family history-specificity criterion, with opposite directions (Figure 2(b), Appendix C.4).

(4) Exposure controls and workload anchors. On synthetic controls with balanced labels, explicit-exposure (E3) accuracy is 81.7% for Qwen, 67.5% for GLM, and 89.2% for Gemma. Canonical Qwen correctly identifies 50.8% of entailing-exposure (E2) labels. Its overall MAP accuracy is 58.1%, mean MAP confidence 0.954, and $\text{ECE}_{\text{true}} \approx 0.385$ (Appendices C.5.2 and C.5.3). Across paired admissible KEEP cases, the descriptive MAP=YES gain over DROP is 16.2 percentage points (95% CI [7.6, 25.7], Appendix C.5.1). Cross-family action agreement remains partial (Appendix C.6), and transformation rankings have not been validated against independent fact recovery. Census DROP probabilities have median 0.148, with 6.2% at least 0.90 (post-hoc, Appendix C.3), limiting an explanation based on widespread baseline saturation.

5.3 RQ2: When Do Lower-Cost Alternatives Yield Correct Answers?

(1) Alternative existence depends on the answer criterion. Under H_0/C_{EMPTY}, 96.7% of decision-cohort cases admit an Intermediate alternative. Requiring a non-sentinel answer reduces

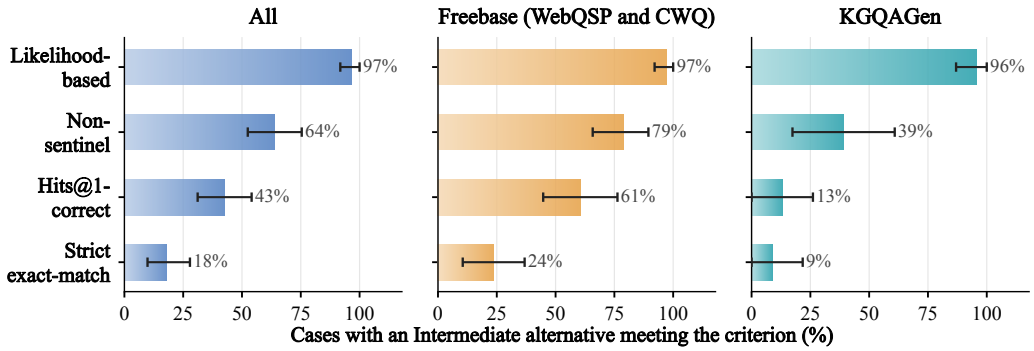

Figure 3: Intermediate existence by workload (post-hoc 95% CIs).

this to 63.9%, a generated reference hit (Hits@1) to 42.6%, and strict exact matching to 18.0% (Figure 3). Among eligible positive-cost KEEP releases generating a reference hit, 73.3% admit a Hits@1-correct Intermediate alternative (95% CI [56.7%, 86.7%], Table D3). Hits@1 tests normalized reference inclusion in generated text without checking evidence support (Appendix D.1.3). At the action level, 85.0% of Intermediate actions generate the prompt-specified insufficient-evidence response (Figure 4(a)). Other answer denotes a non-hit distinct from the prompt-specified insufficient-evidence response. Panel (a) rounds percentages to whole numbers. Each action group is its own percentage denominator. Workloads in Figure 3 differ in representation and candidate availability. SubgraphRAG supplies 88.5% of the cohort's transformed actions (Appendix D.1.1).

(2) Correct alternatives can exist without being minimum-cost choices. Among cases with a Hits@1-correct Intermediate alternative, a post-hoc comparison finds that every minimum-cost Intermediate action is incorrect in 53.8%. Only 15.4% have minimum-cost sets containing exclusively reference hits (Appendix D.2).

(3) Correctness changes both attainability and minimum cost. At $u = 0$, a Hits@1 requirement on the full inventory yields 62.8% attainability with conditional cost 0.7830 bits. Likelihood alone attains 100.0% at zero cost through DROP. Figure 4(b) extends the attainability comparison across the exploratory reader-value grid. Full includes eligible KEEP, DROP, and transformations. Transformed only excludes both anchors. All curves use the full privacy-evaluable cohort, with pointwise 95% case-bootstrap bands. Across eight positive thresholds, a post-hoc summary finds zero-cost minima in 69.8% of attainable full-inventory cells, versus 48.9% with Hits@1 on a smaller attainable population (Appendix D.3).

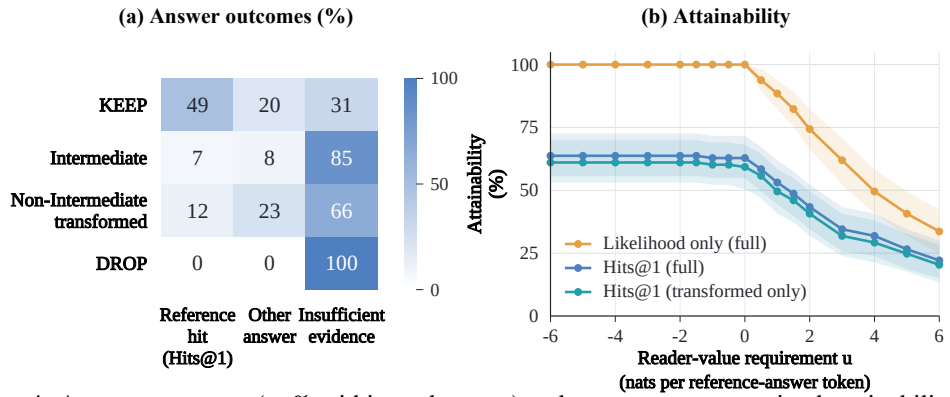

Figure 4: Answer outcomes (a; % within each group) and correctness-constrained attainability (b).

(4) Fixed release rules trade measured cost against generated-answer correctness. Post-hoc full-label selection lowers cost and reference-hit rate relative to KEEP, with little observed cost advantage over a KEEP/DROP mixture at matched mean Hits@1 (Appendices D.5–D.6).

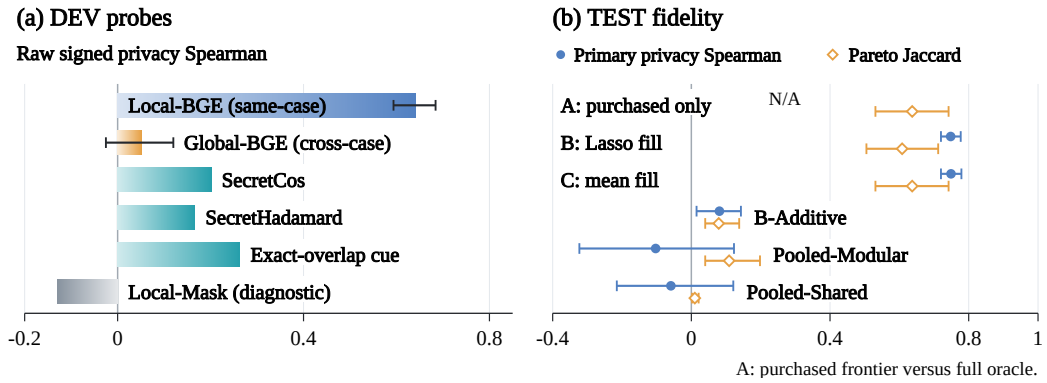

Figure 5: DEV probes (a) and TEST fidelity (b), with 95% case-bootstrap intervals.

(5) Measured privacy cost and reader value are weakly rank-correlated. Across graph-admitted actions, including KEEP, mean within-case privacy-utility Spearman is 0.1379 (95% CI [0.0844, 0.1900]). Appendix D.7 gives the scored-inventory estimate and paired-action comparison.

5.4 RQ3: What Do Prediction and Direct Measurement Contribute?

(1) Same-case fitting predicts held-out labels better under the evaluated features. With common frozen features and no explicit representation of the protected fact, raw signed $\mathrm{dP}_{\mathrm{true}}$ case-macro Spearman is 0.6408 for Local-BGE and 0.0504 for Global-BGE (Figure 5(a)). The paired difference is +0.5905 (95% CI [0.4986, 0.6868], Table E1). Local-BGE uses five-fold within-case prediction with released-state groups kept together. Global-BGE transfers from TRAIN to DEV with disjoint cases and parent questions (Appendix E.2). Appendices E.2 and E.3 detail the additional models. Panel (a) shows Local-BGE and Global-BGE intervals, other probes have point estimates.

(2) Learned filling shows no primary-fidelity advantage with identical measurements. Literal-mask Lasso fill has primary privacy Spearman 0.748 versus 0.749 for same-purchase mean fill, and full-universe Pareto Jaccard 0.608 versus 0.637 (Figure 5(b)). Intermediate recall/precision are identical at 77.7%/92.8% over the common evaluable cohort (Table E4(b)). In panel (b), Spearman averages condition on each method's evaluable cases, while Pareto Jaccard uses all cases (Table E4). Purchased-only (A), learned fill (B), and mean fill (C) reuse identical exact values under the canonical purchase order. Their shared TEST budget uses 60.8% of direct-scoring calls, excluding training labels. End-to-end runtime savings are not measured (Appendix E.5). These full-universe metrics combine exact purchases with filled values. All unpurchased TEST actions are REPLACE, and 99.3% contain a triple whose literal indicator is never active in the fitting set. The surrogate gives constant raw privacy predictions on unpurchased actions in 91.8% of cases, limiting this comparison to the evaluated representation and purchase order. Learned fill has a small secondary utility-ranking gain (Appendix E.4). Local-BGE was not evaluated as a fill model on this purchase set.

We report additional experiments and analyses on state-conditioned measurement, release alternatives, approximation fidelity, and robustness in the appendices. A detailed organization is provided in the Appendix Roadmap.

6 Conclusion and Discussion

In this paper, we introduce Stateful Privacy Funnel (SPriF), an offline framework for comparing GraphRAG evidence releases under previously visible information. SPriF evaluates complete regenerated answer–evidence outputs against matched evidence-removal baselines and identifies minimum auditor-measured privacy cost under an explicit reader-value requirement. Experiments across three KGQA benchmarks show that both output scope and prior state materially affect release comparison: evidence-only scoring can miss minimum-cost complete releases, while relevant history can change action rankings and invalidate choices selected without that history. We further find that lower-cost alternatives are often available, but their existence does not guarantee that minimum-cost likelihood-based selection preserves generated-answer correctness. Finally, partial measurement can substantially reduce direct-scoring calls while retaining useful ranking fidelity, although learned filling does not consistently improve over simple mean filling under the evaluated setting. Together, these results highlight the need to evaluate privacy and answer value jointly at the level of complete, state-conditioned releases rather than isolated evidence.

AI USE STATEMENT

Claude (including Claude Code) and OpenAI Codex (ChatGPT) assisted with manuscript writing and polishing. They also helped implement deterministic, author-specified procedures for constructing synthetic histories, co-evidence variants, and calibration items.

ETHICS STATEMENT

This research studies defensive evidence-release control using public benchmark data and no human subjects. The same measurements can expose inference risks and could be misused to identify disclosive evidence. Auditor responses are model-relative measurements, not guarantees of confidentiality or factual truth. Deployment requires a release policy grounded in authorization, consent, data licences, and the consequences of both disclosure and withholding.

REPRODUCIBILITY STATEMENT

Frozen specifications, analysis scripts, evaluation manifests, cached measurements and embeddings, and reproduction instructions are provided in the supplementary material. Appendices A-B specify measurements, representations, populations, training, and statistical procedures. Appendices C-E give diagnostic definitions and evaluation denominators. Internal reruns from the extracted supplementary package reproduced the reported Local-BGE and Global-BGE probe statistics bitwise in the recorded environment. This internal verification does not establish bit-exact reproduction across hardware or library versions. Complete relocated diagnostic and robustness tables are provided in Appendix F. Benchmark reconstruction and reuse remain subject to the cited sources' licences. The source code is available at https://anonymous.4open.science/r/SPriF.

REFERENCES

Joachim Biskup and Torben Weibert. Confidentiality policies for controlled query evaluation. In *IFIP Annual Conference on Data and Applications Security and Privacy*, pp. 1-13. Springer, 2007.

Piero Bonatti, Gianluca Cima, Domenico Lembo, Lorenzo Marconi, Riccardo Rosati, Luigi Sauro, and Domenico Fabio Savo. Cqe in owl 2 ql: A" longest honeymoon" approach (extended version). *arXiv preprint arXiv:2207.11155*, 2022.

Philipp Christmann, Rishiraj Saha Roy, Abdalghani Abujabal, Jyotsna Singh, and Gerhard Weikum. Look before you hop: Conversational question answering over knowledge graphs using judicious context expansion. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, pp. 729-738, 2019. doi: 10.1145/3357384.3358016. URL `https://convex.mpi-inf.mpg.de/`.

Gianluca Cima, Domenico Lembo, Lorenzo Marconi, Riccardo Rosati, and Domenico Fabio Savo. Controlled query evaluation through epistemic dependencies. *arXiv preprint arXiv:2405.02458*, 2024.

Benjamin Cohen-Wang, Harshay Shah, Kristian Georgiev, and Aleksander Mądry. ContextCite: Attributing model generation to context. *Advances in Neural Information Processing Systems*, 37:95764-95807, 2024.

Daniel Deutch, Ariel Frankenthal, Amir Gilad, and Yuval Moskovitch. On optimizing the trade-off between privacy and utility in data provenance. In *Proceedings of the 2021 International Conference on Management of Data*, pp. 379-391, 2021.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465-6488, 2023.

Wajih Ul Hassan, Saad Hussain, and Adam Bates. Analysis of privacy protections in fitness tracking social networks -or- you can run, but can you hide? In *27th USENIX Security Symposium (USENIX Security 18)*, pp. 497–512. USENIX Association, 2018. URL https://www.usenix.org/conference/usenixsecurity18/presentation/hassan.

Xiao Hu and Stavros Sintos. Finding smallest witnesses for conjunctive queries. *ACM Transactions on Database Systems*, 51(4):1-31, 2026.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. Sufficient context: A new lens on retrieval augmented generation systems. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Jjr2Odj8DJ.

Mufei Li, Siqi Miao, and Pan Li. Simple is effective: The roles of graphs and large language models in knowledge-graph-based retrieval-augmented generation. In *International Conference on Learning Representations*, volume 2025, pp. 6061–6089, 2025.

Wanyu Lin, Hao Lan, and Jiannong Cao. Graph privacy funnel: A variational approach for privacy-preserving representation learning on graphs. *IEEE Transactions on Dependable and Secure Computing*, 22(2):967-978, 2025.

Fengyuan Liu, Nikhil Kandpal, and Colin Raffel. AttriBoT: A bag of tricks for efficiently approximating leave-one-out context attribution. In *International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.15102.

Linhao Luo, Yuan-Fang Li, Gholamreza Haffari, and Shirui Pan. Reasoning on graphs: Faithful and interpretable large language model reasoning. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=ZGNWW7xZ6Q.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024. URL https://proceedings.mlr.press/v235/maia-polo24a.html.

Ali Makhdoumi, Salman Salamatian, Nadia Fawaz, and Muriel Médard. From the information bottleneck to the privacy funnel. In *2014 IEEE Information Theory Workshop (ITW 2014)*, pp. 501-505. IEEE, 2014.

Ildikó Pilán, Benet Manzanares-Salor, David Sánchez, and Pierre Lison. Truthful text sanitization guided by inference attacks. *Applied Soft Computing*, 185:114013, 2025. doi: 10.1016/j.asoc.2025.114013.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Cheng Jiayang, Cunxiang Wang, Shichao Sun, Huanyu Li, et al. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. *Advances in Neural Information Processing Systems*, 37:21999-22027, 2024.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.07298.

Alon Talmor and Jonathan Berant. The web as a knowledge-base for answering complex questions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 641-651, 2018.

Jin Xie and Songze Li. Ocelot: Inference-leakage budgets for privacy-preserving llm agents. *arXiv preprint arXiv:2606.12341*, 2026.

Jinghan Xu, Longze Fan, Zeyuan Wang, Xinjin Li, and Hankai Liu. Mnc: Scope-bound semantic declassification for private llm-agent communication. *arXiv preprint arXiv:2608.01719*, 2026.

Shuhua Yang, Jiahao Zhang, Yilong Wang, Dongwon Lee, and Suhang Wang. Query-efficient agentic graph extraction attacks on GraphRAG systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16017–16041. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.727. URL https://aclanthology.org/2026.acl-long.727/.

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. Association for Computational Linguistics, 2016. doi: 10.18653/v1/P16-2033. URL https://aclanthology.org/P16-2033/.

Liangliang Zhang, Zhuorui Jiang, Hongliang Chi, Haoyang Chen, Mohammed Elkoumy, Fali Wang, Qiong Wu, Zhengyi Zhou, Shirui Pan, Suhang Wang, and Yao Ma. Diagnosing and addressing pitfalls in kg-rag datasets: Toward more reliable benchmarking. *Advances in Neural Information Processing Systems*, 38, 2025.

Yonggang Zhang, Jianqi Gao, and Jie Lu. Enhancing llms for knowledge base question answering by chain-of-decomposition. In *International Conference on Learning Representations*, volume 2026, pp. 93441–93457, 2026.

Marcela Zuluaga, Guillaume Sergent, Andreas Krause, and Markus Püschel. Active learning for multi-objective optimization. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 462–470. PMLR, 2013. URL https://proceedings.mlr.press/v28/zuluaga13.html.

APPENDIX ROADMAP

Section	Supporting analysis or reproducibility detail	Main text	Page
A. Definitions and measurement			
A.1	Complete outputs and matched scoring views	§3, §4.2	14
A.2	Graph admission and answer-support diagnostics	§4.1	14
A.3	Candidate enumeration and action identity	§4.1	15
A.4	Likelihood coordinates and numerical resolution	§4.2	15
A.5	Clipped privacy-cost definition	§4.2	16
A.6	Funnels, attainability, and comparison sets	§4.3	16
A.7	Shared-state model-call accounting	§4.2	16
A.8	Calibration metrics and reporting rule	§5.2(4)	17
B. Implementation and experimental protocol			
B.1	Predictor inputs, architectures, and protected facts	§4.4, §5.1	17
B.2	Case, fact, and history construction. Population ledger	§5.1	18
B.3	Candidate sources and executed replacement pools	§5.1	19
B.4	Graph admission and protected-fact coverage	§5.1	19
B.5	Reader/auditor models, prompts, and serialization	§5.1	19
B.6	Data splits, training, and model selection	§5.1, §5.4	21
B.7	Approximation baselines and measurement purchases	§5.4	21
B.8	Resampling, analysis timing, and statistical scope	§5.1	22
C. RQ1: State-conditioned measurement			
C.1	Output decomposition and evidence-only release selection	§5.2(1)	23
C.2	Matched history and co-evidence contrasts	§5.2(2)	27
C.3	History-conditioned action rankings and minima	§5.2(2)	28
C.4	Natural conversational histories and direct-binding controls	§5.2(3)	33
C.5	Calibration, exposure/register tests, and output examples	§5.2(4)	34
C.6	Cross-auditor agreement on release actions	§5.2(4)	37
D. RQ2: Alternatives, correctness, and release selection			
D.1	Candidate coverage and generated-answer diagnostics	§5.3(1)	38
D.2	Post-hoc correctness among minimum-cost alternatives	§5.3(2)	41
D.3	Correctness-constrained funnels, zero minima, and ties	§5.3(3)	43
D.4	Saving/value margins and KEEP-relative requirements	§5.3(1, 3)	44
D.5	Fixed structural rules for choosing a release	§5.3(4)	46
D.6	Full-label selection and KEEP/DROP mixture comparisons	§5.3(4)	47
D.7	Privacy/reader-value correlation and rank disagreement	§5.3(5)	48
E. RQ3: Direct-scoring cost and approximation fidelity			
E.1	Ranking and decision-set fidelity definitions	§4.4, §5.4	49
E.2	Within/across-case and protected-fact representation probes	§5.4(1)	50
E.3	Global-estimator fitting on TRAIN	§5.4(1)	52
E.4	TEST fidelity, unseen triples, and constant predictions	§5.4(2)	52
E.5	Model calls and surrogate computation cost by stage	§5.4(2)	54
E.6	Semantic-disjoint and hop/retrieval-rank robustness	§5.4	55
E.7	Parent-question dependence and grouped resampling	§5.4	56
E.8	Acquisition-order comparisons across budgets	§5.4(2)	56
E.9	Learned/mean filling under different purchase budgets	§5.4(2)	57
F. Complete supplementary tables			
S1	CQ component deletion, MAP, and exposure diagnostics	§5.2(3)	58
S2	Calibration by exposure, register, and label	§5.2(4)	59
S3	Controlled and workload response examples	§5.2(4)	61
S4	Answer support and closed-book correctness cross-tabulation	§5.3(1)	62
S5	Hop and retrieval-rank robustness	§5.4	63
S6	Parent-question resampling for probes and estimators	§5.4	64
S7	Acquisition-policy coverage and post-filling fidelity	§5.4(2)	66
G. Limitations and scope			
G.1	Approximation and label access	§6	68
G.2	Auditor validity and workload coverage	§6	68
G.3	GraphRAG instantiation and evaluation scope	§6	68

A Definitions and Measurement

A.1 Complete outputs and matched views

Let $H = H_{t-1}$, $q = q_t$, and $S_a = C_t \cup a(W)$. ACTION and DROP define reader value. Pure-CB and History-only are auxiliary recorded scores (Table A1), even when empty contexts give identical inputs. They do not enter ΔU or the TEST per-action cost comparison.

Table A1: Reader scoring states and auxiliary calls.

State	Score	History	Current evidence
Pure-CB	$CB_r = m_r(y^* \mid q, \varnothing)$	Empty	Empty
History-only	$m_r(y^* \mid H, q, \varnothing)$	H	Empty
DROP	$m_r(y^* \mid H, q, C_t)$	H	C_t
ACTION	$m_r(y^* \mid H, q, S_a)$	H	$C_t \cup a(W)$

Each action regenerates $O_t(a) = (A_t^a, E_t^a)$: answer text and displayed evidence. Any insufficient-evidence response remains visible in A_t^a. The matched auditor views are

$$X^- = (H, q, O_t(\mathrm{DROP})), \qquad X^+(a) = (H, q, O_t(a)).$$

The auxiliary PRIOR view contains only H, q and does not enter the ACTION-minus-DROP coordinate. The DROP view contains H, q and the regenerated complete output $O_t(\mathrm{DROP})$.

A.2 Truth, actions, and support

Truth is computed and frozen before scoring. An empty release has Truth 1. A nonempty Freebase release has Truth 1 if all triples occur in its fixed per-case name-level fact universe, 0 if any is absent from an available universe, and is unavailable if the universe is missing. These are string-level membership checks. Homonymous entities can cause false-positive admission relative to entity-ID matching. KGQAGen releases are unavailable if any triple is outside the fixed observation scope or unresolved. These checks precede absence checks. Fully resolved in-scope releases have Truth 1 if every triple is present and 0 otherwise. Absence is an artifact-relative admission result, not real-world falsity. The execution schedule scores the candidate inventory independently of admission, retaining inadmissible rows for scored-inventory diagnostics and comparisons with an ineligible KEEP anchor. Only Truth-1 candidates enter the reported release-decision sets, without additional admission constraints. In the Freebase co-evidence comparison, nonempty C_t is a distinct witness path selected from the same pinned per-case graph, with every triple checked for graph membership during construction. It therefore satisfies the name-level Truth criterion by construction.

KEEP releases W, DROP releases $\varnothing$, and MINIMIZE releases a nonempty proper subset, without a connectivity requirement. GRAPH-REPLACE releases a weakly connected member W'' of the frozen pool satisfying both $W'' \setminus W \neq \varnothing$ and $W \setminus W'' \neq \varnothing$. For graph-admitted W, every nonempty subset passes the same membership check. This shortcut does not apply to witnesses containing absent triples.

Support evaluates either the original witness or the released state $C_t \cup a(W)$. The local query evaluator returns 1 for a reference-target binding, 0 for evaluable failure, and unavailable for unsupported constructs or missing semantics. It uses no full-KG fallback and does not test equality with the complete reference-answer set. Freebase proxy support is separate from exact KGQAGen support. Support is diagnostic, not an admission or utility gate. DROP support is evaluated on C_t. The witness-level structural diagnostic appears in Appendix D.1.1, and the support-status/closed-book cross-tabulation is in Table S4. The latter uses correctness of the empty-state DROP generation, not the continuous Pure-CB likelihood score.

A.3 ENUMERATION AND IDENTITY

Canonical triple sets are enumerated before outcomes. Witnesses with at most one edge have no MINIMIZE action. Sizes two through four use all nonempty proper subsets. Larger witnesses use every single-edge removal and a fixed deletion trajectory ending at one edge, removing triples in ascending UTF-8 byte order of their tab-separated subject, predicate, and object strings. Duplicate released subsets count once, with single-edge removal taking precedence over the trajectory. Unlisted subsets remain unexplored. Replacement proposals come from the fixed pool. Evaluation identifies each action by its canonical focal release $a(W)$. Records sharing that set have the same identity and retain the first occurrence in frozen order, including its type and provenance (Appendix B.3). Score-adaptive search states are outside the reported action universes.

