RESILIENT RAG DATASET INFERENCE VIA ORTHOGONAL DUAL-LAYER WATERMARKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Retrieval-Augmented Generation (RAG) allows unauthorized service providers to exploit proprietary text corpora while hiding provenance behind neural paraphrasing and black-box access, making dataset theft difficult to prove. Lexical watermarks wash away under deep rewriting; injected semantic facts survive paraphrasing but are easily pruned by content filters, or are prohibited in regulated domains such as medicine and law where facts cannot be altered. The Thief's Dilemma formalizes a carrier-conditional bottleneck: under joint carrier contraction, suppressing both channels violates any utility requirement above the carrier-free factual rate. Fact-only reconstruction lies outside this family and can erase regulated-mode provenance. We present orthogonal dual-layer watermarking with a black-box Prober-Verifier. The semantic layer embeds mode-specific invariants: open-domain documents receive long-tail predicate bindings, whereas regulated documents receive truth-preserving coordinate permutations that leave the causal claim graph strictly intact. With semantics frozen, Bidirectional Masked Contextual Resampling (BMCR) introduces tokenizer-independent functional-word biases into disjoint syntactic slots. An active prober uses at most 40 target-side API calls, fusing complementary evidence via a directional Gaussian-copula test. Across a 150-setting sweep of carrier-retaining attacks on the RDI benchmark, our joint auditor maintains at least 98.17% detection power at a 0.1% false-positive rate. Under an aligned 40-call budget, dual-layer attribution reaches 99.35% power against adaptive mixed attacks, exceeding the active semantic channel alone by 15.27 percentage points. Finally, we provide finite text-edit certificates, prove noninferiority on downstream clinical and legal tasks, and show that amortized calibration reduces attribution overhead to 110.68 calls per audit in production monitoring. Code is available at https://anonymous.4open.science/r/rpd_anonymous-F278.

1 INTRODUCTION

Retrieval-Augmented Generation (RAG) anchors neural responses in external knowledge bases (Lewis et al., 2020; Guu et al., 2020; Borgeaud et al., 2022; Gao et al., 2023b). This creates a vulnerability: unauthorized providers can index proprietary documents, exploit their facts, and answer queries through black-box generation while concealing the corpus. Such misappropriation threatens intellectual property rights and raises fair-use concerns (Huang et al., 2023; Min et al., 2024; Henderson et al., 2023; Karamolegkou et al., 2023). Proving whether a suspect black-box system relies on a specific protected corpus defines the problem of RAG dataset inference (Jovanović et al., 2025). Unlike dataset inference for pre-trained weights (Maini et al., 2021; 2024), retrieval auditing faces an asymmetric challenge: legitimate sources often report identical events, and generators paraphrase, synthesize, or compress retrieved context.

Document watermarking faces a technical impasse. Lexical watermarks erode under recursive paraphrasing and lossy neural rewriting (Kirchenbauer et al., 2023; Krishna et al., 2023; Kirchenbauer et al., 2024). Auxiliary semantic facts survive aggressive paraphrasing, but content-relevance filters and graph distillation can prune them. Injecting synthetic facts is unacceptable in clinical medicine, compliance documentation, or financial contracts, where factual changes create legal and operational liability. Prior watermarks treat lexical and semantic signals separately, leaving auditors vulnerable to asymmetric evasion.

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