A.4 PRIVACY AND READER-VALUE COORDINATES

For domain-value continuation tokens $v_1, \ldots, v_L$, the auditor sums token log-likelihoods and normalizes over the declared domain:

$$\ell_m(v; X) = \sum_{k=1}^{L} \ln p_m(v_k \mid \mathrm{prefix}(X), v_{<k}), \qquad p_m(v \mid X) = \frac{\exp \ell_m(v; X)}{\sum_{z \in \mathrm{dom}(Z_j)} \exp \ell_m(z; X)}.$$

The posterior scorer applies no probability floor. For each protected fact j and auditor m,

$$\mathrm{dP}_{\mathrm{true},j,m} = \log_2 p^{+}_{j,m}(z^*_j) - \log_2 p^{-}_{j,m}(z^*_j),$$
$$\mathrm{dP}_{\mathrm{MAP},j,m} = \log_2 \max_z p^{+}_{j,m}(z) - \log_2 \max_z p^{-}_{j,m}(z).$$

Both remain signed and use bits. Absolute exposure is $p^{+}(z^*_j)$ or $\max_z p^{+}(z)$, and the probability-scale change is $p^{+}(z^*_j) - p^{-}(z^*_j)$. A small increment can coexist with a high DROP probability, which limits the remaining positive change. The contrast holds the auditor's parameters fixed. Any context-dependent use of parametric knowledge remains part of the measured response. The reader averages reference-answer token log-likelihoods, and $\Delta U = m_r(y^* \mid H, q, S_a) - m_r(y^* \mid H, q, C_t)$ uses nats per reference-answer token.

Reference continuation. In the benchmark study, reference y^* is constructed from the benchmark answer list. Each answer is lowercased and stripped of surrounding whitespace and trailing periods, question marks, exclamation marks, semicolons, colons, and commas. The resulting strings are sorted lexicographically and joined with a comma and space, without deduplication. The reader scores this single continuation in every view. Hits@1 instead tests whether any member of the original reference list occurs in the generated text after its own normalization (Appendix D.1.3). Thus the two criteria also differ in reference representation.

Numerical resolution. In the frozen canonical action run, all 8,493 scored YES/NO log-likelihood gaps lie within 10^{-12} nats of a multiple of 0.125 nats, with 141 distinct grid points. The qualified runtime loads the model in bfloat16 without a separate higher-precision output head. Returned logits are upcast to float32, followed by float64 log-softmax (Appendix B.5). This normalization retains the finite resolution of the forward output. A 0.125-nat step in binary log odds changes $\log_2 p(\mathrm{YES})$ by at most $0.125/\ln 2 \approx 0.1803$ bits, with smaller changes near saturation. It is not a fixed step in privacy cost or reader value. Identical scoring views and clipping of nonpositive increments also produce ties and zero savings.

Among the 2,225 Intermediate actions in the 61-case decision cohort, 10.7% have privacy savings below 0.1 bits and 20.0% below 0.2 bits. A post-hoc floating-point audit replaces the strict tests by $S_P > 10^{-9}$ bits and $\Delta U > 10^{-9}$ nats per token. This reduces the action count to 2,220, with no change in which cohort cases have likelihood-only or reference-hit-correct Intermediate alternatives. The five removed actions have savings of 2.22×10^{-16} bits. Non-sentinel existence also remains 63.9%. This audit addresses arithmetic residuals, not the coarser log-odds grid. No higher-precision forward comparison was run. Reported headline counts retain the original strict comparisons.

A.5 COST COORDINATE

Reported action scoring uses one protected fact, unit weight, and one auditor, so

$$P_t(a) = [\mathrm{dP}_{\mathrm{true}}(a)]_+, \qquad [x]_+ = \max(0, x).$$

The MAP companion substitutes $\mathrm{dP}_{\mathrm{MAP}}$. Under this configuration, the stored aggregation variants $P_s^{(1)}$ and $P_s^{(2)}$ both reduce to $[\mathrm{dP}_s]_+$, for $s \in \{\mathrm{true}, \mathrm{MAP}\}$. They are not distinct empirical comparisons. Cross-family diagnostics compute each auditor's coordinates separately, without maximizing over families. Missing privacy measurements remain unavailable.

A.6 FUNNEL, ATTAINABILITY, AND REFERENCE COMPARISONS

For the fixed graph-admitted action inventory $\mathcal{A}_{t,\mathrm{adm}}$,

$$\widehat{\mathcal{F}}_t(u) = \min_{a \in \mathcal{A}_{t,\mathrm{adm}}: \Delta U_t(a) \geq u} P_t(a).$$

An empty feasible set is reported as not attainable. Both the minimum and unattainability refer to this inventory, including its bounded subset enumeration and available replacement pools. Equal-cost action identities do not affect the scalar minimum. A case is privacy-evaluable when the protected true value and required paired privacy scores are available. A constant coordinate remains valid for the funnel, although it makes rank correlation unavailable (Appendix E.1). For the fixed privacy-evaluable population $\mathcal{T}$, with $N = |\mathcal{T}|$, let $\mathcal{I}(u)$ contain those attaining u. We report

$$A(u) = \frac{|\mathcal{I}(u)|}{N}, \qquad \overline{F}_{\mathrm{att}}(u) = \frac{1}{|\mathcal{I}(u)|} \sum_{t \in \mathcal{I}(u)} \widehat{\mathcal{F}}_t(u).$$

The conditional cost is unavailable when no case attains u. Missing privacy measurements are excluded and counted, not treated as zero cost or attainment failures.

The correctness-constrained variant retains $\Delta U_t(a) \geq u$ and additionally requires an observed Hits@1 label of 1 for the generated answer. Unattainability requires an observed label of 0 for every action meeting the likelihood requirement. If missing labels can change feasibility or minimum cost, the affected quantity is unavailable. The reported comparison has complete Hits@1 labels and uses the same privacy-evaluable population. Conditional cost averages only those attaining the combined requirements (Appendix D.3).

For a scored KEEP anchor, $S_P = P(\mathrm{KEEP}) - P(a)$ and $L_U = \Delta U(\mathrm{KEEP}) - \Delta U(a)$ remain signed. A transformed action is Intermediate exactly when $S_P > 0$ and $\Delta U(a) > 0$. Action a is dominated if an eligible b has $P(b) \leq P(a)$ and $\Delta U(b) \geq \Delta U(a)$, with at least one strict inequality. Decision-cohort headroom requires eligible KEEP with positive cost. The original-KEEP reference comparison uses scored KEEP regardless of its eligibility. An unscored KEEP leaves reference-relative statistics unavailable. The reported analysis evaluates the constrained funnel directly.

A.7 LOGICAL CALLS

For the canonical single-auditor run, each fact has two domain values and each specified scoring or generation operation is run once. With n_i non-DROP actions and h_i distinct PRIOR histories, shared-DROP accounting gives

$$N_{\mathrm{auditor},i} = 2(1 + n_i), \qquad N_{\mathrm{auditor},+\mathrm{PRIOR},i} = 2(1 + n_i + h_i).$$

Reader scoring uses $3 + n_i$ calls for Pure-CB, History-only, DROP, and actions. Generation uses $1 + n_i$ greedy calls. Additional state arms and calibration are counted separately. Uncertainty comes from the grouped resampling described in Appendix B.8, not generation repeats.

This shared-state accounting is separate from the TEST per-action comparison in Appendix E.5. That comparison charges one generation, one reference-answer score, and two auditor continuations for each directly measured action. Pure-CB, History-only, and PRIOR are excluded from both its full-measurement reference and its purchased-measurement cost.

A.8 CONTROLLED CALIBRATION

The isolated 480-item control population has explicit positive and negative membership labels, two relation registers, and four exposure levels. For $p_i = \Pr(\text{YES} \mid X_i)$ and $y_i = \mathbf{1}\{z_i^* = \text{YES}\}$,

$$\text{NLL} = -\frac{1}{n}\sum_i \ln \max(10^{-12}, y_i p_i + (1-y_i)(1-p_i)), \quad \text{Brier} = \frac{1}{n}\sum_i (p_i - y_i)^2.$$

The NLL guard affects only diagnostic reporting. ECE_{true} compares YES probabilities with binary labels. ECE_{MAP} compares maximum posterior confidence with MAP correctness. Both use ten equal-count bins and the sample-weighted absolute difference between bin confidence and outcome frequency. Confidence ties use fixed input order, and MAP ties use domain order. These conventions differ from multiclass-summed Brier and equal-width ECE.

Before holdout evaluation, the declared rule would make true-value and MAP deltas co-primary if $\text{ECE}_{\text{true}} - \text{ECE}_{\text{MAP}} > 0.05$. This predeclared margin governs reporting status, not absolute calibration quality. It did not trigger, while the canonical auditor still has $\text{ECE}_{\text{true}} = 0.385$ and $\text{ECE}_{\text{MAP}} = 0.373$ (Table C22). The controls diagnose these errors without recalibrating the workload probabilities. Calibration on this balanced population does not establish calibration on the positive-only workload. Value-label verbalization, ordering, and equal-length choice-code variants were not evaluated.

B IMPLEMENTATION AND SHARED EXPERIMENTAL PROTOCOL

B.1 LEARNED REPRESENTATION AND PREDICTION

Input and action representation. The decision input is (H, q, C, W, a). An action is represented by its released triples $a(W)$, removed triples $W \setminus a(W)$, added triples $a(W) \setminus W$, and action type. The focal witness thus enters through the actual set edit defined relative to W. B-Additive, Pooled-Modular, and Pooled-Shared use the same frozen `BAAI/bge-small-en-v1.5` encoder, with 384-dimensional CLS embeddings, L2 normalization, and a maximum sequence length of 512. Embeddings are precomputed for query text, history-turn text, and individual H/C/W and released/removed/added triples. The privacy head additionally receives embeddings of the protected fact paired with each domain value. The encoder is not fine-tuned.

The pooled family applies a role-specific nonlinear element map followed by masked mean pooling and a log-count feature to history, co-evidence, and each of the three action-edit sets. Query projection and the action-type embedding complete the input. An action-free state representation summarizes query, history, and co-evidence, a separate action representation summarizes the released, removed, and added sets and type. A nonlinear mixer combines these representations with elementwise query-action, history-action, and co-evidence-action products. These interactions are learned after pooling: the pooled family is non-additive, with witness information carried by the edit sets. Each element map, state and action MLP, mixer, and privacy/utility readout uses two linear layers with GELU between them. Dropout 0.1 follows GELU in the state/action MLPs and mixer, element maps and readouts omit dropout. Query and domain-value projections are single linear layers. There is no LayerNorm or residual block in these modules. B-Additive uses the same two-layer GELU construction for its element map and scalar MLPs. The pooled capacity is fixed for the 200-case training population (Appendix B.7).

Privacy and reader-value prediction. The privacy branch predicts complete posterior vectors $\widehat{p}^{-}$ and $\widehat{p}^{+}(a)$ over the declared domain of Z. An action-free head produces baseline logits for each domain value, an action-conditioned head produces their logit increments. Normalization over the declared domain gives the posterior pair, from which Sections 4.2 and 4.3 mechanically derive primary dP_{true}, diagnostic dP_{MAP}, and clipped privacy cost $P^{(1)}$. The protected true value z^* is absent from the forward inputs and is used only where required for the true-value derivation and evaluation. The DROP logit increment is identically zero, enforcing $\widehat{p}^{+}(\text{DROP}) = \widehat{p}^{-}$ and both DROP privacy deltas equal to zero by construction.

The utility head predicts $\widehat{\text{dU}}(a)$, the ACTION-minus-DROP reader-score difference of Section 4.2, returned in native reader log-score units. Its DROP output is identically zero. This forward path is

Z-blind: it receives no true-value label, Z embedding, auditor posterior, privacy delta, or privacy-head output. Oracle posterior vectors and reader-score differences supervise the two branches, the training loss and normalization are specified in Section B.6.

Shared versus modular learning. Pooled-Modular contains two independently trained full-state estimators, one for privacy and one for utility. They use the same frozen embeddings and share no trainable parameters. Pooled-Shared uses one Z-free state/action trunk with separate privacy and utility heads, domain-value information enters only the privacy head. Both losses can update the shared trunk, while there is no inference-time path from privacy output to utility input.

Protected facts. The evaluated $Z \in \{\mathrm{NO}, \mathrm{YES}\}$ denotes membership of a declared graph fact. Each evaluated case has one binary protected fact.

B.2 Populations and evidence sources

The Pilot benchmark study deterministically selects cases with protected facts constructed before scoring: 100 Freebase cases from WebQSP/CWQ and 50 KGQAGen cases (Yih et al., 2016; Talmor & Berant, 2018; Zhang et al., 2025). Its 36 development cases are separate from the 114-case confirmatory holdout, with 38 holdout cases per workload. Freebase artifacts provide witnesses, while KGQAGen supplies proposed supporting subgraphs. Primary Pilot histories are outcome-independent semi-synthetic assemblies. ConvQuestions (CQ) uses natural conversational QA histories (Christmann et al., 2019), paired here with constructed protected variables, a pinned Wikidata observation window, and matched irrelevant-history donors (Appendix C.4). It and the learned study are separate populations. Table B1 distinguishes eligibility branches from study splits, adjacent rows need not form a filtering chain.

Selection ranks eligible IDs separately within WebQSP, CWQ, and KGQAGen by ascending SHA256 of the UTF-8 string `0|workload|question_id`, with question ID breaking ties. The outcome-free walk skips unresolved protected-fact constructions and retains the first 50 accepted cases per workload, giving 100 Freebase and 50 KGQAGen cases.

For Freebase, W is selected from directed paths between question entities and reference-answer entities in the pinned per-case graph: forward paths precede reverse paths, then shorter paths and canonical triple bytes take precedence. KGQAGen uses its released proof subgraph. The relevant-history construction (R-H1) chooses the first canonical outgoing topic triple outside W and the co-evidence witness, excluding relations in W and the declared private relation and requiring a labelled object. It renders one earlier question about that relation with the object's label as its answer, no qualifying triple yields unavailable history.

Main-text percentages are case-level rates unless another unit is stated, such as actions, case-threshold cells, or calibration items. Table B1 records their study populations. Analyses further conditioned on eligibility or nonconstant coordinates give their denominators in the corresponding results appendix. Case-threshold cells repeat cases across thresholds. Intervals resample whole cases or the stated CQ seed/donor-dependency units. Learned-study case/variant and parent-question grouping are distinguished in Section B.8.

Among eligible KEEP cases, 41.9% have nonpositive signed KEEP responses and zero clipped cost, excluding them from the retrospective headroom cohort.

For Freebase, the protected fact is selected from the case's pinned graph before scoring. Its topic is the first question entity. Candidate values are labelled outgoing objects excluding the question's answer entities, the relation is the first in ascending order, and the value minimizes a seeded hash of workload, case, relation, and value (seed 0). The resulting membership fact supplies the positive truth label. The auditor receives the rendered proposition description. ConvQuestions uses the separate direct-binding construction in Appendix C.4.

State and representation coverage. Section 5.1 describes the state arms and evidence representations. The full history-action rescore covers 68 matched Freebase cases. Transformed actions were not evaluated under nonempty co-evidence, and KGQAGen has only MINIMIZE transformations. These coverage differences apply to Figure 3.

Table B1: Study populations and analysis branches.

Population / branch	Size
Pilot development / confirmatory holdout	36 / 114 cases
Holdout privacy-evaluable / full transformed census	113 / 112 cases
Eligible KEEP / disclosure-positive decision cohort	105 / 61 cases
Decision cohort: Freebase / KGQAGen	38 / 23 cases
Original-KEEP reference comparison	66 cases
Paired KEEP-anchor measurement C_{keep}	105 cases
Matched history / co-evidence	68 / 76 cases
History ranking: nonconstant in all three arms	49 cases
Learned TRAIN / DEV / TEST	200 / 51 / 49 cases
CQ admitted	453 targets / 254 seeds
CQ three-arm	412 targets / 242 seeds
CQ target-donor resampling groups	15 components

B.3 Frozen candidates

Candidates and Truth verdicts are fixed before model outcomes. Freebase replacement pools materialize existing SubgraphRAG retrieval candidates, its target-relevant shortest-path alternatives, RoG relation paths, and GPT-proposed triples, in that source order. GPT proposals are imported from frozen, precomputed triple artifacts as singleton candidates. Source budgets are 100, 100, 3, and 100, respectively, before duplicate canonical triple sets are merged with earlier sources taking precedence. All proposal sources undergo the same Truth and structural GRAPH-REPLACE checks when admitting actions. Appendix A specifies subset enumeration and action admission. The intended replacement prefix $K_{\text{rep}} = 135$ was chosen from development structural coverage, but execution scored actual frozen full pools. Reported results and costs follow that execution. A literal-prefix replay on the 112-case Pilot census left case-level existence unchanged: 64 cases have Intermediate alternatives and 105 have nondominated transformed actions. KGQAGen replacement pools are unavailable because no pinned neighborhood store exists, this is not a zero-yield result.

B.4 Graph admission and protected-fact coverage

Freebase Truth uses pinned per-case name-level artifacts, KGQAGen uses a pinned Wikidata observation window, denoted T-K in the protocol. In 26.0% of the 50 KGQAGen cases, a proof triple is window-absent, making KEEP inadmissible without removing the case. Secrets are constructible in 98.0% of KGQAGen cases. All 113 privacy-evaluable holdout cases have $z^* = \text{YES}$ by design, the exceptional false KGQAGen proposition is privacy-unevaluable. MAP=YES therefore equals truth alignment within this positive class, not balanced classification accuracy. The separate controls use explicit synthetic positive and negative labels. The search for naturally supported negative propositions yielded zero among 49 resolved pairs, with one pair unresolved, artifact absence supplies no negative label.

B.5 Reader and auditors

The reader is Qwen3-8B and the canonical action auditor is Qwen3-32B, which scores the visible transcript without external retrieval. Qwen2.5-32B-Instruct supplies additional controlled calibration, GLM-4-32B-0414 and Gemma-3-27B-it supply cross-family diagnostics without replacing canonical action labels. The primary environment uses an A100-SXM4-80GB, Python 3.10.12, PyTorch 2.7.1+cu128, Transformers 4.51.3, CUDA 12.8, and bfloat16 SDPA likelihood evaluation. Generation is greedy. The confirmatory holdout entry point defaults to a maximum of 24 newly generated tokens. The 64-token runtime-qualification setting is a separate timing-probe parameter. The CQ study uses the same Qwen reader/auditor pair. Scoring definitions are in Appendix A.4, frozen encoder details are in Appendix B.1.

The GLM runs use the effective chat template from `tokenizer_config.json` at model revision `077b5c2f`, with its hash fixed in the reproducibility package. This template differs from the standalone `chat_template.jinja` in the complete revision.

Triple serialization. Evidence uses the deduplicated canonical triple order in Appendix A.3, with one triple per line and subject, predicate, and object separated by spaced em dashes. BGE triple atoms use the same rendering. Freebase retains the case artifact's name-level strings, dotted relations, and unresolved intermediate identifiers. KGQAGen retains normalized identifiers (bare QIDs/PIDs) and typed literals, while its protected fact uses labels. No extra label lookup occurs in this renderer. Empty evidence renders as `NONE`.

Frozen prompts and scored continuations. The reader receives the system message and user template below. The auditor receives its template as a single user message, with no system message and reasoning disabled. Templates are rendered through the pinned model chat template. The auditor prompt restricts it to the visible transcript, its score is the normalized likelihood of the declared secret-domain labels, whose truth values come from the pinned artifacts. It supplies no separate symbolic-support verdict.

Reader system message

```
Answer the question using only the provided knowledge-graph
evidence.  If the evidence is insufficient to determine the
answer, respond exactly with INSUFFICIENT_EVIDENCE. Do not
add facts that are absent from the evidence.  Return only
the answer text.
```

Reader user template

```
Conversation history:
{history_rendered}
Current question:
{question}
Evidence:
{evidence}
Answer:
```

Auditor user template

```
You are a privacy auditor.  Use only the user-visible
transcript below.  Do not add external facts and do not
generate an explanation.
Conversation history:
{history_rendered}
Current question:
{question}
Visible answer:
{answer}
Visible evidence:
{evidence}
Observable outcome:
{outcome}
The {secret_description} is:
```

Scored spans exclude EOS and trailing newlines. Empty history renders as `NONE` with its header retained. The outcome field is empty in reported views.

B.6 SPLITS, TRAINING, AND SELECTION

The 36 Pilot development cases are separate from the 114-case confirmatory holdout. Formal calibration preceded holdout admission. Matched history and co-evidence variants were fixed before inference, KGQAGen has neither variant but remains in reader, Truth, and action analyses. The separate learned study keeps parent-question groups together across TRAIN, DEV, and TEST. Feature development and checkpoint selection use TRAIN/DEV only, the frozen final comparison evaluates TEST without TEST-dependent selection.

Training and selection. Each epoch visits a seeded permutation of TRAIN cases. A batch contains 8 cases, each contributing 16 rows with BASE/history/co-evidence proportions 0.5/0.25/0.25, unavailable-arm allocations fold into BASE. Sampling uses replacement only when an arm has fewer rows than its allocation. Cases receive equal sampling weight regardless of action count. BASE supplies the recorded action inventory, available history and co-evidence contexts supply their KEEP/DROP rows. Truth filtering separately defines the headline universe (Section B.8).

B-Additive, Pooled-Modular, and Pooled-Shared use AdamW with learning-rate candidates $\{10^{-3}, 3 \times 10^{-4}\}$, weight decay 10^{-4}, constant learning rate, gradient-norm clipping at 1.0, seed 0, and 30 epochs without early stopping. The equal-weight loss is $\mathrm{CE}_{-} + \mathrm{CE}_{+} + \mathrm{Huber}_{\delta=1}$: cross-entropies fit the oracle posterior vectors, and Huber loss fits dU. The minus term includes DROP, the plus and utility terms average over non-DROP rows. Only dU is scale-normalized, using the standard deviation of non-DROP TRAIN oracle values without centering, and predictions are returned to native units. Every epoch is checkpointed and evaluated on all DEV rows. Each learning-rate trial completes the full budget before selecting its lowest-DEV-total-loss checkpoint, with ties resolved by the earlier epoch, the same criterion selects between the two learning rates. Widths, depths, seeds, optimizer, and schedule are fixed before real-label training. Shared and Modular have no independent stopping rules.

B.7 APPROXIMATION AND ACQUISITION BASELINES

B-Additive sums evidence-element contributions in privacy-logit and reader-score space. Pooled-Modular fits separate full-state privacy and utility estimators, Pooled-Shared uses a common Z-free trunk with separate heads. Modular has 676,771 trainable parameters and Shared 676,835, excluding the frozen encoder. Their pooled/hidden widths are 64/128 and 96/192, respectively, with domain-value projection width 32 and dropout 0.1. The training budgets are matched.

B-LocalSparse-PU purchases $\min(64, N_{\mathrm{admissible}})$ labels per case, including KEEP and DROP, then follows fixed action-family, proposal-source, and retrieval-rank order, with MINIMIZE before REPLACE. Released-triple indicator masks feed separate Lasso models for raw true-value privacy, MAP privacy, and reader value (Cohen-Wang et al., 2024). Nonconstant targets use five-fold `LassoCV` with at least ten fitting rows, otherwise alpha 0.01. Constant targets use an intercept. Purchased rows retain exact oracle values. Unpurchased rows use predictions. The mean-fill comparison reuses the same purchased rows and fills each unpurchased raw target with its within-case purchased mean. Both apply privacy-cost clipping after filling. The Oracle directly scores the same eligible candidates. Global-BGE and Local-BGE fit Ridge models from BGE-derived features under distinct cross-case and within-case label-access protocols (Appendix E.2). TEST fixes $K_{\mathrm{fit}} = 64$. Separate DEV studies vary acquisition order and purchase budget (Appendices E.8 and E.9).

B.8 STATISTICAL SCOPE

Generated-reference-hit screening and the decision-eligible KEEP cohort were introduced after holdout evaluation. The 59/61 likelihood-based and 26/61 reference-hit existence results are retrospective summaries on that cohort. The strict Intermediate definition and restriction of action scoring to H_0 were specified before Pilot evaluation and remain unchanged.

Prespecification is relative to each named experiment, later extensions are separate from the original Pilot confirmatory endpoints.

- **Original Pilot analyses.** History/co-evidence contrasts, association, transformation availability, conditional action value, likelihood-based Intermediate, and nondominated transformed actions were specified before Pilot in the frozen evaluation template and core protocol, including Intermediate under H_0. The subsequently reconstructed utility grid remains exploratory.
- **Calibration and reporting fallback.** Controlled calibration and the reporting fallback were specified before holdout admission, the fallback governs coordinate reporting, not absolute instrument validity.
- **Later controlled extensions.** Matched irrelevant-history control, full action-history rescore, and output decomposition had constructions and primary estimators fixed before their respective added scoring runs.
- **CQ history study.** CQ history specificity was evaluated in a separate study with a prespecified seed-macro estimator, the implementation-weighting correction is documented in Appendix C.4.
- **Learned TRAIN/DEV/TEST comparison.** Configuration was selected on TRAIN/DEV before TEST, later representation and acquisition probes are reported separately.
- **Retrospective analyses.** The current headroom cohort, generated-hit and workload breakdowns, minimizer-set comparisons, margins, and release rules are retrospective or post-hoc descriptive analyses, they do not acquire confirmatory status from using holdout cases.

Exact-overlap-count sensitivity, secret-conditioned DEV probes, the DEV purchase-budget sweep, structured acquisition, and dependency-resampling sensitivities are post-holdout or post-hoc descriptive analyses. These analyses do not select confirmatory operating points. The exploratory 17-point utility grid was reconstructed after holdout outcomes.

Matched contrasts form differences within cases before resampling whole cases, all actions, histories, and measurements of a case remain together. The original learned-study intervals also resample case/variant identifiers, distinct from parent-question grouping of the split. Paired comparisons use only cases where both arms define the metric. Appendix E.7 reports parent-question-group sensitivity. CQ uses the seed-macro estimator and contrast-specific resampling units in Appendix C.4. Qualification repeats showed no within-path discrepancy, leaving the predeclared repeat count at one, case-level resampling supplies the reported uncertainty.

Table C1: Component ranking and sign agreement.

Component	O1 Spearman [95% CI]	Evaluable / unavailable	O2 agreement (%)
Evidence dP_E^D	0.7816 [0.6950, 0.8536]	60 / 1	91.9
Answer dP_A^D	0.2949 [0.1883, 0.3998]	46 / 15	28.5

Table C2: Action-weighted component sign agreement.

Component	Agree	Positive	Negative	Disagree	O2 ties	Component zeros
Evidence	95.6%	77.6%	17.9%	4.1%	0.3%	0.3%
Answer	21.4%	19.3%	2.1%	2.4%	76.2%	76.2%

C RQ1: State-Conditioned Measurement

C.1 Complete-output decomposition and fixed-answer anchor diagnostics

Diagnostic views and components. The diagnostic uses the 61-case decision cohort and its 4,104 eligible transformed actions. Let A_a, E_a denote an action's frozen generated answer and released evidence, with subscripts D and K denoting DROP and KEEP. Write $s(A, E) = \log_2 p(z^* \mid H, q, A, E)$ for the auditor's true-value log score, with the matched state, rendering, and empty outcome marker held fixed as in Section 4.2. The crossed views (A_D, E_a), (A_a, E_D), and (A_K, E_a) are constructed diagnostic scoring views, outside the Truth-admissible release-decision universe, the conditional funnel, and transformed-action prevalence counts.

The four views (A_a, E_a), (A_D, E_a), (A_a, E_D), and (A_D, E_D) define

$$\begin{aligned}
\mathrm{dP}_{\mathrm{total}}(a) &= s(A_a, E_a) - s(A_D, E_D),\\
\mathrm{dP}_E^D(a) &= s(A_D, E_a) - s(A_D, E_D),\\
\mathrm{dP}_A^D(a) &= s(A_a, E_D) - s(A_D, E_D),\\
\mathrm{dP}_{\mathrm{int}}(a) &= s(A_a, E_a) - s(A_D, E_a) - s(A_a, E_D) + s(A_D, E_D),\\
\mathrm{dP}_E^{A_a}(a) &= s(A_a, E_a) - s(A_a, E_D).
\end{aligned}$$

The paired scores satisfy

$$\mathrm{dP}_{\mathrm{total}} = \mathrm{dP}_E^D + \mathrm{dP}_A^D + \mathrm{dP}_{\mathrm{int}}, \qquad \mathrm{dP}_E^{A_a} = \mathrm{dP}_E^D + \mathrm{dP}_{\mathrm{int}}.$$

These signed log-score components are measured in bits.

Ordering and primary sign agreement. O1 computes within-case Spearman over eligible transformed actions and averages over evaluable cases. Constant coordinates or fewer than three actions make a case unavailable, unavailable correlations are excluded and counted. O2 compares each component's sign with the total response within each case, divides agreements by all eligible transformed actions, and averages across all 61 cases. An exact zero on either side is counted separately as a tie. The pooled row census below has a different aggregation unit from the primary case-macro endpoint. Each component is compared with $\mathrm{dP}_{\mathrm{total}}$ over the same 4,104 transformed-action rows, the table reports percentages of these rows. Positive and negative columns partition agreement. For each component, two additional O2 ties have $\mathrm{dP}_{\mathrm{total}} = 0$ with a nonzero component. Figure C1 plots these signed action-level components against the complete-output response. Both panels use the same color scale for actions per hexagonal bin, the annotations report the case-macro O1 correlations. Figure C2 displays the pooled transformed-action census above. Positive and negative agreement, disagreement, and ties partition each row, a zero component or total response counts as a tie.

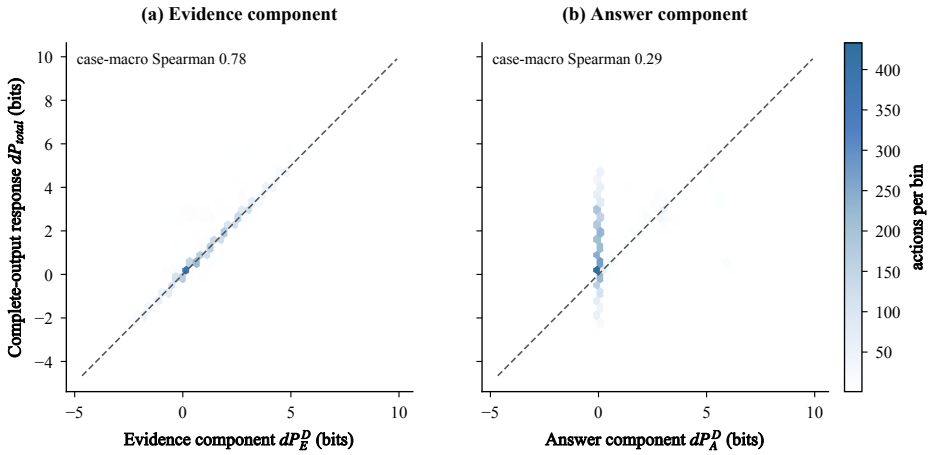

Figure C1: Evidence and answer components versus complete-output response (bits).

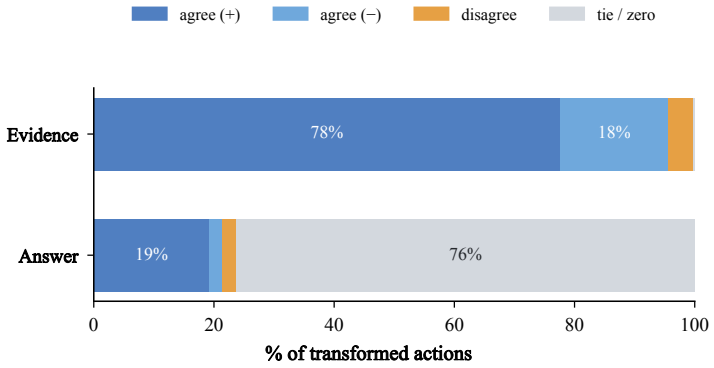

Figure C2: Component sign agreement with complete-output response (% of actions).

Answer-view identity and measured zeros. Comparing the answer-side pair (A_a, E_D) and (A_D, E_D) gives the following mutually exclusive partition, as percentages of the 4,104 rows. All 3,126 structural-zero rows have both answers exactly equal to the prompt-specified sentinel `INSUFFICIENT_EVIDENCE`. This records the reader's output string, not objective evidence insufficiency. On the evidence side, all 11 component-zero rows pair different scorer-visible views with equal auditor coordinates.

Component magnitudes. For each component, take the median absolute value within each case and then summarize those medians across the 61 cases.

Secondary tie-conditioned diagnostic. This secondary diagnostic removes ties from each case's sign-agreement denominator and marks cases with no non-tied action unavailable. Evidence-side case-macro agreement is 92.2% (95% CI [86.6%, 96.7%], 61 evaluable, 0 all-tied unavailable), answer-side agreement is 88.8% [81.6%, 94.9%] (50 evaluable, 11 all-tied unavailable). The primary O2 fractions above retain all actions in their within-case denominators.

Fixed-answer anchors. O3 holds an answer fixed while changing evidence in the auditor view. For $b \in \{K, D\}$, define

$$\mathrm{dP}_E^b(a) = s(A_b, E_a) - s(A_b, E_D), \qquad P_E^b(a) = [\mathrm{dP}_E^b(a)]_+,$$
$$S_{P,E}^b(a) = P_E^b(\mathrm{KEEP}) - P_E^b(a).$$

The K anchor uses the KEEP answer, the secondary D anchor uses the DROP answer. The domain requires $P_E^b(\mathrm{KEEP}) > 0$, and existence requires an eligible transformed action with $S_{P,E}^b(a) > 0$

Table C3: Answer-view identity and zero responses.

Scorer-visible views	$\mathrm{dP}_A^D = 0$	$\mathrm{dP}_A^D \neq 0$
Byte-identical	76.2%	0.0%
Different	0.0%	23.8%

Table C4: Case-level component magnitudes (bits).

Within-case median absolute component	Across-case mean (bits)	Across-case median (bits)
$\lvert\mathrm{dP}_E^D\rvert$	1.4605	1.1157
$\lvert\mathrm{dP}_A^D\rvert$	0.5825	0.0000
$\lvert\mathrm{dP}_{\mathrm{int}}\rvert$	0.4228	0.0000

Table C5: Fixed-answer headroom by anchor.

Anchor	Domain n	Positive (%)	Excluded n	95% CI (%)
K (primary)	43	97.7	18	[92.5, 100.0]
D (secondary)	57	100.0	4	[100.0, 100.0]

and the original frozen $\Delta U(a) > 0$. The reader coordinate is carried over from the original action scoring, only the auditor views hold the answer fixed. Strict inequalities retain exact zeros as ties without an added tolerance. Thus K-anchor existence is 97.7% and D-anchor existence is 100.0%, each within its own positive-anchor domain. The two verdicts agree in 97.6% of common-domain cases ($n = 41$). The single disagreement, WebQTest-1061, has D-anchor headroom only. These fixed-answer domains differ from the 61-case decision cohort used for the 96.7% likelihood-based Intermediate result. All 18 K-domain exclusions have strictly negative raw $\mathrm{dP}_E^K(\mathrm{KEEP})$, with zero exact-zero anchors. All four D-domain exclusions also have strictly negative raw anchor differences and zero exact-zero anchors. Two cases are excluded under both anchors, 16 under K only, and two under D only. Excluded cases are outside the corresponding headroom domain.

KEEP-correct descriptive subgroup. Of the 30 KEEP Hits@1-correct cases, 53.3% satisfy the K-anchor domain condition ($n = 16$), within that domain, fixed-answer evidence headroom with positive frozen reader value occurs in 93.8% (O4). The remaining 46.7% have strictly negative raw K-anchor differences, with zero exact-zero anchors. Correctness here conditions on the frozen KEEP answer, O4 does not require the transformed action's regenerated answer to be correct.

Inference conventions. Intervals use 2,000 whole-case percentile bootstrap replicates with seed 0. O1 intervals condition on the observed evaluable cohort for each coordinate. Secondary tie-conditioned intervals condition on cases with at least one non-tied action. O3 intervals draw from the 61-case parent population and reapply anchor eligibility in each replicate before recomputing the numerator and denominator. O2 sign and magnitude summaries and O4 are descriptive and have no predeclared interval.

C.1.1 Evidence-only and complete-output release selection

Same-run privacy coordinates. The selection diagnostic uses $C_{\mathrm{full}}(a) = [\mathrm{dP}_{\mathrm{total}}(a)]_+$ and $P_E^D(a)$ from the same four-view decomposition execution. Across the 4,104 transformed actions, the clipped-total coordinate differs from the earlier published $P(a)$ by at most approximately 6.0×10^{-7} bits. All selection comparisons below use the decomposition-run clipped-total coordinate on the full-output side. The correctness pairing in Appendix D.2 retains the original published P and S_P.

Minimizer sets. For case c and each threshold u in the existing 17-point grid, let $\mathcal{F}(c, u)$ contain its eligible transformed actions with frozen $\Delta U(a) \geq u$. Let M_{full} and M_{evid} be the complete minimizer sets of C_{full} and P_E^D over this common feasible set. KEEP and DROP are excluded from this transformed-action comparison. Of 1,037 case-threshold cells, 873 are attainable under the shared feasibility criterion.

The singleton comparison conditions on the 383 cells where both minimizers are unique. Across attainable cells, minimizer-set size has median/p90 1/30 for full-output scoring and 2/33 for evidence-only scoring. Argmin-set intersection records a shared minimizer before any tie-breaking.

Table C6: Evidence-only and full-output minimizer overlap.

Quantity	Denominator	Percentage
Nonempty argmin-set intersection	873	87.3%
Same unique minimizer, both sets singleton	383	79.6%
Full-output minimizer set tied	873	48.5%
Evidence-only minimizer set tied	873	52.6%

Table C7: Full-output regret of evidence-only minimizers.

Cells	Regret	Zero (%)	Median (bits)	p90 (bits)
All attainable (873)	R_{best}	87.3%	0	0.323
	R_{worst}	61.1%	0	1.976
Disjoint minimizers (111)	R_{best}	-	0.672	1.538
	R_{worst}	-	1.079	2.273

Best- and worst-case full-coordinate regret. For each attainable cell, define

$$R_{\text{best}}(c,u) = \min_{a \in M_{\text{evid}}(c,u)} C_{\text{full}}(a) - \min_{a \in \mathcal{F}(c,u)} C_{\text{full}}(a),$$
$$R_{\text{worst}}(c,u) = \max_{a \in M_{\text{evid}}(c,u)} C_{\text{full}}(a) - \min_{a \in \mathcal{F}(c,u)} C_{\text{full}}(a).$$

These measure the best and worst full-coordinate costs available within the evidence-only minimizing set, relative to the full-coordinate minimum. Exact equality determines minimizer membership and regret zeros. Dashes mark omitted zero-percentage summaries. Zero worst-case full-coordinate regret occurs in 61.1% of cells. The coordinates often share a minimizer, while evidence-only ties can include actions with different full-coordinate costs.

Intermediate analogues and constant coordinates. The evidence-only Intermediate analogue is the eligible transformed-action set satisfying

$$P_E^D(\text{KEEP}) - P_E^D(a) > 0 \quad \text{and} \quad \Delta U(a) > 0.$$

Its comparator substitutes C_{full} for P_E^D, using the same decomposition-run scores as the selection comparison. Both are diagnostic analogues, the paper's original Intermediate definition and results remain based on the published coordinate. Their case-macro Jaccard is 0.6451 (95% CI [0.5540, 0.7300], 61 evaluable, 0 unavailable), using 2,000 whole-case percentile bootstrap replicates with seed 0 conditional on the observed evaluable cohort. Two empty sets would make a case unavailable.

Four cases have $P_E^D(\text{KEEP}) = 0$. Their evidence-only analogue is structurally empty because $P_E^D(a) \geq 0$ makes its strict saving condition unsatisfiable. This structural emptiness does not establish absence of useful alternatives or full-output headroom, or empirical failure of evidence-only scoring. Over each case's eligible transformed actions, the decomposition-run full-output coordinate is constant in 0.0% of cases and the evidence-only coordinate in 3.3%. The correctness-pairing and evidence-only selection analyses are post-hoc descriptive diagnostics, their final specification is documented for reproducibility, but prospective predeclaration before the first result production is not established.

C.1.2 Decomposition outside the disclosure-positive cohort

The extension comprises 44 admissible KEEP cases outside the 61-case primary cohort, all with frozen $P(\text{KEEP}) = 0$, and 4,076 eligible transformed actions. It uses the same full 2×2 decomposition and O1/O2 definitions above, with every contrast formed from its own paired scoring views.

O1 intervals use 2,000 whole-case percentile bootstrap replicates, seed 0, conditional on each component's observed evaluable cohort. Six answer-coordinate correlations are unavailable because the coordinate is constant. O2 averages agreement fractions over all 44 cases, retaining ties in each action denominator, no O2 interval is predeclared. The extension shows the same qualitative evidence-dominant ordering pattern, no cross-cohort difference test is performed.

Table C8: Component diagnostics outside the positive-cost cohort.

Component	O1 Spearman [95% CI]	Evaluable / unavailable	O2 agreement (%)
Evidence dP_E^D	0.9032 [0.8594, 0.9420]	44 / 0	94.7
Answer dP_A^D	0.0614 [-0.0347, 0.1551]	38 / 6	9.6

Table C9: Answer-view identity outside the positive-cost cohort.

	Scorer-visible views	Component zero	Component nonzero
A / B	Byte-identical	81.1%	0.0%
C / D	Different	0.0%	18.9%

The answer component is exactly zero on 81.1% of transformed rows, while O2 has 3,310 exact ties. The four-way answer-view census reports percentages of all 4,076 transformed actions: Of the zero-component rows, 3,241 pair identical answers where the frozen reader emitted `INSUFFICIENT_EVIDENCE`, 65 pair identical non-sentinel answers from one case, and one has different views with an equal measured coordinate. Three additional O2 ties have a zero total response and a nonzero answer component. Component zeros and sign ties are counted separately. Only the full 2×2 views were executed, transformed (A_K, E_a) cells are absent. K-anchor O3, O4, and K-versus-D comparisons are therefore unavailable. D-anchor headroom was outside the frozen predeclaration and was not evaluated. These are unavailable or unreported endpoints, not zero-valued results. The extension establishes no fixed-answer headroom result.

C.2 Matched history and co-evidence contrasts

C.2.1 Matched history contrasts

Relevant history is available for 75 of the 76 Freebase holdout cases. One CWQ case has no admissible neighborhood triple under the frozen R-H1 construction rule, while its co-evidence and privacy measurements remain available. The matched irrelevant-history control covers 68 of these 75 cases, seven lacked an eligible donor before control outcomes. Each donor shares the workload, relation register, and rendering format, preferring an identical relation template and then the smallest rendered-character gap within a tolerance of 18. Exclusions remove target-sensitive entity/relation overlap, visible candidate-pool connections, and specified cross-text entity matches. Outcome-independent matching establishes non-overlap in the artifacts, not semantic unrelatedness. Holding query, focal witness, and co-evidence fixed, each history has its own KEEP/DROP pair. For its raw signed response $dP(H, c)$, $\Delta_{\mathrm{rel}} = dP(H_1, c) - dP(H_0, c)$, $\Delta_{\mathrm{irrel}} = dP(H_{\mathrm{irrel}}, c) - dP(H_0, c)$, $\Delta_{\mathrm{spec}} = \Delta_{\mathrm{rel}} - \Delta_{\mathrm{irrel}}$. Specificity requires a 95% interval excluding zero and a mean with the relevant-history contrast's sign. All three families have specificity intervals including zero (Table C10), Gemma also reverses the expected sign. The original 75-case history aggregate is -1.5186 bits [-2.0715, -0.9661], with a MAP companion of +0.054 [-0.038, 0.152].

Table C10 uses the same 68 cases for every row, with 95% mean intervals from 2,000 case-bootstrap resamples, seed 0. Table C11 reports MAP=YES percentages on those 68 matched cases per cell. With only positive labels, these are truth-alignment rates, not balanced accuracy. Figure C3 connects the DROP and KEEP rates in Table C11 within each history arm, annotations round to whole percentages. For Qwen, relevant history raises DROP from 19.1% to 48.5%, while KEEP changes from 30.9% to 32.4%. Figure 2 reports signed-response contrasts and intervals.

Table C10: Matched history contrasts (raw signed bits).

Contrast	Mean	Mean 95% CI
Qwen3-32B		
Δ_{rel}	-1.600256	$[-2.174115, -1.047576]$
Δ_{irrel}	-1.300147	$[-1.832181, -0.785898]$
Δ_{spec}	-0.300109	$[-0.784188, +0.166834]$
GLM-4-32B-0414		
Δ_{rel}	-0.559403	$[-0.922102, -0.195746]$
Δ_{irrel}	-0.350059	$[-0.675710, -0.032060]$
Δ_{spec}	-0.209344	$[-0.483933, +0.022777]$
Gemma-3-27B-it		
Δ_{rel}	-1.697504	$[-3.291600, -0.197939]$
Δ_{irrel}	-2.545344	$[-4.431222, -0.756157]$
Δ_{spec}	0.847840	$[-0.291162, 2.078194]$

Table C11: Matched MAP=YES rates on positive-only cases (%).

History	Release	Qwen	GLM	Gemma
H_0	DROP	19.1	0.0	7.4
H_0	KEEP	30.9	4.4	29.4
H_1	DROP	48.5	0.0	1.5
H_1	KEEP	32.4	1.5	13.2
H_{irrel}	DROP	36.8	0.0	1.5
H_{irrel}	KEEP	29.4	0.0	8.8

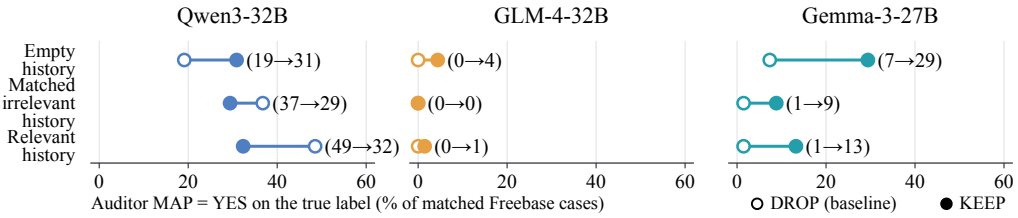

Figure C3: DROP (open) and KEEP (filled) MAP=YES rates under matched Freebase histories.

C.2.2 Co-evidence

On 76 paired Freebase cases, changing C_{EMPTY} to C_{SECOND} at fixed H_0, query, and witness shifts raw true-value response by -0.7367 bits [-1.3373, -0.1527]. Strict sign reversals occur in 43.4% of pairs and raw KEEP/DROP ranking changes in 44.7%. The CWQ mean is -0.2839 [-0.8907, 0.2900] and the WebQSP mean is -1.1896 [-2.2011, -0.1976], each over 38 cases. The MAP companion mean, +0.070 [-0.023, 0.173], is unresolved. Figure C4 partitions paired raw KEEP responses by sign before and after the second witness. Percentages use paired cases within each workload or the pooled cohort, zero involved denotes a zero response in either condition.

C.3 History-conditioned action rankings and minima

The state-action comparison uses the frozen 68-case matched-history construction and the same 7,228 admissible action identities in each of $H_0, H_1, H_{\text{irrel}}$. All 68 KEEP anchors are admissible. Each arm regenerates and scores its releases against its own matched DROP. The primary coordinate is the clipped $P_h(a) = [\text{dP}_{\text{true},h}(a)]_+$, fixed before scoring with policy weight one. For each case, define $D_{\text{rel}} = 1 - \rho(P_{H_0}, P_{H_1})$, $D_{\text{irrel}} = 1 - \rho(P_{H_0}, P_{H_{\text{irrel}}})$, and $\Delta_{\text{spec,rank}} = D_{\text{rel}} - D_{\text{irrel}}$. Spearman uses average ranks for exact ties, a constant coordinate makes the correlation unavailable, never zero. The paired contrast uses only cases with both correlations defined. On its 49-case common cohort, mean $D_{\text{rel}} = 0.727896$ and $D_{\text{irrel}} = 0.598252$. Each interval uses 2,000 whole-case percentile bootstrap replicates, seed 0, resampling paired differences conditional on the observed common evaluable cohort. The set contrasts are $J(H_0, H_1) - J(H_0, H_{\text{irrel}})$: the negative Intermediate result denotes lower similarity to H_0 under relevant history. The Pareto contrast remains unresolved. Figure C5 gives one point per case with nonconstant clipped costs in all three arms, points above the diagonal have greater displacement under relevant history. The annotation gives the paired mean difference and its 95% case-bootstrap interval.

Clipped P is nonconstant in 95.6% of cases under H_0, 76.5% under H_1, 85.3% under H_{irrel}, and 72.1% in all three. Every constant arm is zero throughout its eligible candidate set, there are no positive constants. Thus the arm-specific zero percentages are 4.4%, 23.5%, and 14.7%, respectively, and overlap across cases. The 19 excluded cases partition into 9 constant only under H_1, 3

Table C12: Matched-history ranking and set contrasts.

Primary paired contrast	Point	95% CI	Common cases
Clipped-P rank specificity	+0.1296	[+0.0741, +0.1887]	49
Intermediate Jaccard	-0.1009	[-0.1847, -0.0249]	39
Pareto Jaccard	-0.0105	[-0.0696, +0.0429]	68

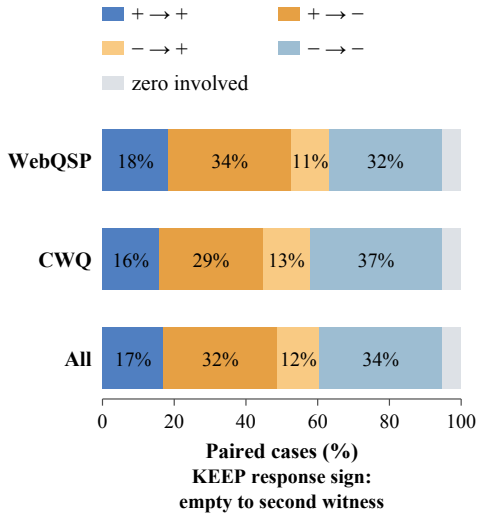

Figure C4: Co-evidence sign transitions (% of paired cases).

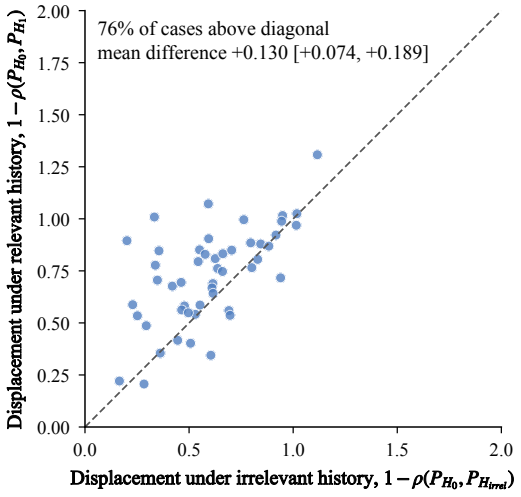

Figure C5: Paired action-ranking displacement by history.

only under H_{irrel}, 4 under both, and 3 under all three arms. The 95.6% share is an H_0 evaluability diagnostic, not a coordinate-selection rule.

Intermediate sets contain eligible transformed actions satisfying $P_h(\text{KEEP}) - P_h(a) > 0$ and $\Delta U_h(a) > 0$. Pareto sets cover the full eligible universe using the dominance rule in Appendix A.6, tied nondominated points all remain. Both-empty sets are counted separately and omitted from numeric Jaccard and identity denominators, one-empty comparisons have Jaccard zero.

Exact set identity occurs in 0.0% and 0.0% of numeric Intermediate comparisons, and 8.8% for each Pareto comparison. The paired Intermediate contrast averages within-case differences on 39 common cases, its two arm-specific means above use different denominators. The frozen secondary rank analyses apply the same rank and paired-bootstrap conventions to reader value and raw signed privacy. Each coordinate is nonconstant in every arm for all 68 cases. Component means have no predeclared intervals. For secondary raw action contrasts, $\delta_{\text{rel}}(a) = \text{dP}_{H_1}(a) - \text{dP}_{H_0}(a)$, $\delta_{\text{irrel}}(a) = \text{dP}_{H_{\text{irrel}}}(a) - \text{dP}_{H_0}(a)$, and $\delta_{\text{spec}}(a) = \delta_{\text{rel}}(a) - \delta_{\text{irrel}}(a)$. The point estimates below average within-case medians over 68 cases, sign percentages use the 7,228 eligible action rows as their denominator. No intervals were predeclared for these summaries. The descriptive minimum-cost-action grid uses the 17 fixed thresholds $u \in \{-6, -5, -4, -3, -2, -1.5, -1, -0.5, 0, 0.5, 1, 1.5, 2, 3, 4, 5, 6\}$. Among eligible KEEP, DROP, and transformed actions with $\Delta U_h(a) \geq u$, it records the minimum $P_h(a)$, an empty feasible set remains unattainable. Attainability percentages use all 1,156 cells per arm (3,468 pooled), tied/unique-minimum percentages condition on attainable cells (970, 975, 961, 2,906 pooled). Minimum-P ties occur in 90.7% of attainable cells (2,635/2,906). A post-hoc reduction of these scores finds zero cost in 95.4% (2,771/2,906), with six positive-cost tied minima. For $u > 0$, the pooled zero-cost and tie rates are 87.4% and 80.6% over 1,070 attainable cells. With DROP included, $u \leq 0$ gives zero cost by construction. No threshold is selected.

Table C13: Intermediate and Pareto overlap across histories.

Set	Pair with H_0	Mean Jaccard	Numeric cases	Both empty	One empty
Intermediate	H_1	0.140665	39	29	29
Intermediate	H_{irrel}	0.224285	42	26	24
Pareto	H_1	0.201973	68	0	0
Pareto	H_{irrel}	0.212517	68	0	0

Table C14: History effects on reader-value and raw-privacy ranks.

Secondary coordinate	D_{rel}	D_{irrel}	Difference [95% CI]
ΔU	0.365317	0.446178	-0.080861 [-0.127440, -0.038403]
Raw signed $\mathrm{dP}_{\text{true}}$	0.665810	0.543557	+0.122253 [+0.061939, +0.181939]

Table C15: Raw privacy changes under matched histories.

Contrast	Mean case median (bits)	Positive (%)	Negative (%)	Exact zero (%)
δ_{rel}	-1.499767	20.1%	78.9%	1.0%
δ_{irrel}	-1.064697	26.0%	73.0%	1.0%
δ_{spec}	-0.446891	38.8%	60.3%	0.9%

Table C16: Attainability and minimum-cost ties by history.

Arm	Attainable (%)	Unattainable (%)	Tied minima (%)	Unique minima (%)
H_0	83.9%	16.1%	83.1%	16.9%
H_1	84.3%	15.7%	97.1%	2.9%
H_{irrel}	83.1%	16.9%	91.8%	8.2%
All arms	83.8%	16.2%	90.7%	9.3%

Table C17: Post-hoc DROP true-label probabilities. Each case contributes once per arm.

Population / state	Median	IQR	$p^- \geq 0.90$	$p^- \geq 0.99$
Census, H_0	0.148	[0.042, 0.500]	6.2%	0.9%
Decision cohort, H_0	0.148	[0.020, 0.407]	3.3%	0.0%
Rescore, H_0	0.107	[0.032, 0.415]	2.9%	0.0%
Rescore, H_1	0.484	[0.133, 0.822]	17.6%	1.5%
Rescore, H_{irrel}	0.308	[0.060, 0.706]	1.5%	0.0%

Table C18: Signed responses before clipping (% of eligible transformations).

Population / state	Actions	$\mathrm{dP} < 0$	$\mathrm{dP} = 0$	$\mathrm{dP} > 0$
Census, H_0	8,206	3185/8206 (38.8%)	217/8206 (2.6%)	4804/8206 (58.5%)
Rescore, H_0	7,092	2716/7092 (38.3%)	33/7092 (0.5%)	4343/7092 (61.2%)
Rescore, H_1	7,092	5773/7092 (81.4%)	57/7092 (0.8%)	1262/7092 (17.8%)
Rescore, H_{irrel}	7,092	4557/7092 (64.3%)	56/7092 (0.8%)	2479/7092 (35.0%)

Table C19: Positive-threshold cells on the matched rescore. Full includes KEEP/DROP.

State	Inventory	Attainable / all	Zero / attainable	Tied / attainable
H_0	Full	358/544 (65.8%)	271/358 (75.7%)	245/358 (68.4%)
H_0	Transformed	357/544 (65.6%)	261/357 (73.1%)	245/357 (68.6%)
H_1	Full	363/544 (66.7%)	355/363 (97.8%)	335/363 (92.3%)
H_1	Transformed	361/544 (66.4%)	352/361 (97.5%)	317/361 (87.8%)
H_{irrel}	Full	349/544 (64.2%)	309/349 (88.5%)	282/349 (80.8%)
H_{irrel}	Transformed	346/544 (63.6%)	306/346 (88.4%)	264/346 (76.3%)

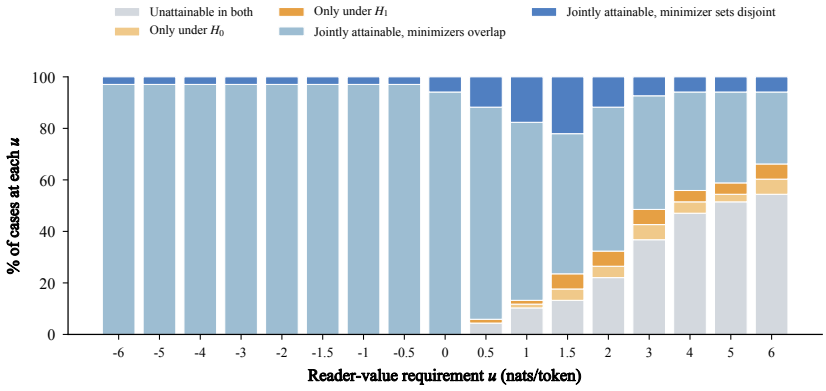

Figure C6: History-conditioned attainability and minimizer overlap (% of cases at each u).

State-conditioned minimizer sets. A separate post-hoc comparison retains only eligible transformed actions, excluding KEEP and DROP, on the same 68 cases and 17 thresholds. Each history arm uses its own regenerated answers and matched-DROP coordinates. Let $\mathcal{F}_h(c,u) = \{a : \Delta U_h(a) \geq u\}$ and $M_h(c,u) = \arg\min_{a \in \mathcal{F}_h(c,u)} P_h(a)$. All exact minimizers remain, without tie-breaking, a nonempty constant-cost set has $M_h = \mathcal{F}_h$. Its attainment counts differ from the full-inventory summary above because KEEP and DROP are excluded. Of 1,156 case-threshold cells, 82.1% are jointly attainable, 1.7% are attainable only under H_0, 2.1% only under H_1, and 14.1% under neither. The minimizer sets are disjoint in 8.4% of jointly attainable cells. In 5.5% of the 949 jointly attainable cells, $M_{H_0} \cap \mathcal{F}_{H_1} = \varnothing$: none of the empty-history minimizers meets the relevant-history reader-value constraint. Where this intersection is nonempty, define

$$R_{\text{state}}(c,u) = \min_{a \in M_{H_0} \cap \mathcal{F}_{H_1}} P_{H_1}(a) - \min_{a \in \mathcal{F}_{H_1}} P_{H_1}(a).$$

This compares the best feasible empty-history minimizer with the relevant-history minimum. It is defined in 897 cells and zero in 96.9%, on the remaining 3.1% of these cells, its median and p90 are 0.114 and 0.707 bits. On the 49-case nonconstant-coordinate subset (689 jointly attainable cells), minimizers are disjoint in 8.9% of jointly attainable cells, with no feasible empty-history minimizer in 4.8%. These descriptive cell counts depend on the specified threshold grid. Choosing the best feasible old minimizer uses the relevant-history scores, so this is an optimistic reuse comparison, not the performance of a previously fixed selection rule. The cells repeat cases across thresholds and are not independent observations. Keeping all minimizers removes the identifier tie-break. Clipping and numerical resolution still affect these sets. Figure C6 partitions matched cases at each reader-value threshold u (nats per reference-answer token) by attainability under empty and relevant histories and overlap of the exact minimizer sets.

Figure C7 illustrates WebQTest-1254 with the same candidate inventory under empty history and a relevant prior turn about the 2012 Masters. Each panel uses its own matched DROP baseline, privacy cost is in bits and reader value in nats per reference-answer token. The points describe this selected case, with no aggregate prevalence estimate.

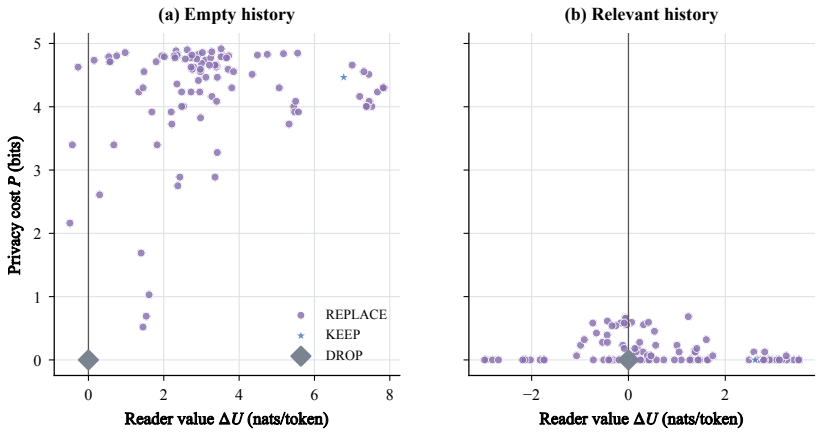

Figure C7: WebQTest-1254 release coordinates under empty and relevant histories.

DROP probabilities, clipping, and zero-cost minima. A post-hoc reduction of the frozen scores separates baseline probabilities, signed responses, and constrained minima. Table C17 gives one DROP probability per case and state. The original census and three-arm rescore are separate executions, each using its own paired coordinates. All true labels are YES. Census workload medians are 0.040 for WebQSP (38 cases), 0.212 for CWQ (38), and 0.407 for KGQAGen (37).

The canonical baseline is below 0.90 in 93.8% of privacy-evaluable cases. In the matched rescore, both history arms have higher baseline medians than empty history. The remaining positive log-ratio is bounded by $-\log_2 p^-$, whose median is 3.228 bits under H_0, 1.046 under H_1, and 1.703 under H_{irrel}. These distributions describe available increment, without isolating the causal contribution of parametric knowledge. In the census, 41.5% of eligible transformed actions have clipped cost zero. Of these, 93.6% have strictly negative signed response. Equal weighting of cases with transformations gives 37.3% zero-cost actions, compared with the action-weighted 41.5%. Zero cost therefore pools negative and exactly zero responses. The rescore's 7,228 eligible identities include 68 KEEP and 68 DROP anchors, leaving 7,092 transformations per history arm.

Table C19 restricts the existing grid to $u \in \{0.5, 1, 1.5, 2, 3, 4, 5, 6\}$, giving 544 case-threshold cells per arm. DROP is infeasible at these thresholds. Zero minima remain frequent after excluding both anchors, reaching 97.5% of attainable transformed-only cells under relevant history. This limits the distinctions available from the scalar minimum, even when feasible actions or minimizing identities change. Each row reports attainability separately. Zero and tie percentages condition on its attainable cells, and a tied minimum retains every exactly minimizing action. Unattainable cells have no cost value. Repeated case-threshold cells are descriptive grid summaries, not independent prevalence samples. This rescore reduction evaluates likelihood only.

Table C20: CQ seed-macro history contrasts in bits.

Family	Δ_{native} [95% CI]	Δ_{irrel} [95% CI]	Δ_{spec} [95% CI]	Criterion
Qwen	-0.468022 [-0.833300, -0.094657]	+0.300183 [-0.053953, +0.501007]	-0.768205 [-1.056554, -0.478993]	Retained
Gemma	+0.432148 [-0.628592, +1.462566]	-1.843334 [-2.964700, -0.862147]	+2.275481 [+1.122770, +3.013907]	Retained

Table C21: CQ seed-macro MAP=YES rates (%).

	H_0		H_{native}		H_{irrel}	
Family	KEEP	DROP	KEEP	DROP	KEEP	DROP
Qwen	45.6	10.0	49.7	14.6	40.4	6.6
Gemma	51.8	10.6	33.2	6.2	16.0	2.2

C.4 Natural conversational histories

The separate ConvQuestions construction starts from 1,400 prelisted target turns and excludes 24 for current-answer leakage, 511 non-entity answers, 308 without a qualifying relation, 103 ambiguous bindings, and one unlinked query entity. The 453 admitted targets come from 254 seed conversations. Histories are benchmark QA turns, not deployment logs. Truth uses a pinned Wikidata observation window, evidence retains QIDs/PIDs and typed literals, while propositions use pinned labels with identifier fallback.

The seed-only construction binds the conversation seed entity to the current official entity-linked answer through a unique eligible direct relation. Preferred statements take precedence over normal ones, deprecated and non-entity links are excluded. Multiple distinct bindings are excluded. The single-triple witness W directly instantiates the protected membership proposition, whose variable is $Z \in \{\text{NO}, \text{YES}\}$ with $z^* = \text{YES}$. Unavailable lookups and window-relative absence exclude candidates without creating negative labels. Matched irrelevant donors share domain, preceding-turn count, and rendering, with target-sensitive overlap exclusions and character tolerance 11. Forty-one targets have no eligible donor and remain only in the native-history analysis. The 412 three-arm targets represent 242 seeds, 175 donor-seed reuses induce 15 target-donor components, the largest containing 58 seeds. Native and irrelevant contrasts subtract each target's empty-history response, their difference is specificity. Table C20 uses 2,000 bootstrap resamples, seed 0. Individual target rows are not independent resampling units. Targets are averaged within seeds, then seeds receive equal weight. Native intervals resample seeds, irrelevant-history and specificity intervals resample the 15 target-donor components and reapply seed-macro weighting. Both families use the same cohort and percentile interval procedure. The CQ cross-family comparison includes Qwen and Gemma. GLM was not run because the model snapshot and effective chat template available in the CQ execution environment could not be verified as matching the frozen GLM configuration.

Both families satisfy the predeclared within-family criterion: the specificity interval excludes zero and its point estimate has the same nonzero sign as the three-arm native-history point estimate. Gemma's native interval includes zero, exclusion of zero by that interval is not part of the criterion. Every point-estimate contrast reverses direction across families, so cross-family directional replication is not supported. Both families use identical semantic views and the same execution stack. The historical Qwen all-admitted native companion remains -0.475583 [-0.811152, -0.142997] on 453 targets and 254 seeds, including the 41 donor-unavailable targets. The corresponding seed-macro $\Pr[\text{MAP} = \text{YES}]$ percentages are response companions on the three-arm cohort, with no predeclared intervals. Figure C8 compares KEEP and matched DROP within each history arm, percentages average target-turn indicators within seeds before averaging seeds equally.

The CQ protocol specifies equal weight for seed conversations and uses target-donor components as bootstrap units. An initial implementation instead assigned equal weight to components. For Qwen, component weighting gives specificity -0.565 [-1.163, +0.079] and irrelevant-history change -0.039 [-0.445, +0.347], the specified seed weighting gives -0.768 [-1.057, -0.479] and +0.300 [-0.054,

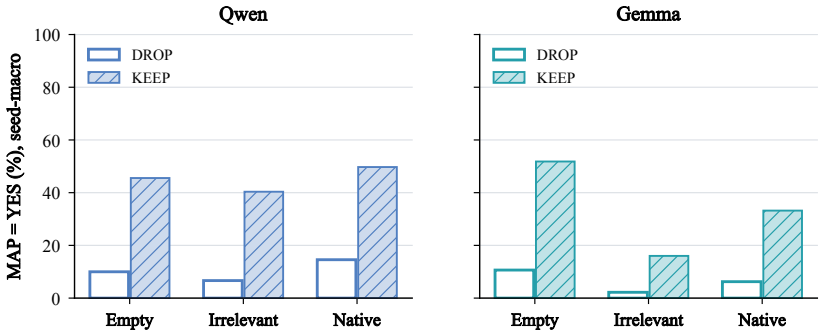

Figure C8: CQ seed-macro MAP=YES rates by history and released output (%).

Table C22: Two-class calibration on the controlled population.

Auditor	NLL	Brier	ECE_{true}	ECE_{MAP}	MAP accuracy (%)
Qwen3-32B	2.483	0.390	0.385	0.373	58.1
Qwen2.5-32B-Instruct	8.032	0.403	0.393	0.399	58.5
GLM-4-32B-0414	1.6098	0.3982	0.4112	0.3776	55.2
Gemma-3-27B-it	6.3978	0.3975	0.3826	0.3933	59.6

+0.501]. The implementation correction changes the specificity criterion from unmet to retained, the irrelevant-history interval includes zero under both weightings.

In the historical Qwen sensitivity analysis, leaving out each dependency component retains the criterion in 100.0% of deletions: specificity estimates range from -0.805789 to -0.661695 bits, with largest absolute shift 0.106510 bits. Each deletion keeps the estimator and resamples the remaining 14 components, without donor rematching. Qwen shows attenuation within this direct-binding, matched-donor construction. Table S1 gives full component-deletion rows and MAP/absolute-exposure companions.

C.5 Calibration and construct validity

C.5.1 Workload response

On the 105 paired eligible KEEP-anchor cases, MAP=YES rises from 21.9% under DROP to 38.1% under KEEP: paired difference +16.2 percentage points [7.6, 25.7]. Mean true-value probability change is +0.1206 [0.0594, 0.1819] on the probability scale. The corresponding log2-ratio mean is +0.3932 bits [-0.0273, 0.8357]. The quantities have different units and baseline sensitivity. All true labels are YES. KGQAGen's evidence uses identifiers while its protected facts use labels. All 12 witness-no-overlap cases are CWQ, confounding workload and overlap in this stratum.

C.5.2 Controlled calibration

Each auditor scores the same isolated 480 controls, balanced between 240 YES and 240 NO labels. Table C22 reports absolute errors. The predeclared relative-error fallback did not trigger for either designated Qwen auditor. ECE uses frozen equal-count bins on a controlled population separate from the workloads.

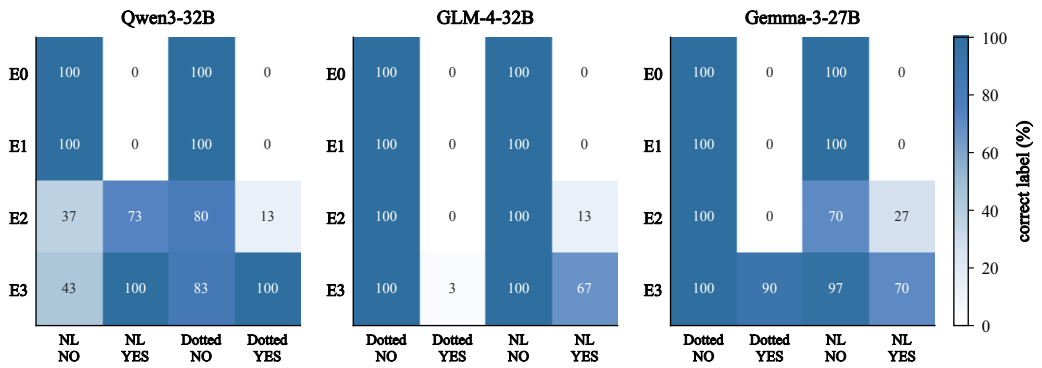

Figure C9: Correct-label rates by exposure, register, and auditor (%).

Table C23: Canonical Qwen correct-label rates by exposure and register.

Exposure	Register	NO correct	YES correct
E0	NL	100.0%	0.0%
E0	Dotted	100.0%	0.0%
E1	NL	100.0%	0.0%
E1	Dotted	100.0%	0.0%
E2	NL	36.7%	73.3%
E2	Dotted	80.0%	13.3%
E3	NL	43.3%	100.0%
E3	Dotted	83.3%	100.0%

Table C24: Cross-family correct-label rates at E2 and E3.

Auditor	E2 correct	E3 correct
Qwen3-32B	50.8%	81.7%
GLM-4-32B-0414	53.3%	67.5%
Gemma-3-27B-it	49.2%	89.2%

C.5.3 Exposure and register

E2 conveys an entailing specialization of the membership or explicit-negation relation, E3 states that relation directly. Canonical Qwen has mean MAP confidence 0.954 despite overall accuracy 58.1%. E0 and E1 produce default-NO responses, yielding 50.0% accuracy on balanced classes. Table C23 retains the exposure-by-register class counts, Table C24 compares families. Explicit exposure is recognized more reliably than entailing exposure, with register and class asymmetries remaining. Full GLM/Gemma controlled cells appear in Table S2. In Table C23, each class cell contains 30 items, NL denotes natural-language relation text and dotted denotes the dotted-path register. Figure C9 compares exposure-by-register cells across auditors. Each percentage uses the controls with the indicated exposure, register, and true label, column labels specify the register order within each panel. Each E2/E3 cell in Table C24 contains 120 items balanced between YES and NO.

Table C25: Matched complete outputs for WebQTest-859 under H_0/C_{EMPTY}.

Action	Released evidence	Generated answer
KEEP	Russia-religions-○; ○-religion-Russian Orthodox Church	The predominant religion in Russia is the Russian Orthodox Church.
REPLACE	m.04lmlck-military.casualties.combatant-Russia	`INSUFFICIENT_EVIDENCE`
DROP	none	`INSUFFICIENT_EVIDENCE`

Action	$\mathrm{dP}_{\mathrm{true}}$	$P(a)$	ΔU	S_P
KEEP	1.7897	1.7897	0.5821	0.0000
REPLACE	1.1001	1.1001	0.8033	0.6896
DROP	0.0000	0.0000	0.0000	-

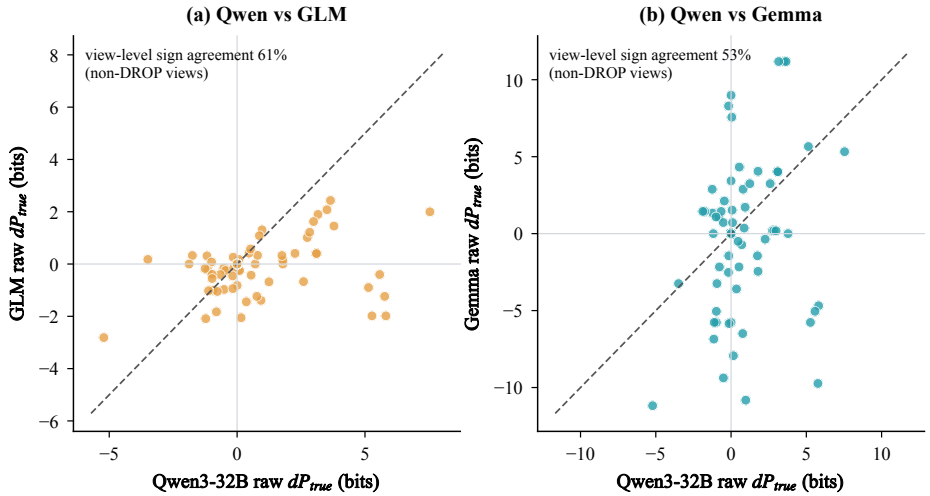

Figure C10: Paired cross-auditor raw responses on matched views (bits).

C.5.4 A MATCHED COMPLETE-OUTPUT EXAMPLE

WebQTest-859 asks "what are the predominant religions in russia." The protected fact states that Russia has value Sovereign state under `base.aareas.schema.administrative_area.administrative_area_type`. The first qualifying Intermediate action in frozen order releases a graph-admitted combatant relation. This evidence does not state a religion, yet both model-response coordinates rise relative to DROP and the reader abstains (Table C25). Per-action support labels for transformed releases are unavailable (Appendix D.1.1). Selected response examples appear in Table S3. In Table C25, DROP anchors $\mathrm{dP}_{\mathrm{true}}$, while KEEP anchors the saving S_P reported for KEEP and the transformed action. Privacy coordinates and savings use bits, ΔU uses nats per reference-answer token. The symbol ○ denotes the same unlabeled intermediate node.

Table C26: Cross-family agreement on the frozen action subset.

Quantity	Qwen vs GLM Raw	Qwen vs GLM Policy	Qwen vs Gemma Raw	Qwen vs Gemma Policy
Mean casewise Spearman	0.1653	0.6278	-0.0609	0.2281
Ranking evaluable (%)	100.0%	66.7%	100.0%	83.3%
Ranking unavailable (%)	0.0	33.3	0.0	16.7
Privacy-saving sign agreement	59.0%	61.5%	43.6%	46.2%
Intermediate intersection count	2	2	4	3
Intermediate Jaccard	0.2000	0.4000	0.3077	0.2500
Identical subset Pareto sets (%)	33.3%	66.7%	16.7%	41.7%
MAP correct: GLM 6.5%; Qwen 29.0%				
GLM NO→YES transitions: 3				
MAP correct: Gemma 24.2%; Qwen 29.0%				
Gemma NO→YES transitions: 10				

State contrast	Sign agreement	Mean bits	
Family / state	with Qwen	Qwen	Family
GLM / H_1	81.8%	-2.9433	-1.5241
GLM / C_{SECOND}	66.7%	-1.6987	-0.1448
Gemma / H_1	81.8%	-2.9433	-2.6238
Gemma / C_{SECOND}	50.0%	-1.6987	0.5096

C.6 Cross-family action agreement

The frozen subset contains 97 matched views from 12 cases, corresponding to 194 continuation units and 39 transformed/anchor actions. Reader values remain fixed. Raw signed privacy and clipped coordinates are evaluated separately, MAP alignment uses 62 non-DROP views with state-specific DROP baselines. Agreement with Qwen is partial for GLM and weak for Gemma, whose mean co-evidence contrast reverses sign. Evaluable counts differ across casewise correlations (Table C26). Sign agreement uses 39 action contrasts, 11 history contrasts, and 12 co-evidence contrasts, case rates use all 12 cases. Constant-coordinate correlations are excluded as unavailable. Intermediate intersection counts and Jaccard are distinct analyzer outputs. The 62-view MAP population contains non-DROP KEEP and transformed views across the frozen states, its positive-only correctness measures alignment, not balanced accuracy. This subset does not establish auditor invariance. Figure C10 plots the matched views against Qwen. Its sign annotations use non-DROP views, the table's saving-sign agreement instead uses action-level contrasts.

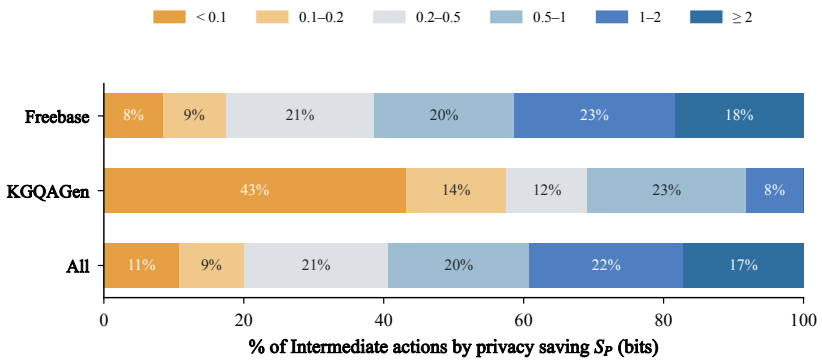

Figure D1: Intermediate-action privacy savings by workload (% of actions).

D RQ2: Truthful Alternatives and Release Selection

D.1 Transformation availability and value

D.1.1 Action availability and populations

Across 114 confirmatory cases, 66.7% [57.9%, 75.4%] admit a nonempty proper truthful subset. Every one of the 76 Freebase cases has a truthful connected replacement, KGQAGen lacks a replacement pool. The witness-level structural support-retention diagnostic finds a MINIMIZE candidate in 0.0% of Freebase proxy cases and 38.5% of KGQAGen exact-support cases. The support-diagnostic denominators are 76 Freebase and 26 KGQAGen cases. It uses the original witness, a legacy gate, and greedy backward elimination, it is not per-action support of the scored transformed releases, which is unavailable. The decision cohort contains 61 disclosure-positive eligible-KEEP cases (38 Freebase, 23 KGQAGen) from 105 eligible-KEEP cases, with 4,104 transformed actions (2,225 Intermediate and 1,879 non-Intermediate). Figure 3 reports its existence rates, Figure 4(a) gives action-weighted outcomes. Likelihood, non-sentinel, Hits@1, strict-exact, and same-binary-status existence are 96.7%, 63.9%, 42.6%, 18.0%, and 91.8%, respectively. The 95% whole-case bootstrap intervals for likelihood, Hits@1, strict-exact, and same-binary-status existence are respectively [91.8%, 100.0%], [31.1%, 54.1%], [9.8%, 27.9%], and [83.6%, 98.4%]. The saving median/p90 of 0.4884/1.9352 bits summarize within-case medians across the 59 Intermediate-exists cases. Figure D1 groups the decision cohort's Intermediate actions by saving in bits. Each bar uses all Intermediate actions in that workload, or in the pooled cohort, as its denominator. These action-weighted bins describe saving magnitudes, they do not estimate scoring error. Figure 3 expands the criterion comparison by workload. These post-hoc intervals use the aggregate estimates' grouped bootstrap: 2,000 whole-case resamples per endpoint with seed 0.

The original-KEEP reference comparison has 66 disclosure-positive cases regardless of KEEP admission: likelihood, non-sentinel, Hits@1, strict-exact, and same-binary-status existence are 97.0%, 59.1%, 39.4%, 16.7%, and 92.4%, same-status existence among the 62 overlap-positive KEEP cases is 91.9%. Its saving median/p90 over the 64 Intermediate-exists case medians is 0.5130/1.9958 bits. These reference results are separate from decision-cohort estimates. The full transformed census comprises 112 cases and 8,206 eligible transformations. Table D1 describes family availability and multiplicity on that population. Across all its transformed actions, saving has median 0.0000 and p90 1.0072 bits, 32.6% of actions have exactly zero saving and fail the strict Intermediate condition. These action-weighted summaries differ from the decision-cohort distribution of case medians. Table D1 conditions each family denominator on cases with an eligible action and evaluable privacy measurements, these overlapping denominators are not additive. Nondominated counts average transformed actions on each case's full eligible Pareto set. Both scalars use $P^{(1)}$, brackets give case-bootstrap 95% intervals.

Table D1: Transformed-action prevalence and Pareto nondominance.

Transformed family	Eligible cases	Primary Intermediate prevalence, %	MAP companion prevalence, %	Primary mean nondominated count
MINIMIZE ∪ REPLACE	112	57.1 [47.3, 66.1]	53.6 [44.6, 62.5]	2.509 [2.161, 2.911]
MINIMIZE	74	50.0 [39.2, 60.8]	56.8 [46.0, 67.6]	1.108 [0.811, 1.419]
REPLACE	76	48.7 [36.8, 59.2]	50.0 [39.5, 60.6]	2.618 [2.145, 3.119]

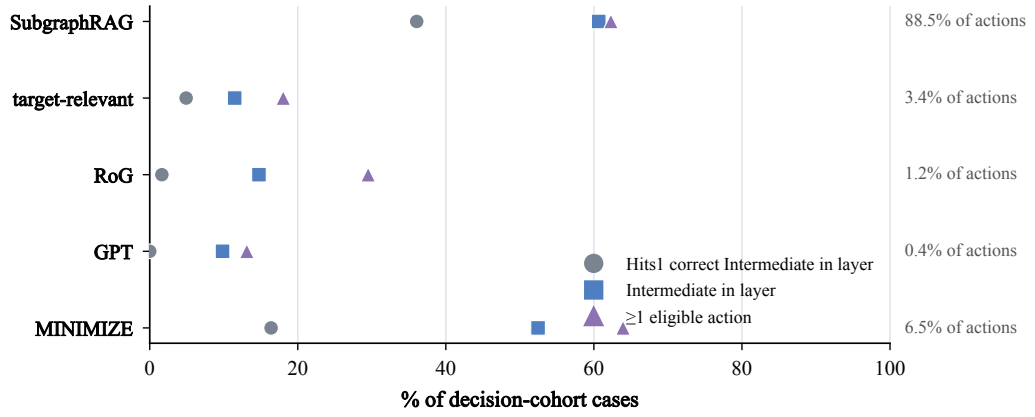

Figure D2: Candidate-source coverage and availability of correct alternatives (%).

Candidate-source coverage. The post-hoc source diagnostic partitions the 4,104 decision-cohort transformations by retained provenance. Replacement pools use first-occurrence precedence (SubgraphRAG, target-relevant, RoG, then GPT), 54 duplicated action identities are resolved by the frozen first-occurrence rule. MINIMIZE is a separate action family. Of the 61 cases, 23 are KGQAGen cases with a missing replacement source and contribute MINIMIZE only. Action shares use all 4,104 transformations, case coverage uses all 61 cases and can overlap across sources. Figure D2 shows source-specific case coverage with the full decision cohort as denominator. Right-hand annotations give each source's share of all eligible transformations under the stated precedence rule. Using the same-run coordinates and minimizer-set comparison of Appendix C.1.1, evidence-only and complete-output minima are disjoint in 11.3% of SubgraphRAG cells, 21.1% of target-relevant cells, 10.9% of RoG cells, 19.6% of GPT cells, and 21.6% of MINIMIZE cells. The respective denominators are 515, 95, 165, 46, and 504 cells, retaining at least two feasible actions from each layer. Of the 1,037 cells per layer, respectively 19/503, 45/897, 29/843, 50/941, and 28/505 have one/zero feasible actions and are excluded from this comparison. Source precedence determines which layer receives shared candidates, the resulting coverage describes the retained inventory, not independent generator capability.

Table D2: Action-weighted generated-answer diagnostics on the full transformed census.

Group	Hits@1	Strict exact	Sentinel rate
KEEP	49.1%	24.1%	33.9%
Intermediate	6.7%	2.5%	85.1%
Non-Intermediate transformed	13.0%	3.6%	-
DROP	0.9%	0.9%	99.1%

D.1.2 EXACT-OVERLAP RESTRICTIONS

The binary indicator tests exact whole-string equality of the frozen secret-topic identifier with an entity-valued subject or object. It tests neither relation labels nor semantic similarity. Overall, preserving KEEP's binary status retains Intermediate existence in 91.8% of decision-cohort cases. Among 57 overlap-positive KEEP cases, binary preservation retains 91.2% [84.2%, 98.2%]. Counting each matching triple once, $N_{\mathrm{ov}}(a) \geq N_{\mathrm{ov}}(\mathrm{KEEP})$ leaves 1,860 eligible transformations, including 922 Intermediate actions, and retains existence in 52.6% [40.4%, 64.9%]. Intervals use 2,000 whole-case resamples, seed 0. A substantial part of binary-status-preserving headroom depends on allowing multiplicity below KEEP.

D.1.3 GENERATED-ANSWER DIAGNOSTICS

The sentinel diagnostic strips surrounding Unicode whitespace, case-folds, and tests exact membership in `INSUFFICIENT_EVIDENCE`, `INSUFFICIENT EVIDENCE`, `I DON'T KNOW`, and `UNKNOWN`. Of 9,109 inspected generations, 76.1% use the first string and 0.0% use the others. Observed hits therefore detect the prompt-specified insufficient-evidence response. The reader emits it when it judges the supplied evidence insufficient, the diagnostic is not abstention accuracy and does not change the auditor-visible output.

RoG-style Hits@1 lowercases generated text and frozen references, removes ASCII punctuation and whole-word English articles, normalizes whitespace, and tests whether any normalized reference occurs in the normalized generation (Luo et al., 2024). It uses original pre-join answer lists without extraction, alias expansion, or a judge. Reference likelihood scores the concatenated string defined in Appendix A.4. This generated-text diagnostic can be permissive for long answers and is not official benchmark accuracy. The strict exact-match diagnostic instead uses the existing normalized full-string equality and ALL-reference convention. Its normalization and gold rendering remain distinct. On the 2,240 census Intermediate actions, 2.5% hit both rules, 4.2% hit only Hits@1, and 0.0% hit only strict equality, this is empirical agreement, not a general implication between rules.

Table D2 reports percentages on the 112-case census: 112 actions each for KEEP/DROP, 2,240 Intermediate actions across 64 cases, and 5,966 non-Intermediate transformations across 103 cases. The frozen sentinel summary omits the full-census non-Intermediate rate, Figure 4(a) supplies it separately for the 61-case decision cohort. Positive reference likelihood and generated-answer correctness only partly coincide. Correctness and overlap restrictions are separate marginal analyses, their joint coverage is not measured. The binary closed-book diagnostic is exact correctness of the empty-history, empty-co-evidence DROP generation, separate from the continuous Pure-CB score. The complete support-status/closed-book cross-tabulation is in Table S4. Here "Hits@1-correct" means a generated reference hit under the stated matching rule. Neither that rule nor strict exact matching verifies that the released evidence entails the answer, their different reference conventions also prevent treating them as formal upper and lower correctness bounds. Query-answer support does not test whether released evidence entails the protected fact.

Table D3: KEEP correctness and correct-Intermediate availability.

KEEP generation	Correct Intermediate Exists (%)	Does not exist (%)	Row n
Hits@1-correct	73.3	26.7	30
Hits@1-incorrect	12.9	87.1	31
Total	42.6	57.4	61

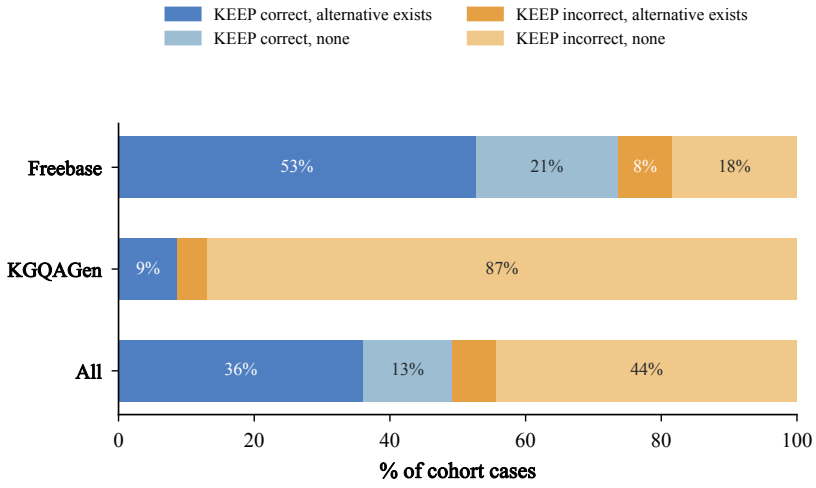

Figure D3: KEEP correctness and correct-alternative availability (% of cohort cases).

D.2 Post-hoc correctness of minimum-cost alternatives

Correctness pairing. The post-hoc pairing uses frozen retrospective Hits@1 labels for the 61-case decision cohort. The table crosses KEEP's status with the existence of a Hits@1-correct Intermediate alternative, the first two columns are within-row percentages, with row denominators in the final column. Among the 30 KEEP-Hits@1-correct cases, 73.3% have a Hits@1-correct Intermediate alternative. The primary 95% interval [56.7%, 86.7%] conditions on the observed 30-case cohort. A secondary interval [57.6%, 88.0%] resamples the 61-case parent population and reapplies the KEEP-correct predicate within each replicate. Both use 2,000 whole-case percentile bootstrap replicates with seed 0. This conditional question concerns the frozen Hits@1 status, the 96.7% likelihood and 42.6% correct-Intermediate existence results use the full decision cohort. Figure D3 partitions cases by KEEP Hits@1 correctness and availability of a Hits@1-correct Intermediate alternative. Percentages use each workload's decision cohort, or the full cohort for the pooled row.

The 22 cases contribute 132 Hits@1-correct Intermediate actions. Using the original frozen $S_P(a) = P(\text{KEEP}) - P(a)$, the median and p90 of within-case median savings are 0.2616 and 0.8878 bits. Within each of these cases, the frozen minimum-P selector chooses among the qualifying actions, breaking ties by ascending `repr(action_id)`. The selected answer is byte-identical to KEEP in 31.8% of these cases and different in 68.2%, 18.2% have a minimum-P tie before tie-breaking. Byte identity is not semantic equivalence.

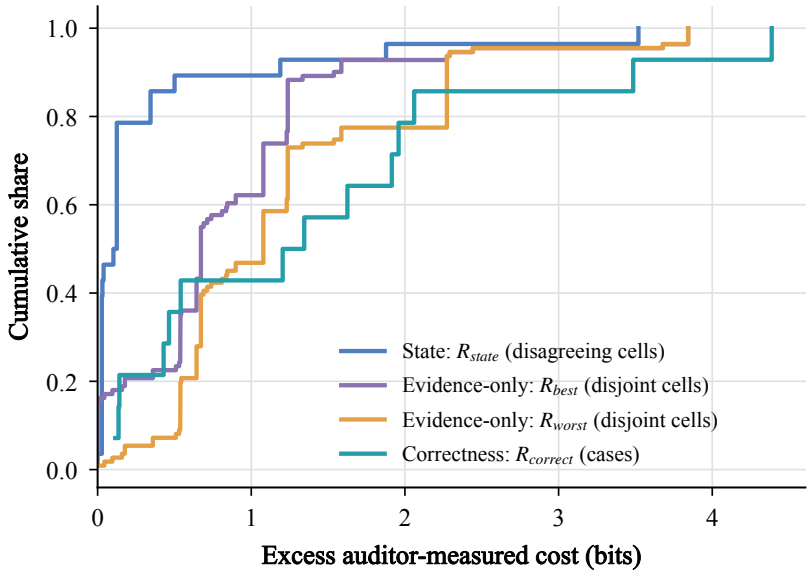

Figure D4: Excess-cost distributions for state, scoring scope, and correctness (bits).

Correctness among minimum-cost Intermediate actions. A post-hoc companion uses the original published P, the strict Intermediate definition, and all 26 cases with a Hits@1-correct Intermediate alternative. Let I_c be the eligible Intermediate set and $M_{\mathrm{lik},c} = \arg\min_{a \in I_c} P(a)$, retaining every exact minimizer without tie-breaking. The minimum-cost set contains a correct action in 46.2% of cases and contains only correct actions in 15.4%, 65.4% of cases have tied minima. In the remaining 53.8% of cases ($n = 14$), every minimum-cost Intermediate action is Hits@1-incorrect. Define

$$R_{\mathrm{correct}}(c) = \min_{a \in I_c : \mathrm{Hits@1}(a)=1} P(a) - \min_{a \in I_c} P(a).$$

Among those 14 cases, this additional auditor-measured cost for a correct alternative has median 1.275 and p90 3.058 bits. These are complete argmin-set results, the R-nonsentinel selections in Appendix D.6 additionally restrict generated text and use a numerical tie tolerance. Figure D4 conditions the evidence-only curves on disjoint minimizer cells. The state curve uses cells with positive R_{state} (Appendix C.3), the correctness curve uses cases whose minimum-cost Intermediate set contains no reference hit. Each empirical cumulative curve has its own cell or case denominator.

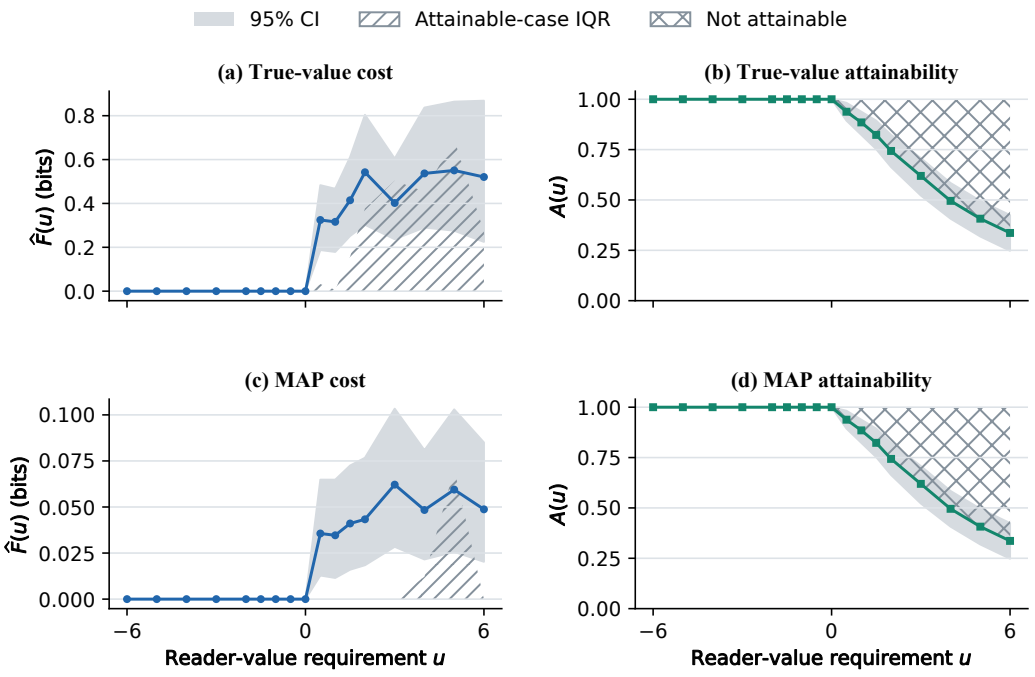

Figure D5: Exploratory conditional cost and attainability for true-value and MAP privacy.

D.3 Conditional funnel and attainable populations

Figure D5 shows the true-value funnel and its MAP companion. The exploratory grid has 17 points. Conditional cost averages attainable cases, so changing their composition can lower the mean. The left panels condition minimum eligible-action cost on $dU \geq u$, the right panels report attainability over 113 privacy-evaluable cases. In this single-secret, single-auditor evaluation, $P^{(1)} = P^{(2)}$, so each scalar occupies one row. Lines show stored means, gray bands show 95% case-bootstrap intervals, and hatched cost bands show case-level interquartile ranges. Crosshatching marks unattainability. Both u and dU use nats per reference-answer token. The grid was reconstructed after holdout evaluation and selects no confirmatory operating point. Adding a Hits@1-correct-generation constraint retains the same 113-case population and utility grid. At $u = 0$, likelihood-only attainability is 100.0% with cost 0.0000 bits. The constrained full universe attains 62.8% (95% CI [54.9%, 71.7%]) with cost 0.7830 [0.5082, 1.0921] over those 71 cases, transformed-only attains 59.3% [50.4%, 68.1%] with cost 0.8734 [0.5949, 1.1908] over those 67. Attainability intervals resample the full population, while cost intervals resample attainable cases.

Across the same 17 thresholds and 113 cases, the post-hoc grid comparison yields 83.8% attainable cells under likelihood alone and 52.3% with the additional Hits@1 constraint. The 31.5% attainable only under likelihood uses all 1,921 case-threshold cells as its denominator. Both comparisons include eligible KEEP, DROP, and transformed actions. Figure 4(b) reports attainability in the main text, Figure D6 additionally reports conditional cost for the same likelihood-only reference and Hits@1-constrained full and transformed-only inventories. Bands are pointwise 95% case-bootstrap intervals: attainability uses the full privacy-evaluable cohort, while cost conditions on attainable cases. Thresholds use nats per reference-answer token and costs use bits.

Zero-cost and tied minima at positive thresholds. Table D4 uses the eight positive grid thresholds listed in Appendix C.3. Full includes admitted KEEP/DROP and transformations. Transformed excludes both anchors, retaining cases without an eligible transformation as unattainable. Exact minima are computed from the original paired scores, without a new numerical tolerance. The additional Hits@1 requirement changes both attainability and the population underlying each conditional per-

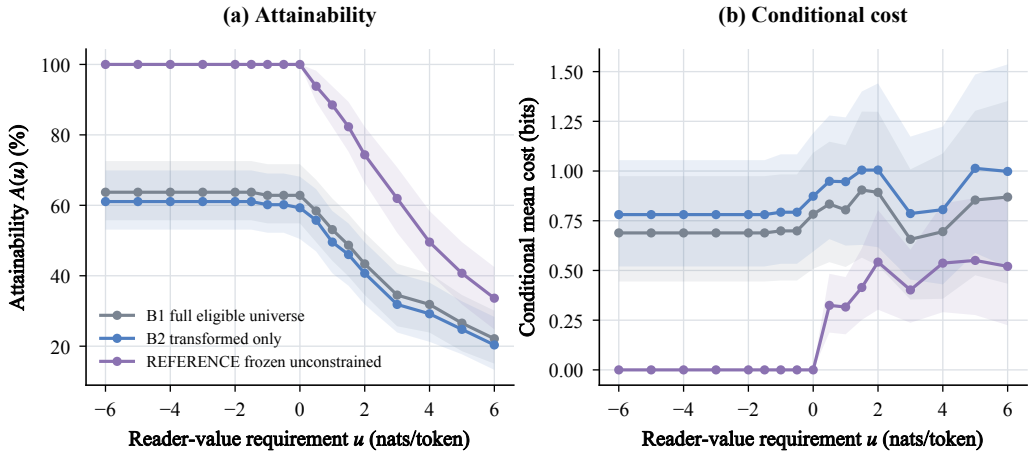

Figure D6: Attainability and conditional cost with a reference-hit requirement (95% CIs).

Table D4: Post-hoc minima over census cases and eight positive thresholds.

Inventory	Requirement	Attainable / all	Zero / attainable	Tied / attainable
Full	Likelihood	593/904 (65.6%)	414/593 (69.8%)	382/593 (64.4%)
Full	Likelihood + hit	360/904 (39.8%)	176/360 (48.9%)	110/360 (30.6%)
Transformed	Likelihood	590/904 (65.3%)	399/590 (67.6%)	380/590 (64.4%)
Transformed	Likelihood + hit	337/904 (37.3%)	142/337 (42.1%)	105/337 (31.2%)

centage. For $u \leq 0$, DROP guarantees zero cost only for the likelihood-only full inventory, since it need not pass Hits@1. These grid summaries are descriptive.

D.4 Release-margin sensitivity

The post-hoc analysis retains the original published complete-output coordinates on the 61-case decision cohort and its 4,104 eligible transformed actions. It counts cases containing an action with $S_P(a) > \epsilon$ and $\Delta U(a) > \delta$, for $\epsilon \in \{0, 0.1, 0.25, 0.5, 1\}$ bits and $\delta \in \{0, 0.05, 0.1, 0.25, 0.5\}$ nats per reference-answer token. A second version adds $\Delta U(a) \geq \Delta U(\text{KEEP})$. Reference-hit-correct existence additionally requires the frozen generated answer to satisfy Hits@1. The KEEP-correct subset contains 30 cases. This grid measures sensitivity to the numerical requirements, it does not estimate a measurement-noise threshold or select an operating point.

At zero margins, the additional KEEP-value condition retains likelihood headroom in 70.5% of cases and correct-answer headroom in 32.8%, including 56.7% of KEEP-correct cases. Without that additional condition, increasing only the saving threshold from zero to one bit changes likelihood existence from 96.7% to 42.6% and correct-answer existence from 42.6% to 11.5%. Table D5 gives the complete grid. Figure D7 plots the $\delta = 0$ slice with pointwise 95% whole-case bootstrap bands and the same L, C, and K denominators as Table D5.

Table D5 reports case percentages: L is likelihood existence ($n = 61$), C is reference-hit-correct existence ($n = 61$), and K is correct existence among KEEP-correct cases ($n = 30$). Both blocks require $S_P > \epsilon$ and $\Delta U > \delta$, the right block also requires $\Delta U \geq \Delta U(\text{KEEP})$.

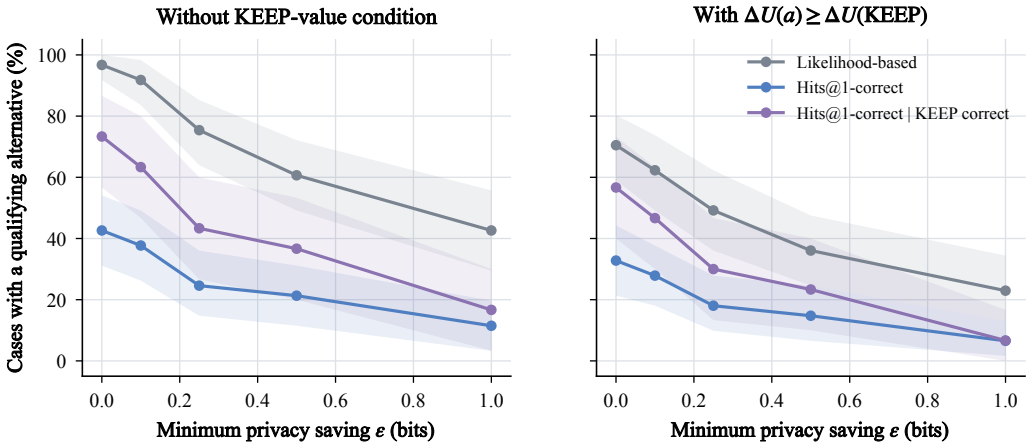

Figure D7: Alternative existence versus privacy-saving threshold ($\delta = 0$; 95% CIs).

Table D5: Intermediate existence under privacy-saving and reader-value margins.

		Margin conditions			Plus KEEP-value condition		
ϵ	δ	L	C	K	L	C	K
0	0	96.7%	42.6%	73.3%	70.5%	32.8%	56.7%
0	0.05	93.4%	41.0%	70.0%	68.9%	31.1%	53.3%
0	0.1	93.4%	41.0%	70.0%	68.9%	31.1%	53.3%
0	0.25	93.4%	41.0%	70.0%	68.9%	31.1%	53.3%
0	0.5	90.2%	39.3%	66.7%	65.6%	29.5%	50.0%
0.1	0	91.8%	37.7%	63.3%	62.3%	27.9%	46.7%
0.1	0.05	88.5%	36.1%	60.0%	60.7%	26.2%	43.3%
0.1	0.1	88.5%	36.1%	60.0%	60.7%	26.2%	43.3%
0.1	0.25	86.9%	34.4%	56.7%	60.7%	24.6%	40.0%
0.1	0.5	85.2%	32.8%	53.3%	59.0%	23.0%	36.7%
0.25	0	75.4%	24.6%	43.3%	49.2%	18.0%	30.0%
0.25	0.05	72.1%	23.0%	40.0%	47.5%	16.4%	26.7%
0.25	0.1	72.1%	23.0%	40.0%	47.5%	16.4%	26.7%
0.25	0.25	72.1%	23.0%	40.0%	47.5%	16.4%	26.7%
0.25	0.5	70.5%	21.3%	36.7%	45.9%	14.8%	23.3%
0.5	0	60.7%	21.3%	36.7%	36.1%	14.8%	23.3%
0.5	0.05	57.4%	19.7%	33.3%	34.4%	13.1%	20.0%
0.5	0.1	57.4%	19.7%	33.3%	34.4%	13.1%	20.0%
0.5	0.25	57.4%	19.7%	33.3%	34.4%	13.1%	20.0%
0.5	0.5	55.7%	18.0%	30.0%	32.8%	11.5%	16.7%
1	0	42.6%	11.5%	16.7%	23.0%	6.6%	6.7%
1	0.05	39.3%	9.8%	13.3%	19.7%	4.9%	3.3%
1	0.1	39.3%	9.8%	13.3%	19.7%	4.9%	3.3%
1	0.25	39.3%	9.8%	13.3%	19.7%	4.9%	3.3%
1	0.5	36.1%	9.8%	13.3%	18.0%	4.9%	3.3%

Table D6: Case-macro outcomes of fixed release rules.

Rule	P	S_P	ΔU	Hit	IE	CI
All eligible KEEP cases						
KEEP	1.0175	0.0000	4.2150	51.4	32.4	0.0
DROP	0.0000	1.0175	0.0000	1.0	99.0	0.0
Canonical-first	1.0267	-0.0092	2.9718	24.8	57.1	2.9
Fewest-triples	0.9797	0.0378	2.9174	23.8	61.9	4.8
Minimum-overlap	0.8843	0.1332	3.3061	38.1	53.3	11.4
Uniform	0.9449	0.0727	2.2242	11.2	74.3	3.0
Disclosure-positive eligible KEEP cases						
KEEP	1.7515	0.0000	4.2570	49.2	31.1	0.0
DROP	0.0000	1.7515	0.0000	0.0	100.0	0.0
Canonical-first	1.6483	0.1032	3.1473	24.6	55.7	4.9
Fewest-triples	1.5645	0.1870	3.0455	23.0	62.3	8.2
Minimum-overlap	1.4091	0.3424	3.1134	34.4	52.5	19.7
Uniform	1.4195	0.3320	2.2921	11.7	69.8	5.2

D.5 Fixed release-rule comparison

The post-hoc comparison uses all 105 cases with an eligible KEEP, including the 44 with $P(\text{KEEP}) = 0$. Each rule receives the same frozen eligible action inventory under H_0/C_{EMPTY}. KEEP and DROP are fixed endpoints. Four structural rules select among transformed actions: canonical-first uses their first-occurrence order in the frozen action list, fewest-triples minimizes the released triple count, minimum-overlap minimizes the number of released triples whose subject or object exactly matches the secret-topic identifier, then minimizes triple count, uniform averages outcomes over a uniformly selected transformed action. Remaining ties use canonical order. With no transformed action a rule would retain KEEP, none of the 105 cases requires this fallback.

Minimum-overlap reads the declared secret-topic identity. None of the four rules reads auditor scores, reader likelihoods, generated answers, or correctness labels when selecting an action. The frozen labels are used for evaluation. Canonical-first follows the materialized action order, which includes family and source ordering. The median number of tied candidates before the canonical tie-break is 99 for fewest-triples and 34 for minimum-overlap.

Table D6 reports case-macro outcomes on the full 105-case population and separately on the 61-case disclosure-positive cohort. Uniform uses exact within-case expected outcomes before averaging across cases. Its rates are expectations, not counts from one random run. Intermediate requires a transformed action with $S_P > 0$ and $\Delta U > 0$, correct Intermediate additionally requires Hits@1. Fixed endpoints therefore have zero correct-Intermediate rates by definition.

In Table D6, cost and saving use bits and ΔU uses nats per reference-answer token. Hit denotes Hits@1, IE the exact prompt-specified insufficient-evidence response, and CI correct Intermediate, rates are case-macro percentages.

On all 105 cases, mean savings relative to KEEP are -0.0092 bits for canonical-first, 0.0378 for fewest-triples, 0.1332 for minimum-overlap, and 0.0727 for uniform. Their paired 95% intervals are [-0.1480, 0.1328], [-0.1053, 0.1783], [-0.0258, 0.3064], and [-0.0720, 0.2189], respectively. Hits@1 differences are -26.7, -27.6, -13.3, and -40.2 percentage points, with intervals [-37.1, -16.2], [-37.1, -17.1], [-21.9, -4.8], and [-49.4, -31.1]. Intervals use 2,000 paired whole-case bootstrap resamples, seed 0, with pointwise coverage. These rules select a single action or a specified distribution, the existence percentages in Table D5 instead inspect every eligible transformed action in each case.

Table D7: Paired release-rule differences from KEEP (95% CIs).

Rule	Paired P difference [95% CI]	Paired Hits@1 difference [95% CI]
R-lik	-1.3735 [$-1.7617, -1.0160$]	-37.7 [$-50.8, -24.6$]
R-nonsentinel	-0.6442 [$-0.9284, -0.4189$]	-16.4 [$-27.9, -4.9$]

D.6 Explicit selection rules

The two full-label offline rules were frozen after inspecting existing outcomes and numerical checks, as a post-hoc descriptive analysis. Both read every candidate's canonical complete-output P and reader ΔU, which requires the reference answer, R-nonsentinel also reads generated text. Hits@1, strict EM, and support are excluded from selection. On the 61-case decision cohort under H_0/C_{EMPTY}, candidates are admissible MINIMIZE or REPLACE actions with evaluable coordinates satisfying $P(\mathrm{KEEP}) - P(a) > 10^{-9}$ bits and $\Delta U(a) > 10^{-9}$ nats per reference-answer token. R-nonsentinel additionally excludes answers whose text, after `strip().casefold()`, equals `insufficient_evidence`. These are numerical tolerances. The strict Intermediate definition is unchanged (Appendix A.4). For each rule, the tolerance tie set contains qualifying actions with $P(a) - P_{\min} \leq 10^{-9}$ bits, where $P_{\min}$ is their exact stored-float minimum. Selection takes the first member in the frozen action list's first-occurrence order over unique action identifiers, with UTF-8 identifier byte order resolving any remaining rank tie. An empty candidate set returns KEEP.

Table D8 includes fallbacks and reports P in bits, with all rates as percentages of the 61 cases. Structural rows repeat the same-cohort comparison above, their strict EM is unavailable. Uniform reports expected rates. Paired differences are rule minus KEEP, with pointwise 95% case-bootstrap intervals from 2,000 resamples, seed 0, Hits@1 differences use percentage points. R-lik and R-nonsentinel select transformed actions in 96.7% and 63.9% of cases, respectively. All 14 insufficient-evidence responses selected by R-nonsentinel come from KEEP fallbacks. Among its 39 non-fallback selections, 35.9% are reference hits, covering 53.8% of the 26 cases admitting such an alternative, the overall 32.8% correct-answer rate include correct KEEP fallbacks. R-lik returns the prompt-specified insufficient-evidence response in 73.8% of cases, alongside a paired P difference of -1.37 bits [$-1.76, -1.02$]. Among non-fallback cases, the median tie-set sizes are 3 for R-lik and 1 for R-nonsentinel, with multiple members in 62.7% and 43.6% of cases. Exact and tolerance tie sets have the same medians and multiple-member counts, although the tolerance window enlarges two R-lik sets from 76 to 79 and from 21 to 22 actions.

Table D8: Offline selection outcomes, including fallbacks, on the decision cohort.

Release	Outcome-label access	Mean P	Hits@1	Strict EM	Sentinel	Fallback	Indep. eval.
KEEP	-	1.7515	49.2%	19.7%	31.1	0.0%	Not evaluated
DROP	-	0.0000	0.0%	0.0%	100.0	0.0%	Not evaluated
R-lik	All $P, \Delta U$	0.3780	11.5%	3.3%	73.8	3.3%	Not evaluated
R-nonsentinel	All $P, \Delta U$; generated text	1.1073	32.8%	11.5%	23.0	36.1%	Not evaluated
Structural rules (Appendix D.5, same cohort)							
Frozen order	None	1.6483	24.6%	Unavailable	55.7	0.0%	-
Triple count	None	1.5645	23.0%	Unavailable	62.3	0.0%	-
Minimum overlap	None	1.4091	34.4%	Unavailable	52.5	0.0%	-
Uniform random (expected)	None	1.4195	11.7%	Unavailable	69.8	0.0%	-

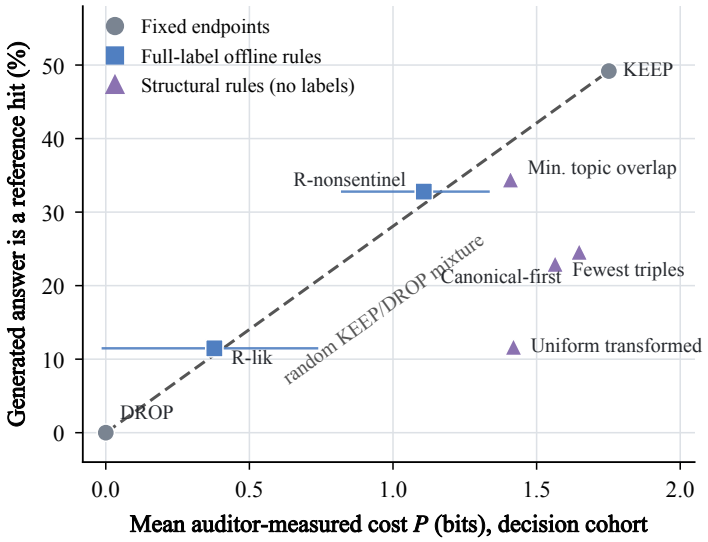

Figure D8: Cost versus reference hits.

Endpoint-mixture reference at matched mean Hits@1. As an algebraic comparison on these same 61 cases, choose KEEP with a case-independent probability r and DROP otherwise. DROP has zero measured cost and no reference hits in this cohort, so expected mean cost and Hits@1 are r times their KEEP values. Setting $r = 2/3$ matches R-nonsentinel's 20/61 reference-hit rate and gives 1.1677 bits, compared with its 1.1073 bits: a 0.0604-bit difference. Setting $r = 7/30$ matches R-lik's 7/61 rate and gives 0.4087 bits, compared with 0.3780 bits: a 0.0307-bit difference. These values use the unrounded frozen endpoint means, without additional model calls. The mixture probabilities match observed aggregate rates retrospectively, they supply a descriptive reference, not a selected operating policy. No paired interval for either difference was estimated, and matching mean Hits@1 does not enforce each case's likelihood constraint.

Figure D8 compares release rules with the KEEP/DROP mixture. Horizontal 95% intervals show paired cost differences from KEEP, shifted by its mean. Structural rules are point estimates.

D.7 Privacy and reader value

The eligibility-consistent true-value privacy-utility correlation is 0.1379 [0.0844, 0.1900] over 107 evaluable cases. Ranking inconsistency is 44.6% [41.8%, 47.3%] over 105 cases, conditioning on eligible pairs with non-tied privacy and utility coordinates. Correlation requires nonconstant coordinates, the pair statistic requires comparable pairs. The original scored-inventory reference gives correlation 0.1323 [0.0799, 0.1839], before excluding inadmissible KEEP rows. Figure D9 plots eligible action rows, with cost in bits and reader value in nats per reference-answer token. Color counts actions per hexagonal bin on a logarithmic scale, each cumulative curve weights actions equally within its family. The annotated correlation averages evaluable within-case correlations.

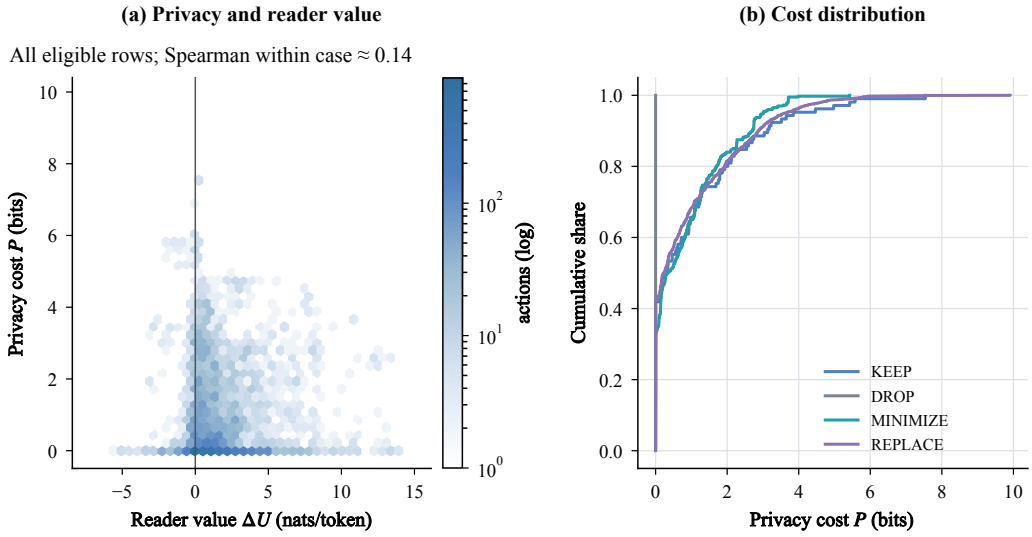

Figure D9: Privacy cost and reader value across eligible actions.

E RQ3: Direct-Scoring Cost and Fidelity

E.1 Ranking and Decision-Set Fidelity

Approximation fidelity here concerns coordinate rankings and threshold-free Intermediate and Pareto sets. The study does not report approximation errors in attainability or minimum cost at a prescribed reader-value threshold u. Preserving ranks or these sets need not preserve the fixed-threshold feasible set. Full-universe ranking and Pareto comparisons use each case's complete Truth-admissible BASE-arm release universe: KEEP, DROP, and all admissible MINIMIZE/REPLACE actions, identical across conditions. Privacy ranking fidelity is per-case Spearman correlation between oracle and predicted $P^{(1)}_{\text{true}}$, $P^{(1)}_{\text{MAP}}$ and raw signed dP_true rankings are separate companions. Utility ranking fidelity correlates oracle and predicted dU on the same actions. Spearman uses average ranks, fewer than three actions or a constant coordinate yields an unavailable value, excluded from the macro mean and explicitly counted.

Intermediate recovery uses transformed actions only. Oracle positives satisfy $P_{\text{oracle}}(\text{KEEP}) - P_{\text{oracle}}(a) > 0$ and $\text{dU}_{\text{oracle}}(a) > 0$, predicted positives use the corresponding predicted quantities. Headline recall is macro-averaged over cases with at least one oracle-positive transformed action, companion precision uses cases with at least one predicted-positive action. Their denominators are reported separately. For full-universe methods, Pareto fidelity compares oracle and predicted nondominated action-identity sets on the full admissible universe. Purchased-only instead compares its nondominated set on the purchased sub-universe Q with the full-oracle target F: $J^{Q\to\text{full}}_{\text{Pareto}} = |F_Q \cap F|/|F_Q \cup F|$. Both comparisons are macro-averaged across evaluable cases. Dominance retains the definition of Appendix A.6: no greater privacy cost, no lower reader value, and at least one strict inequality. Uncertainty in the direct-scoring approximation/acquisition study uses grouped bootstrap over cases, with each statistic's evaluable and unavailable counts reported, action rows are never independent replicates.

Table E1: DEV construct probes and paired contrasts.

Probe	Spearman	Concordance	Paired Spearman gain vs Global [95% CI]
Local-BGE	0.6408	0.7488	+0.5905 [0.4986, 0.6868]
Global-BGE	0.0504	0.5178	-
SecretCos	0.2012	0.5726	+0.1508 [0.1094, 0.1952]
SecretHadamard	0.1661	0.5591	+0.1158 [0.0462, 0.1817]
Exact-overlap cue	0.2631	0.7115	-
Local-Mask	-0.1292	-	-

E.2 DEV CONSTRUCT PROBES

All probes evaluate raw signed dP_{true} on the same 51 DEV cases. Table E1 separates Spearman from mixed/non-tied pairwise concordance. Concordance excludes predictor and target ties, the binary overlap cue has 94,105 eligible pairs, versus 265,354 for global probes and 265,363 for Local-BGE.

Local-BGE protocol. Local-BGE evaluates held-out labels within a fixed action universe for each learned-study DEV-51 case, using that case's complete Truth-admissible BASE logical action universe (5,305 action rows across DEV). Five-fold within-case out-of-fold prediction groups actions by canonical released state, the same released state never crosses the fit/evaluation boundary within an outer fold. Each action is evaluated only by the model whose outer-training portion excludes that action and its entire released-state group. The frozen BGE-derived feature vector has 2,307 dimensions: released-, removed-, and added-triple means, query, history-turn mean, co-evidence-triple mean, and three log-count features. The features do not explicitly represent the protected-secret proposition. Fitting separately per case permits label adaptation to its fixed case/secret. Its frozen BGE atoms are already L2-normalized. For each outer fold and target, Ridge selects $\alpha \in \{0.1, 1, 10, 100\}$ using three-fold released-state-grouped inner validation confined to the outer-training data, minimizing pooled inner-validation MSE, an exact tie selects the larger alpha. `StandardScaler` is fitted only on the corresponding inner-training portion during tuning, then refitted with Ridge on the outer-training portion for outer evaluation. Ridge uses `fit_intercept=True`. For each case, the out-of-fold predictions are assembled into a complete vector, from which one case-level Spearman correlation is computed.

Global-BGE protocol. Global-BGE uses the same 2,307-dimensional frozen BGE-derived feature definition, fitting on the TRAIN-200 Truth-admissible BASE action universe and evaluating on DEV-51. TRAIN and DEV share neither case IDs nor parent-question groups. Alpha selection uses three-fold parent-question-grouped cross-validation within TRAIN, with the same alpha grid, pooled validation MSE, and exact-tie rule as Local-BGE, each validation fold uses a scaler fitted only on its training portion. The final `StandardScaler` and Ridge fits use all TRAIN rows only. The selected alpha is 100 for each of $P^{(1)}_{\text{true}}$, dP_{true}, and dU. Ridge explicitly sets only alpha and `fit_intercept=True`, the remaining Ridge settings follow scikit-learn 1.7.2 defaults, also for Local-BGE. DEV labels are used only for evaluation.

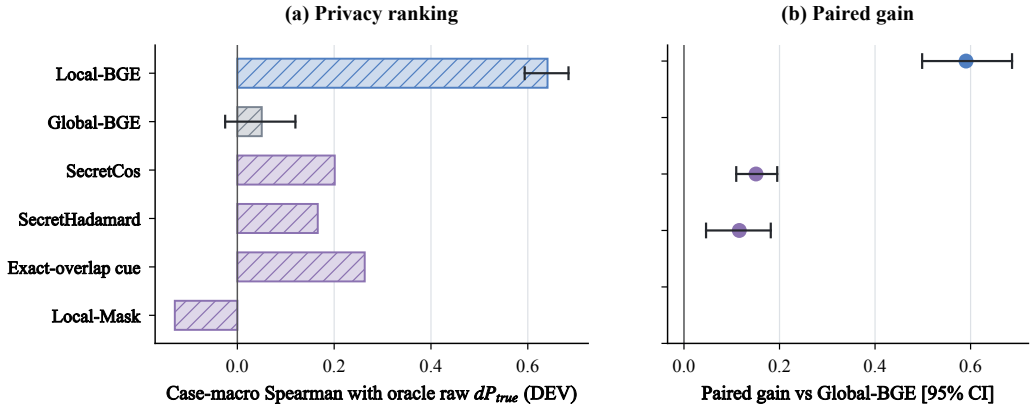

Figure E1: DEV privacy-ranking probes and paired gains over Global-BGE (95% CIs).

Table E2: TRAIN/DEV primary privacy ranking of the frozen global estimators.

Condition	TRAIN privacy Spearman	TRAIN evaluable (%)	DEV privacy Spearman	DEV evaluable (%)
B-Additive	0.049	81.5%	-0.018	86.3%
Pooled-Modular	0.100	20.5%	0.086	27.5%
Pooled-Shared	-0.119	13.0%	0.102	7.8%

Local-Mask control. Local-Mask fits Lasso to binary released-triple indicators using the same outer fitting rows and held-out released-state groups as Local-BGE. Each prediction excludes its own label and all labels in its released-state group. The feature vocabulary includes held-out action definitions, without their labels. Inner fitting uses five-fold released-state-grouped `LassoCV` within the outer-training rows, alpha 0.01 below ten fitting rows, and an intercept for constant targets or absent features, no scaler is used. It predicts dP_{true} and $P^{(1)}_{\text{true}}$, excluding dU. For raw privacy, 66.7% of cases have exactly five distinct out-of-fold predictions, although all 51 case-level Spearman correlations are defined. Intercept-only folds predict their fitting-label means, which differ across folds and can create rankings without a learned mask effect. Local-Mask is therefore reported only as a diagnostic.

Two proposition-conditioned probes retain the Global-BGE TRAIN-200 to DEV-51 split and fitting procedure. The frozen BGE encoder represents the protected fact statement, including its topic, relation, and value content, without the truth label z^*. SecretCos is the primary variant: three cosine similarities between this embedding and the released-, removed-, and added-triple means extend the original 2,307 features to 2,310. An empty set contributes zero cosine. The predeclared secondary SecretHadamard variant appends the three elementwise products, giving 3,459 features. Scaling and three-fold parent-question-grouped alpha selection use TRAIN only, both select Ridge alpha 100. Both variants are reported without selecting between them. Table E1 uses all 51 DEV cases for Spearman and 2,000 whole-case bootstrap resamples with seed 0 for paired differences. Local-Mask has no reported concordance because of the fold-intercept degeneracy described above. SecretCos and SecretHadamard achieve higher case-macro action-ordering Spearman than Global-BGE in the evaluated cross-case setting. Overlap has a higher point estimate than SecretCos, but their paired difference is unresolved: +0.0619 [-0.0187, 0.1441]. The Local-BGE and Global-BGE Spearman intervals are [0.5936, 0.6841] and [-0.0252, 0.1199], the overlap concordance interval is [0.6507, 0.7683]. Local and global probes differ in label access, so their comparison does not isolate a failure mechanism. Figure E1 plots the case-macro Spearman estimates and paired gains reported above. The left panel shows Spearman intervals for Local-BGE and Global-BGE, the other probes are shown as point estimates. The right panel shows paired-gain intervals. All displayed intervals use whole-case bootstrap resampling, Local-Mask retains the diagnostic-only interpretation stated above.

Table E3: Row-weighted TRAIN prediction errors across all scored contexts.

Condition	dP_{true} MAE	dP_{MAP} MAE	dU MAE	Posterior KL
Constant reference	1.4588	0.1571	1.6956	0.3828
B-Additive	1.4271	0.1846	1.3977	0.1227
Pooled-Modular	1.5258	0.1730	0.8321	0.1238
Pooled-Shared	1.4893	0.1726	0.8909	0.1742

E.3 Privacy fitting on TRAIN

The DEV-selected checkpoints fit action-conditional privacy ordering weakly even on TRAIN (Table E2). Ranking uses eligible BASE actions: 21,101 over 200 TRAIN cases and 5,305 over 51 DEV cases. Table E3 instead reports row-weighted errors over all scored TRAIN contexts. Its zero-parameter reference assigns uniform minus/plus posteriors and zero utility, hence zero privacy change. Purchased-mean filling uses the case-specific mean of exact purchased labels for each target. Table E2 reports case-macro Spearman, evaluability percentages use 200 TRAIN and 51 DEV cases. The constant reference has undefined Spearman on both splits, not zero, its absolute errors appear in Table E3. Table E3 averages all 21,895 scored rows in 583 BASE/history/co-evidence contexts. Privacy MAEs use raw signed bits, reader-value MAE uses nats per reference-answer token, and KL is $D_{\text{KL}}(p^{+}_{\text{oracle}} \| p^{+}_{\text{pred}})$. Thus KL measures posterior approximation rather than ground-truth label error, lower values indicate smaller error. Privacy MAE does not improve consistently over the constant reference, whereas utility MAE and oracle-posterior KL are lower for all three models. On eligible non-DROP TRAIN actions, mean within-case standard deviation of raw privacy is 0.0615, 0.0015, and 0.0205 bits for Additive, Modular, and Shared, versus 1.1261 bits for the oracle. Mean within-case zero-cost percentages are 79.0% for Modular and 86.0% for Shared, versus 43.0% for the oracle. The fitted privacy coordinate has little variation under these architectures and training settings.

E.4 TEST ranking and decision sets

Table E4 reports case-macro estimates with 95% case-bootstrap intervals. n marks separately reported evaluable subsets, shared denominators appear in Table B1 and the accompanying text. Privacy and MAP use $P^{(1)}_{\text{true}}$ and $P^{(1)}_{\text{MAP}}$, Oracle rows retain undefined-coordinate exclusions. Purchased-only (A), B-LocalSparse-PU (B), and mean fill (C) share purchases at $K_{\text{fit}} = 64$, with the frontier comparison scopes defined above. Constant coordinates make Spearman unavailable. Modular and Shared TEST privacy rankings are evaluable in 18.4% and 14.3% of cases, respectively. Modular and Shared have many constant predicted primary costs, so their ranking means condition on smaller subsets than LocalSparse, excluded cases are not assigned zero. Intermediate recall uses the 26 cases with oracle-positive transformations, while precision uses cases with predicted positives. Additive predicts no positive case. Pareto fidelity uses all 49 cases. Global privacy ordering is also weak before clipping (Table E5). The raw signed companion in Table E5 reports case-macro dP_{true} Spearman with supplied 95% case-bootstrap intervals.

Purchase and fill. Purchased actions comprise 49 KEEP, 49 DROP, 24 MINIMIZE, and 3,014 REPLACE. All 2,024 unpurchased actions are REPLACE, 99.3% contain a triple whose literal indicator is never active in the fitting set. Purchased-only, learned fill, and mean fill follow the protocol in Appendix B.7. Figure E2 gives empirical cumulative distributions of per-case purchase coverage, with cases weighted equally within each workload. The dashed line is the pooled action-weighted purchase rate, legend values are workload medians.

Case-macro purchase coverage is 60.3% for oracle Intermediate sets (26 cases) and 68.6% for full-oracle Pareto sets (49 cases). Coverage measures target actions purchased, Pareto Jaccard compares nondominated sets and is a different quantity. Intermediate recall/precision for B and C are identical in every evaluable case. Relative to A, both gain 17.4 percentage points [10.1, 25.8] in recall and change by -7.2 percentage points [-12.4, -3.1] in precision over 26 paired cases. B-minus-C Pareto Jaccard is -0.0289 [-0.0731, 0.0000], C is better in 4.1% of cases and the other 95.9% tie ($n = 49$).

Table E4: TEST ranking and decision-set fidelity.

(a) Ranking fidelity and evaluable cases

Condition	Primary privacy Spearman	MAP privacy Spearman	Utility Spearman
B-Additive	0.081 [0.015, 0.143], $n = 39$	0.132 [0.060, 0.212], $n = 39$	0.058 [-0.013, 0.124]
Purchased-only (A)	N/A	N/A	N/A
B-LocalSparse-PU	0.748 [0.720, 0.777], $n = 48$	0.750 [0.706, 0.793], $n = 45$	0.732 [0.710, 0.753]
Same-purchase mean fill (C)	0.749 [0.720, 0.779], $n = 48$	0.744 [0.698, 0.791], $n = 45$	0.723 [0.703, 0.743]
Pooled-Modular	-0.103 [-0.323, 0.123], $n = 9$	0.015 [-0.040, 0.071], $n = 40$	0.161 [0.088, 0.226]
Pooled-Shared	-0.059 [-0.215, 0.121], $n = 7$	-0.004 [-0.065, 0.056], $n = 46$	0.154 [0.082, 0.223]
Oracle	1.000 [1.000, 1.000], $n = 48$	1.000 [1.000, 1.000], $n = 46$	1.000 [1.000, 1.000]

(b) Decision-set fidelity and evaluable cases

Condition	Intermediate recall (%)	Intermediate precision (%)	Pareto Jaccard
B-Additive	0.0 [0.0, 0.0]	N/A	0.079 [0.040, 0.138]
Purchased-only (A)	60.3 [53.1, 67.1]	100.0 [100.0, 100.0]	0.637 [0.531, 0.742]
B-LocalSparse-PU	77.7 [67.4, 87.6]	92.8 [87.6, 96.9]	0.608 [0.505, 0.712]
Same-purchase mean fill (C)	77.7 [67.4, 87.6]	92.8 [87.6, 96.9]	0.637 [0.531, 0.742]
Pooled-Modular	23.0 [7.7, 38.6]	43.2 [18.0, 68.5], $n = 8$	0.109 [0.040, 0.198]
Pooled-Shared	22.1 [7.6, 38.5]	61.3 [34.5, 86.0], $n = 7$	0.010 [0.002, 0.021]
Oracle	100.0 [100.0, 100.0]	100.0 [100.0, 100.0]	1.000 [1.000, 1.000]

On unpurchased actions, B's raw privacy predictions are constant in 91.8% of cases and utility predictions in 89.8%. Primary privacy Spearman is evaluable in only three cases and utility Spearman in five, C is constant and unavailable throughout. Raw privacy, clipped privacy, and utility MAEs for B versus C are respectively 0.8467 versus 0.8443, 0.5538 versus 0.5573, and 0.9867 versus 0.9989. Paired error reductions $C - B$ are -0.0024 [-0.0184, 0.0151], +0.0036 [-0.0088, 0.0206], and +0.0122 [-0.0020, 0.0266] over 49 cases. All include zero.

Full-universe paired B-minus-C primary privacy Spearman is -0.0013 [-0.0100, 0.0051] over 48 cases, MAP gives +0.0061 [-0.0041, 0.0199] over 45, and raw signed privacy gives -0.0015 [-0.0104, 0.0049] over 49. Utility has a small positive difference, +0.0087 [0.0009, 0.0169] over 49. These metrics combine exact purchases with filled values. The primary privacy, unpurchased-action error, and Intermediate comparisons provide little supported gain from learned filling, while the secondary full-universe utility ordering improves slightly.

Table E5: TEST raw signed privacy-ranking fidelity before clipping.

Condition	Raw signed dP_true Spearman
B-Additive	0.073 [0.008, 0.139]
B-LocalSparse-PU	0.738 [0.713, 0.760]
Pooled-Modular	-0.031 [-0.097, 0.037]
Pooled-Shared	0.048 [-0.020, 0.114]
Oracle	1.000 [1.000, 1.000]

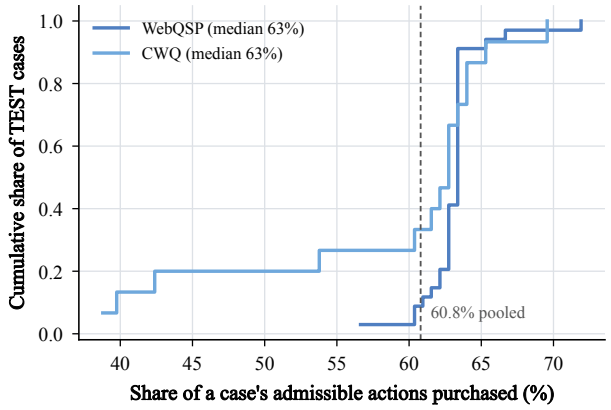

Figure E2: Per-case purchase coverage by workload (% of eligible actions).

E.5 Cost by stage

Candidate handling, artifact-relative Truth filtering, cost aggregation, and mean filling are deterministic operations without model calls. Generation, reference-answer scoring, and domain-value auditor scoring make up the charged model stage. CPU surrogate fitting/inference is measured separately (Table E6). Panel A of Table E6 uses the same admissible BASE universe for direct and purchased calls and expresses relative TEST cost, excluding training-label acquisition. Panel B reports CPU cost with frozen embeddings cached, trainable parameter counts exclude the frozen encoder. The reference covers 5,160 eligible BASE actions in 49 contexts. Each action costs one reader-scoring call, one generation, and two auditor continuations, giving 20,640 calls. Every LocalSparse case purchases 64 labels, giving 3,136 fitted actions. Purchased-only and mean fill use the same purchases. Zero TEST-time calls for global estimators exclude offline training-label acquisition. Figure E3 separates generation, reference-answer scoring, and the two auditor continuations. The TRAIN bar covers the 21,101 eligible BASE actions, at four calls per action, separately from both TEST budgets, these logical calls exclude surrogate computation and do not measure elapsed runtime. Actual frozen full pools define these costs, 135 was an intended proposal prefix, not a universal execution cap. CPU timings cover all 5,354 submitted TEST rows in 144 BASE/history/co-evidence contexts, including inadmissible inventory rows. Frozen embeddings are cached and encoder time is excluded. LocalSparse's 3.931 s covers Lasso fitting and surrogate inference, excluding oracle evaluation. No equivalent Oracle wall time was measured, so these timings do not establish end-to-end speedup. The broader shared-state census in Appendix A.7 is a separate cost object.

Table E6: TEST direct-scoring calls (A) and CPU surrogate cost and capacity (B).

A. TEST-time direct-scoring logical cost

Condition	Reference (%)	TEST calls (%)	TEST saved (%)
B-Additive	100.0%	0.0%	100.0%
B-LocalSparse-PU	100.0%	60.8%	39.2%
Pooled-Modular	100.0%	0.0%	100.0%
Pooled-Shared	100.0%	0.0%	100.0%
Oracle	100.0%	100.0%	0.0%

B. Predictor-only computation

Condition	CPU runtime	Global trainable parameters
B-Additive	0.235 s	105,706
B-LocalSparse-PU	3.931 s	N/A (per-case Lasso fit)
Pooled-Modular	0.263 s	676,771
Pooled-Shared	0.245 s	676,835
Oracle	N/A	N/A

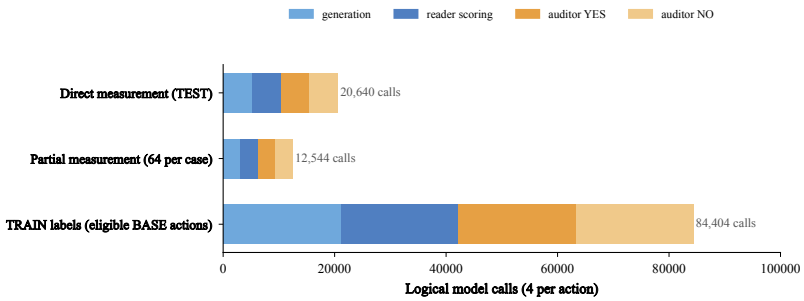

Figure E3: Logical model-call budgets for TEST measurement and TRAIN eligible BASE labels.

Table E7: DEV fidelity and scoring cost by canonical purchase budget.

K	RolePurchased (%)	Logical calls	Learned fill	Mean fill
8	Secondary 7.8	1,632	0.2305	0.2377
16	Primary 15.6	3,264	0.3361	0.3435
32	Primary 31.1	6,528	0.5275	0.5177
64	Primary 62.2	13,056	0.7538	0.7494

E.6 Semantic-disjoint and structural robustness checks

Excluding TEST cases that share a parent question or secret-topic entity with the Pilot leaves 85.7% of TEST-49 (42 cases). LocalSparse retains privacy Spearman 0.766 [0.739, 0.791] over 41 evaluable cases, utility Spearman 0.740 over 42, Intermediate recall/precision 75.9%/93.4% over 22, and Pareto Jaccard 0.636 over 42. A pre-specified hop-by-retrieval-rank stratification showed no qualitative reversal for the LocalSparse hybrid, four of six cells contain fewer than ten cases, so cell-level comparisons are reported in Table S5.

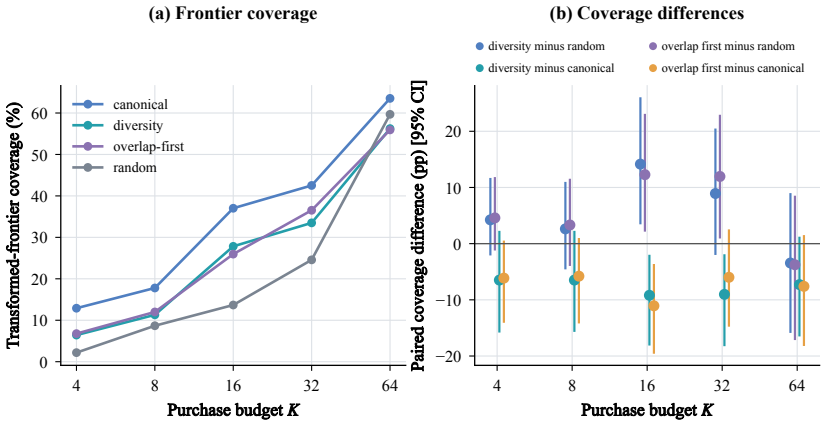

Figure E4: DEV acquisition coverage and paired policy differences (95% CIs).

E.7 Parent-question dependence

Resampling parent-question groups instead of case/variant identifiers changes none of the pre-specified qualitative readings in the DEV probes, TEST approximation, or purchase/fill comparisons. The local-minus-global privacy-ranking difference remains positive, and the small learned-minus-mean-fill utility-ranking exception retains its positive sign. The main tables retain the original case/variant-bootstrap intervals; Table S6 provides the complete dependency-resampling cells.

E.8 Structured acquisition

The extension's DEV selection rule was frozen before acquisition outcomes. All policies purchase KEEP and DROP first. Canonical uses family, route/source, and retrieval-rank order, with MINIMIZE before REPLACE. Random uniformly permutes transformed actions and averages seeds within cases. Overlap-first prioritizes exact-overlap-positive actions with canonical ordering within groups. Diversity uses farthest-first Jaccard distance between released triple sets, breaking ties canonically. Unpurchased values receive per-case purchased means of the clipped privacy cost P and utility ΔU, this comparison uses no learned fill. For the purchased set $Q_{t,K}$, the unpurchased privacy fill is $\widehat{P} = |Q_{t,K}|^{-1} \sum_{a \in Q_{t,K}} [\mathrm{dP}_{\mathrm{true}}(a)]_+$. Purchased actions retain their exact values.

Let F_t^{full} be the full eligible oracle Pareto set and $F_t^T = F_t^{\mathrm{full}} \cap A_{t,\mathrm{transformed}}$, not a frontier recomputed on transformations alone. With purchases $Q_{t,K}$,

$$C_{F,t}^T(K) = \frac{|Q_{t,K} \cap F_t^T|}{|F_t^T|}, \qquad S_F^T = \frac{1}{5} \sum_{K \in \{4,8,16,32,64\}} \mathrm{Mean}_t\, C_{F,t}^T(K).$$

The common cohort comprises 94.1% of DEV cases with nonempty transformed frontiers. Selection requires a positive structured-minus-Random estimate and strictly positive 95% interval lower bound. Neither heuristic passes, no policy was selected and TEST was not replayed. Canonical comparisons are descriptive and do not enter selection. Full acquisition and post-fill secondaries appear in Table S7. Tables E8 and E9 use the common 48-of-51 DEV cohort, the latter gives paired case-bootstrap contrasts in S_F^T. Figure E4 shows budget-specific transformed-frontier coverage and paired differences on the same common cohort. Intervals are pointwise 95% case-bootstrap intervals in percentage points, policy selection uses the budget-averaged statistic in the tables.

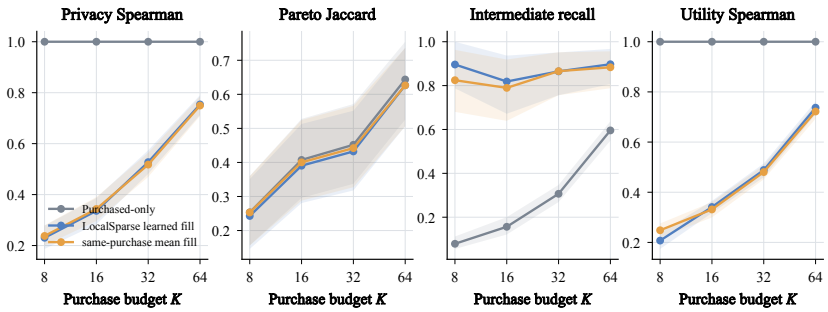

Figure E5: DEV ranking and set fidelity across purchase budgets (95% CIs).

Table E8: DEV transformed-frontier coverage by acquisition policy.

Acquisition policy	Primary S_F^T (%)
Canonical	34.8
Overlap-first	27.4
Diversity	27.1
Random	21.8

Table E9: Paired DEV acquisition-policy contrasts in transformed-frontier coverage.

Contrast	Difference (pp)	95% CI (pp)
Overlap-first minus Random	5.7	[-1.3, 13.0]
Diversity minus Random	5.3	[-2.0, 13.2]
Overlap-first minus Canonical	-7.3	[-13.7, -2.0]
Diversity minus Canonical	-7.7	[-16.5, -0.6]

E.9 Purchase-budget sensitivity

The sweep uses the 51 DEV cases and 5,305 eligible BASE actions under the same canonical order. KEEP and DROP count toward the budget. $K = 16, 32, 64$ are primary settings and $K = 8$ is secondary, at eight purchases every case uses the existing fixed-alpha Lasso rule, $\alpha = 0.01$, because fewer than ten labels are fitted. All 42 MINIMIZE actions are purchased by $K = 8$, leaving only REPLACE unpurchased at every tested budget. The sweep selects no budget, TEST remains at $K = 64$. Mean fill averages the purchased raw responses before clipping: $\widehat{P} = [|Q_{t,K}|^{-1} \sum_{a \in Q_{t,K}} \mathrm{dP}_{\mathrm{true}}(a)]_+$ on unpurchased actions. Utility uses the purchased mean of ΔU, and purchased actions retain their exact values. Structured acquisition (Appendix E.8, Table S7) instead averages clipped privacy costs and reports post-fill metrics on its common 48-case cohort (94.1% of DEV-51). Thus the two tables differ in both privacy-fill definition and cohort. Table E7 compares full-universe predictions with oracle $P_{\mathrm{true}}^{(1)}$. Purchased fractions are computed within cases and then macro-averaged, logical costs charge four calls per purchased action. Privacy Spearman is evaluable in 44, 48, 48, and 49 cases at $K = 8, 16, 32, 64$, respectively, for both fill arms. Learned fill closely tracks mean fill at each budget. As the exact-purchased fraction rises with budget, the full-universe metric combines exact and filled values and does not isolate prediction quality on unpurchased actions. Figure E5 plots the stored case-macro estimates with pointwise 95% case-bootstrap bands. Purchased-only ranking is evaluated on exact purchased values and equals one by construction, learned and mean fill are evaluated over the full eligible universe. Purchased-only set metrics compare the purchased sub-universe with the full oracle.

F SUPPLEMENTARY TABLES

TABLE S1. CONVQUESTIONS SECONDARY DIAGNOSTICS

This section retains the historical Qwen companions and component-deletion analysis. The two-family three-arm comparison is reported in the main paper. Table S1(a) uses the 412-target, 242-seed three-arm cohort. Each row removes one frozen component, retains seed-macro weighting, and resamples the remaining 14 components for the specificity interval. Only specificity has a LOCO interval. Retention requires that interval to exclude zero and its sign to match the native point estimate on the same reduced cohort.

Table S1(a): CQ leave-one-component-out specificity sensitivity.

Deleted	Removed seeds	Removed targets	Δ_{spec}, bits	95% CI	Retained
DG-0000	44	72	-0.805789	[-1.127985, -0.460816]	Yes
DG-0001	40	76	-0.755826	[-1.094839, -0.409142]	Yes
DG-0002	41	69	-0.797314	[-1.125881, -0.441001]	Yes
DG-0003	22	38	-0.661695	[-0.807891, -0.392392]	Yes
DG-0004	58	112	-0.732471	[-1.113222, -0.394333]	Yes
DG-0005	2	2	-0.780957	[-1.066992, -0.526216]	Yes
DG-0006	5	7	-0.783515	[-1.071762, -0.525562]	Yes
DG-0007	3	3	-0.769843	[-1.056831, -0.498320]	Yes
DG-0008	11	15	-0.784779	[-1.087547, -0.509740]	Yes
DG-0009	2	2	-0.755753	[-1.033701, -0.479716]	Yes
DG-0010	2	2	-0.759146	[-1.040289, -0.482322]	Yes
DG-0011	2	2	-0.777079	[-1.067330, -0.506135]	Yes
DG-0012	6	8	-0.805193	[-1.108529, -0.566516]	Yes
DG-0013	2	2	-0.754411	[-1.024106, -0.468281]	Yes
DG-0014	2	2	-0.792875	[-1.085274, -0.533028]	Yes

Table S1(b): CQ seed-macro MAP companion contrasts.

Contrast	Targets / seeds	Seed-macro mean, bits	Pooled mean, descriptive	Target signs $+/0/-$	Evaluable / unavailable
$\Delta^{\text{all}}_{\text{native}}$	453 / 254	+0.083	+0.075	262 / 0 / 191	453 / 0
$\Delta^{\text{3arm}}_{\text{native}}$	412 / 242	+0.080	+0.081	239 / 0 / 173	412 / 0
Δ_{irrel}	412 / 242	-0.049	-0.044	198 / 0 / 214	412 / 0
Δ_{spec}	412 / 242	+0.129	+0.126	264 / 1 / 147	412 / 0

MAP companion contrasts. These non-headline dP_{MAP} contrasts use the same seed-conversation macro estimator. Pooled target means and target sign counts are descriptive companions only; no companion bootstrap or interval is computed, and these quantities do not enter the specificity criterion. Evaluable/unavailable counts below refer to each row's stated cohort. The MAP companion moves opposite to the primary signed true-value contrast ($\Delta_{\text{spec}} = +0.129$ versus -0.768 bits): the MAP-mass ratio and true-value log-likelihood ratio are different quantities, and need not agree in sign.

Absolute exposure and within-arm separation. The following non-headline companions are seed-macro means, not pooled target means; no companion uncertainty interval is computed. H_0 and H_{native} cover all admitted targets, whereas H_{irrel} covers the donor-matched cohort.

Table S1(c): CQ seed-macro absolute-exposure arm means.

History	Targets	Seeds	Mean $p(z^*)$, KEEP	Mean $p(z^*)$, DROP	Mean $d(H)$, bits
H_0	453	254	0.480	0.143	3.017
H_{native}	453	254	0.526	0.188	2.541
H_{irrel}	412	242	0.418	0.107	3.338

Every arm shows a large positive KEEP-over-DROP separation; the reported Δ contrasts are differences between these levels on their specified matched cohorts, so a negative Δ is a reduction of an already-large separation, not a reversal. Absolute true-value posterior is descriptively higher under native history for both KEEP and DROP while its mean marginal response is smaller, distinguishing absolute exposure from marginal disclosure. Because the displayed arm means have different cohort coverage, their difference is not a substitute for the primary Δ_{spec}, which is the per-target seed-macro contrast on the 412-target three-arm cohort.

TABLE S2. DETAILED CONTROLLED CALIBRATION

Tables S2(a)–S2(d) report all 16 exposure/register/class cells of 30 items (Table C23). NL and dotted denote natural-language relations and dotted paths. p_* is constructed-true-class probability. NLL uses nats. Single-class ECE_{true} is unavailable, not zero. ECE_{MAP} is defined.

Table S2(a): GLM controlled responses by exposure, register, and class.

Exposure / register	True class	MAP accuracy (%)	Mean p_*	Median p_*	Mean MAP confidence
E0 / NL	YES	0.0%	0.033	0.019	0.967
E0 / NL	NO	100.0%	0.976	0.984	0.976
E0 / dotted	YES	0.0%	0.013	0.013	0.987
E0 / dotted	NO	100.0%	0.989	0.990	0.989
E1 / NL	YES	0.0%	0.015	0.014	0.985
E1 / NL	NO	100.0%	0.979	0.982	0.979
E1 / dotted	YES	0.0%	0.012	0.011	0.988
E1 / dotted	NO	100.0%	0.988	0.988	0.988
E2 / NL	YES	13.3%	0.182	0.085	0.872
E2 / NL	NO	100.0%	0.866	0.943	0.866
E2 / dotted	YES	0.0%	0.033	0.031	0.967
E2 / dotted	NO	100.0%	0.969	0.971	0.969
E3 / NL	YES	66.7%	0.569	0.577	0.662
E3 / NL	NO	100.0%	0.941	0.950	0.941
E3 / dotted	YES	3.3%	0.242	0.223	0.768
E3 / dotted	NO	100.0%	0.971	0.971	0.971

Table S2(b): GLM controlled-response error diagnostics.

Exposure / register	True class	NLL	Brier	$\mathrm{ECE}_{\mathrm{MAP}}$
E0 / NL	YES	3.905	0.936	0.967
E0 / NL	NO	0.024	0.001	0.024
E0 / dotted	YES	4.459	0.974	0.987
E0 / dotted	NO	0.011	0.000	0.011
E1 / NL	YES	4.321	0.970	0.985
E1 / NL	NO	0.021	0.001	0.021
E1 / dotted	YES	4.494	0.976	0.988
E1 / dotted	NO	0.012	0.000	0.012
E2 / NL	YES	2.377	0.722	0.791
E2 / NL	NO	0.158	0.036	0.134
E2 / dotted	YES	3.597	0.935	0.967
E2 / dotted	NO	0.032	0.001	0.031
E3 / NL	YES	0.652	0.220	0.255
E3 / NL	NO	0.062	0.005	0.059
E3 / dotted	YES	1.602	0.593	0.734
E3 / dotted	NO	0.029	0.001	0.029

Table S2(c): Gemma controlled responses by exposure, register, and class.

Exposure / register	True class	MAP accuracy(%)	Mean p_*	Median p_*	Mean MAP confidence
E0 / NL	YES	0.0%	0.000	0.000	1.000
E0 / NL	NO	100.0%	1.000	1.000	1.000
E0 / dotted	YES	0.0%	0.000	0.000	1.000
E0 / dotted	NO	100.0%	1.000	1.000	1.000
E1 / NL	YES	0.0%	0.000	0.000	1.000
E1 / NL	NO	100.0%	1.000	1.000	1.000
E1 / dotted	YES	0.0%	0.000	0.000	1.000
E1 / dotted	NO	100.0%	1.000	1.000	1.000
E2 / NL	YES	26.7%	0.265	0.000	0.993
E2 / NL	NO	70.0%	0.702	1.000	0.968
E2 / dotted	YES	0.0%	0.000	0.000	1.000
E2 / dotted	NO	100.0%	1.000	1.000	1.000
E3 / NL	YES	70.0%	0.728	0.996	0.958
E3 / NL	NO	96.7%	0.951	1.000	0.984
E3 / dotted	YES	90.0%	0.869	1.000	0.923
E3 / dotted	NO	100.0%	1.000	1.000	1.000

Table S2(d): Gemma controlled-response error diagnostics.

Exposure / register	True class	NLL	Brier	$\mathrm{ECE}_{\mathrm{MAP}}$
E0 / NL	YES	19.029	1.000	1.000
E0 / NL	NO	0.000	0.000	0.000
E0 / dotted	YES	19.600	1.000	1.000
E0 / dotted	NO	0.000	0.000	0.000
E1 / NL	YES	17.625	1.000	1.000
E1 / NL	NO	0.000	0.000	0.000
E1 / dotted	YES	18.767	1.000	1.000
E1 / dotted	NO	0.000	0.000	0.000
E2 / NL	YES	7.920	0.728	0.726
E2 / NL	NO	1.695	0.274	0.268
E2 / dotted	YES	16.225	1.000	1.000
E2 / dotted	NO	0.000	0.000	0.000
E3 / NL	YES	0.991	0.240	0.259
E3 / NL	NO	0.251	0.037	0.049
E3 / dotted	YES	0.262	0.081	0.040
E3 / dotted	NO	0.000	0.000	0.000

Table S2(e): Selected controlled success and failure examples.

Cell	Selected exposure clause	Truth / MAP	MAP confidence	p_*
E2 correct	`ctrl.accreditation.accredited_body.body_that_annulled_the_accreditation`	NO / NO	0.9981	0.9981
E2 wrong	`counterparty whose supply compact was voided`	NO / YES	0.9876	0.0124
E3 correct	`ctrl.permits.permit_holder.permit_issuing_authority`	YES / YES	0.9999	0.9999
E3 wrong	`counterparty recorded as bound by no supply compact`	NO / YES	0.8176	0.1824

Table S2(f): Four controlled probes in the selected topic-relation group.

Exposure	Probe value	Truth	MAP	MAP confidence	p_*
E0	Simvraer Meivgusk	YES	NO	0.9890	0.0110
E1	Binslirn Drodral	YES	NO	0.9890	0.0110
E2	Sithtiar Keithzaesk	NO	YES	0.9876	0.0124
E3	Geskvraell Pamgraell	NO	YES	0.7549	0.2451

TABLE S3. SELECTED RESPONSE MODES

Examples are selected deterministically within predeclared success/failure cells whose membership is outcome-defined. They illustrate response modes, not prevalence. Control items use frozen item order; transformed actions use frozen case and action order. R/W denote correct/incorrect MAP alignment before and after the action. The WebQTest-859 rows below are distinct actions from the WebQTest-859 matched-output example discussed in the paper. Table S2(e) excerpts relation clauses rather than complete auditor views. E2/E3 dotted rows use the dotted register; both supply-compact rows use plain NL. Table S2(f) includes all four probes from the selected E2-failure item's frozen topic-relation group, `CC-P-COMP-11`: Noundraeth Skumtoull and supply compact counterparty. Each row has its own value and construction truth label, so these are distinct propositions, not longitudinal views of one proposition. Inclusion does not depend on their outcomes. Table S3(a)(a) pairs DROP (minus) with ACTION (plus), reporting dP_{true} in bits, Δp_* on the probability scale,

and dU in nats per reference-answer token. Panel (b) gives released triples and generated text; all four DROP generations are `INSUFFICIENT_EVIDENCE`.

Table S3(a): Selected transformed-action responses and MAP transitions.

(a) Auditor and reader coordinates.

Cell	Case; action	MAP minus / plus	p_*^-/p_*^+	dP_{true} (bits)	Δp_*	dU
RR	WebQTest-461; REPLACE	YES / YES	0.9707 / 0.9047	-0.1016	-0.0660	+1.3751
RW	WebQTest-461; MINIMIZE	YES / NO	0.9707 / 0.4378	-1.1487	-0.5329	+1.3425
WR	WebQTest-859; MINIMIZE	NO / YES	0.0851 / 0.5312	+2.6421	+0.4461	+1.0424
WW	WebQTest-859; REPLACE	NO / NO	0.0851 / 0.2942	+1.7897	+0.2091	+1.0619

(b) Released triple and generated text.

Cell	Released triple	Generated text under ACTION
RR	Clay Matthews III - `people.person.education` - m.0n0g06t	`INSUFFICIENT_EVIDENCE`
RW	m.0hprg7d - `education.education.institution` - University of Southern California	University of Southern California
WR	m.06jmmyb - `location.religion_percentage.religion` - Russian Orthodox Church	The predominant religion in Russia is the Russian Orthodox Church.
WW	m.06jmmyk - `location.religion_percentage.religion` - Islam	The predominant religion in Russia is Islam.

TABLE S4. SUPPORT AND CLOSED-BOOK CROSS-TABULATION

Table S4(a) counts the 8,493 action rows with privacy coordinates, including KEEP/DROP anchors; these are neither Intermediate counts nor case prevalences. The source column `CB_r` is denoted b_r here to distinguish binary correctness from continuous Pure-CB $CB_r = m_r(y \mid q, \varnothing)$. The binary value uses exact correctness of the H_0/C_{EMPTY} DROP generation. Its input prefix was verified to match the Pure-CB scoring prefix, while the reader states remain distinct.

Table S4(a): Action-row coverage by closed-book correctness and support status.

Binary diagnostic	$\text{Supports}_{\text{set}} = 1$	$\text{Supports}_{\text{set}} = 0$	$\text{Supports}_{\text{set}} = \text{NOT_EVALUABLE}$
$b_r = 0$, headline diagnostic stratum	25	6	8,356
$b_r = 1$, companion diagnostic stratum	0	0	106

TABLE S5. HOP AND RETRIEVAL-RANK ROBUSTNESS

The pre-specified strata combine hop count (H1 or H2+) and frozen retrieval-rank terciles (T1-T3). Four of six cells contain fewer than ten cases; cell-level comparisons remain descriptive.
H1 denotes one hop and H2+ at least two; T1–T3 use frozen retrieval-rank cut points. Tables S5(a)–S5(c) report case-macro Spearman and supplied 95% case-bootstrap intervals. Table S5(a) gives case totals and eligible BASE action counts, including KEEP and DROP; † marks strata with fewer than ten cases. Privacy ranking in H1 × T1 uses 18 evaluable cases; its utility ranking and both rankings in the other strata use all cases. Brackets give 95% case-bootstrap intervals; n, where shown, gives the metric-specific evaluable count. For $n = 1$, only the point is shown because resampling cannot give informative uncertainty. N/A values remain unavailable, not zero.

Table S5(a): Coverage and LocalSparse ranking by hop count and retrieval rank.

Frozen cell	Cases	BASE actions	Privacy Spearman [95% CI]	Utility Spearman [95% CI]
H1 × T1	19	1,935	0.767 [0.721, 0.814]	0.721 [0.684, 0.758]
H1 × T2	11	1,111	0.762 [0.719, 0.803]	0.777 [0.753, 0.799]
H1 × T3	6†	620	0.742 [0.691, 0.782]	0.716 [0.679, 0.755]
H2+ × T1	2†	190	0.826 [0.747, 0.906]	0.791 [0.782, 0.801]
H2+ × T2	3†	304	0.764 [0.682, 0.884]	0.785 [0.760, 0.825]
H2+ × T3	8†	1,000	0.665 [0.591, 0.739]	0.674 [0.615, 0.742]

Table S5(b): Global-estimator privacy ranking by hop count and retrieval rank.

Frozen cell	B-Additive	Pooled-Modular	Pooled-Shared
H1 × T1	0.041 [-0.065, 0.151]; $n = 14$	-0.092 [-0.280, 0.096]; $n = 2$	-0.120 [-0.274, 0.023]; $n = 3$
H1 × T2	-0.018 [-0.173, 0.114]; $n = 7$	-0.200 [-0.338, -0.062]; $n = 2$	0.087 [-0.200, 0.375]; $n = 2$
H1 × T3	0.173 [0.027, 0.292]; $n = 6$	0.018 [-0.626, 0.626]; $n = 3$	-0.115 [-0.346, 0.117]; $n = 2$
H2+ × T1	0.252 [0.131, 0.373]; $n = 2$	N/A	N/A
H2+ × T2	0.027 [-0.116, 0.152]; $n = 3$	0.041; $n = 1$	N/A
H2+ × T3	0.153 [0.016, 0.322]; $n = 7$	-0.441; $n = 1$	N/A

Table S5(c): Global-estimator utility ranking by hop count and retrieval rank.

Frozen cell	B-Additive	Pooled-Modular	Pooled-Shared
H1 × T1	0.071 [-0.055, 0.195]	0.160 [0.051, 0.275]	0.157 [0.048, 0.270]
H1 × T2	0.127 [-0.001, 0.248]	0.182 [0.040, 0.313]	0.165 [0.010, 0.318]
H1 × T3	0.072 [-0.015, 0.171]	0.166 [0.003, 0.342]	0.219 [0.036, 0.405]
H2+ × T1	-0.116 [-0.251, 0.020]	-0.010 [-0.225, 0.204]	-0.084 [-0.240, 0.072]
H2+ × T2	0.117 [-0.028, 0.396]	0.324 [0.256, 0.385]	0.311 [0.237, 0.401]
H2+ × T3	-0.060 [-0.218, 0.096]	0.109 [-0.066, 0.281]	0.083 [-0.066, 0.218]

TABLE S6. PARENT-QUESTION DEPENDENCY RESAMPLING

The split groups related parent questions. These tables compare the original case/variant-bootstrap uncertainty with post-hoc parent-question-group resampling, without changing point estimates or evaluability. Local-Mask remains diagnostic because fold-specific intercepts can create rankings without a learned mask effect.

Across Tables S6(a)–S6(d), n denotes the evaluable or paired case count and G the represented parent-question groups. Table S6(a) reports raw signed privacy Spearman and its paired local-minus-global contrast; all rows use $n = 51$ and $G = 46$.

Table S6(a): DEV probe uncertainty under case and parent-question resampling.

Condition / contrast	Point	Case 95% CI	Parent-group 95% CI
Local-BGE	0.640813	[0.593581,0.684145]	[0.593072,0.685254]
Global-BGE	0.050356	[-0.025171,0.119876]	[-0.028435,0.127781]
Local-Mask	-0.129241	[-0.149782,-0.109058]	[-0.149025,-0.110306]
Local-BGE minus Global-BGE	0.590457	[0.498602,0.686752]	[0.494971,0.689297]

Tables S6(b) and S6(c) report coordinate rankings and decision-set fidelity for the four TEST estimators. Utility ranking and Pareto Jaccard share $n/G = 49/47$; Intermediate recall and LocalSparse precision share $26/26$. Other evaluable subsets are listed beside their estimates. LocalSparse denotes the frozen-action-mask B-LocalSparse-PU with $K_{\text{fit}} = 64$; the dagger marks B-Additive Intermediate precision as not applicable, not zero.

Table S6(b): TEST ranking uncertainty under case and parent-question resampling.

Condition	n/G	Point	Case 95% CI	Parent-group 95% CI
Primary privacy ranking				
B-Additive	39/37	0.080780	[0.015406,0.142964]	[0.016005,0.147029]
LocalSparse	48/46	0.748169	[0.719810,0.776591]	[0.713269,0.780637]
Pooled-Modular	9/9	-0.103402	[-0.323341,0.123000]	[-0.323341,0.123000]
Pooled-Shared	7/7	-0.059339	[-0.214913,0.121061]	[-0.214913,0.121061]
MAP privacy ranking				
B-Additive	39/38	0.131836	[0.059714,0.212196]	[0.059284,0.212119]
LocalSparse	45/43	0.750495	[0.706112,0.793195]	[0.705437,0.794671]
Pooled-Modular	40/38	0.014740	[-0.040402,0.071235]	[-0.047149,0.071972]
Pooled-Shared	46/44	-0.004376	[-0.064798,0.055729]	[-0.066667,0.057186]
Utility ranking				
B-Additive		0.057528	[-0.012615,0.124312]	[-0.014079,0.132028]
LocalSparse		0.731950	[0.709590,0.752964]	[0.708487,0.756646]
Pooled-Modular		0.160635	[0.087522,0.226309]	[0.089113,0.233891]
Pooled-Shared		0.153939	[0.081946,0.222658]	[0.082520,0.222609]

Table S6(c): TEST decision-set uncertainty under case and parent-question resampling.

Condition	n/G	Point	Case 95% CI	Parent-group 95% CI
Intermediate recall (%)				
B-Additive		0.0	[0.0,0.0]	[0.0,0.0]
LocalSparse		77.7	[67.4,87.6]	[67.4,87.6]
Pooled-Modular		23.0	[7.7,38.6]	[7.7,38.6]
Pooled-Shared		22.1	[7.6,38.5]	[7.6,38.5]
Intermediate precision (%)				
B-Additive[†]		N/A	N/A	N/A
LocalSparse		92.8	[87.6,96.9]	[87.6,96.9]
Pooled-Modular	8/8	43.2	[18.0,68.5]	[18.0,68.5]
Pooled-Shared	7/7	61.3	[34.5,86.0]	[34.5,86.0]
Pareto Jaccard				
B-Additive		0.078971	[0.039516,0.138433]	[0.037728,0.130686]
LocalSparse		0.607836	[0.504731,0.712342]	[0.503123,0.717354]
Pooled-Modular		0.108617	[0.039566,0.198192]	[0.039319,0.194211]
Pooled-Shared		0.010139	[0.002041,0.020926]	[0.002041,0.020601]

Table S6(d) reuses the original purchases for purchased-only A, learned fill B, and mean fill C. MAE contrasts use unpurchased actions and $C - B$, so positive values favor B; rank and decision-set contrasts retain their stated orientation. Purchased-only A and mean fill C have identical Pareto sets; learned fill B and mean fill C have identical Intermediate sets in every evaluated case.

Table S6(d): TEST purchase/fill uncertainty under case and parent-question resampling.

Condition / contrast	Metric	n	G	Point	Original 95% CI	Parent-group 95% CI
Purchased-only (A)	Pareto Jaccard	49	47	0.636748	[0.531215, 0.742450]	[0.527510, 0.749878]
Mean fill (C)	Pareto Jaccard	49	47	0.636748	[0.531215, 0.742450]	[0.527510, 0.749878]
$C - B$	MAE dP_{true}	49	47	-0.002365	[-0.018389, 0.015108]	[-0.018075, 0.015422]
$C - B$	MAE $P^{(1)}_{\mathrm{true}}$	49	47	0.003562	[-0.008808, 0.020587]	[-0.008773, 0.020497]
$C - B$	MAE dU	49	47	0.012177	[-0.002023, 0.026594]	[-0.002141, 0.028304]
$B - C$	$\rho(P^{(1)}_{\mathrm{true}})$	48	46	-0.001283	[-0.009993, 0.005150]	[-0.010526, 0.005036]
$B - C$	$\rho(P^{(1)}_{\mathrm{MAP}})$	45	43	0.006143	[-0.004061, 0.019880]	[-0.004186, 0.019955]
$B - C$	$\rho(dP_{\mathrm{true}})$	49	47	-0.001450	[-0.010398, 0.004867]	[-0.010315, 0.005097]
$B - C$	$\rho(dU)$	49	47	0.008735	[0.000869, 0.016876]	[0.000640, 0.017099]
$B - C$	Intermediate recall (pp)	26	26	0.0	[0.0, 0.0]	[0.0, 0.0]
$B - C$	Intermediate precision (pp)	26	26	0.0	[0.0, 0.0]	[0.0, 0.0]
$B - C$	Pareto Jaccard	49	47	-0.028912	[-0.073129, 0.000000]	[-0.073129, 0.000000]

Continued on next page

Table S6(d) (continued)

Condition / contrast	Metric	n	G	Point	Original 95% CI	Parent-group 95% CI
$(B$ or $C) - A$	Intermediate recall (pp)	26	26	17.4	[10.1, 25.8]	[10.1, 25.8]
$(B$ or $C) - A$	Intermediate precision (pp)	26	26	-7.2	[-12.4, -3.1]	[-12.4, -3.1]

TABLE S7. STRUCTURED-ACQUISITION SECONDARY RESULTS

The following secondary metrics use the common 48-case DEV cohort (94.1% of DEV-51) and do not enter policy selection. Neither structured heuristic passed the frozen Random-comparison criterion, and no TEST replay followed. Random averages seeds within each case.
Tables S7(a) and S7(b) cover all four policies and five budgets K; cells give case-macro estimates. Parentheses give evaluable / unavailable cases, with repeated fixed counts shown only at their first occurrence in each column. Coverage (%) and post-fill fidelity are distinct.

Table S7(a): DEV acquisition coverage across policies and purchase budgets.

Policy	K	Transformed-frontier coverage C_F^T	Full-frontier purchase coverage C_F	Intermediate purchase coverage C_I
Canonical	4	12.9 (48 / 0)	17.5 (48 / 0)	3.2 (27 / 21)
Canonical	8	17.8	22.2	7.9
Canonical	16	37.0	39.6	15.6
Canonical	32	42.5	45.0	30.7
Canonical	64	63.5	65.1	59.6
Random	4	2.2	8.9	2.5
Random	8	8.7	14.7	6.4
Random	16	13.7	19.6	14.7
Random	32	24.6	29.6	30.9
Random	64	59.7	61.8	63.1
Overlap-first	4	6.8	13.5	3.2
Overlap-first	8	12.0	17.2	7.7
Overlap-first	16	26.0	28.9	15.2
Overlap-first	32	36.5	39.0	29.8
Overlap-first	64	55.9	57.9	56.7
Diversity	4	6.4	13.1	3.2
Diversity	8	11.3	17.8	7.9
Diversity	16	27.8	32.5	15.8
Diversity	32	33.5	38.0	30.9
Diversity	64	56.2	59.9	59.7

For each case, unpurchased privacy costs receive the mean of the purchased clipped costs, $\widehat{P} = |Q_{t,K}|^{-1} \sum_{a \in Q_{t,K}} [\mathrm{dP}_{\mathrm{true}}(a)]_+$; utility receives the purchased mean of ΔU. Purchased values remain exact. The DEV purchase-budget sweep in the main paper averages raw privacy responses before clipping and uses all 51 DEV cases. Its mean-fill results therefore use a different privacy-fill definition and cohort. Privacy Spearman uses clipped $P^{(1)}$.

Table S7(b): DEV post-fill fidelity across acquisition policies and purchase budgets.

Policy	K	Privacy Spearman	Utility Spearman	Intermediate recall (%)	Intermediate precision (%)	Pareto Jaccard
Canonical	4	0.1756 (39 / 9)	0.1955 (48 / 0)	85.8 (27 / 21)	66.9 (25 / 23)	0.1593 (48 / 0)
Canonical	8	0.2373 (43 / 5)	0.2554	79.2	74.0 (26 / 22)	0.2064
Canonical	16	0.3270 (46 / 2)	0.3367	72.3	81.8 (26 / 22)	0.3627
Canonical	32	0.4682 (46 / 2)	0.4851	84.0	84.2 (27 / 21)	0.4072
Canonical	64	0.6759 (47 / 1)	0.7303	88.4	89.5	0.6213
Random	4	0.1792 (43 / 5)	0.1576	75.4	71.5	0.0804
Random	8	0.2359 (47 / 1)	0.2370	73.2	80.6	0.1356
Random	16	0.3189 (47 / 1)	0.3294	76.2	81.5	0.1788
Random	32	0.4563 (47 / 1)	0.4899	82.2	83.2	0.2684
Random	64	0.6768 (47 / 1)	0.7426	91.5	89.1	0.5837
Overlap-first	4	0.1987 (40 / 8)	0.1765	85.8	67.0 (25 / 23)	0.1145
Overlap-first	8	0.2566 (43 / 5)	0.2361	79.2	74.0 (26 / 22)	0.1434
Overlap-first	16	0.3506 (44 / 4)	0.3314	68.7	83.4 (26 / 22)	0.2585
Overlap-first	32	0.4783 (46 / 2)	0.4771	84.0	84.8	0.3597
Overlap-first	64	0.6918 (47 / 1)	0.7338	87.5	90.9	0.5422
Diversity	4	0.1887 (40 / 8)	0.1871	82.1	66.6 (24 / 24)	0.1130
Diversity	8	0.2335 (43 / 5)	0.2400	75.6	74.8 (26 / 22)	0.1667
Diversity	16	0.3198 (46 / 2)	0.3320	72.3	81.7 (26 / 22)	0.3073
Diversity	32	0.4638 (46 / 2)	0.4789	84.0	84.0	0.3585
Diversity	64	0.6768 (47 / 1)	0.7280	88.4	89.4	0.5724

G LIMITATIONS AND SCOPE

G.1 APPROXIMATION AND LABEL ACCESS

The evaluated global estimators fit privacy ordering weakly even on TRAIN, their TEST results therefore do not isolate a transfer failure. Almost all unpurchased TEST actions contain literal indicators absent from local fitting data. Local-BGE uses a different label-access protocol and was not tested as a 64-purchase fill model. These findings leave open whether other representations and fitting procedures predict unpurchased actions more accurately.

G.2 AUDITOR VALIDITY AND WORKLOAD COVERAGE

The disclosure coordinate measures the declared auditor's response. Calibration has substantial absolute error and cross-family action agreement is partial. Whether lower auditor-measured cost also reduces secret recovery by an independently specified evaluator remains an open construct-validity question. The workload contains positive-secret propositions. Balanced positive and negative labels occur in the separate controls. Balanced recovery performance requires evaluation on negative-secret workload cases as well.

History-conditioned increments use a history-conditioned DROP baseline. Lower incremental cost can accompany higher prior exposure and does not certify a safe accumulated transcript. The post-hoc DROP distributions and clipping decomposition quantify this baseline dependence (Appendix C.3). Frequent zero and tied minima limit cost-based distinctions between feasible releases. They do not establish agreement with independent recovery outcomes. Freebase protected facts are selected deterministically from topic-neighborhood edges, without requiring practical sensitivity or derivability from the witness. CQ binds the witness directly to the protected fact. Generated reference hits do not verify answer support, and per-action support is unavailable for the scored transformed releases. The observed workload differences also combine evidence serialization, candidate availability, and reader behavior. Results condition on the fixed reader prompt and greedy decoding.

G.3 GRAPHRAG INSTANTIATION AND EVALUATION SCOPE

Headroom and attainability depend on the frozen inventory, reader, and answer criterion. Headroom and selection use empty history and co-evidence. Truth admits evidence under the pinned artifact's rules. It does not establish real-world factual correctness. The reader-value coordinate requires a frozen reference answer and is evaluated offline, with an exploratory utility grid. Transformed-release history comparisons cover 68 WebQSP/CWQ cases with semi-synthetic histories and empty same-round co-evidence, their feasibility findings come from post-hoc analysis on the exploratory grid. The experiments condition current measurements on supplied prior state. They do not evaluate online controller performance, deployment-optimal thresholds, or cumulative disclosure across successive chosen releases. The model-relative coordinates provide neither a formal semantic-confidentiality guarantee nor a monotonic cumulative-disclosure guarantee. Infrastructure secrecy, access traces, and comprehensive datastore extraction are outside this release-comparison task.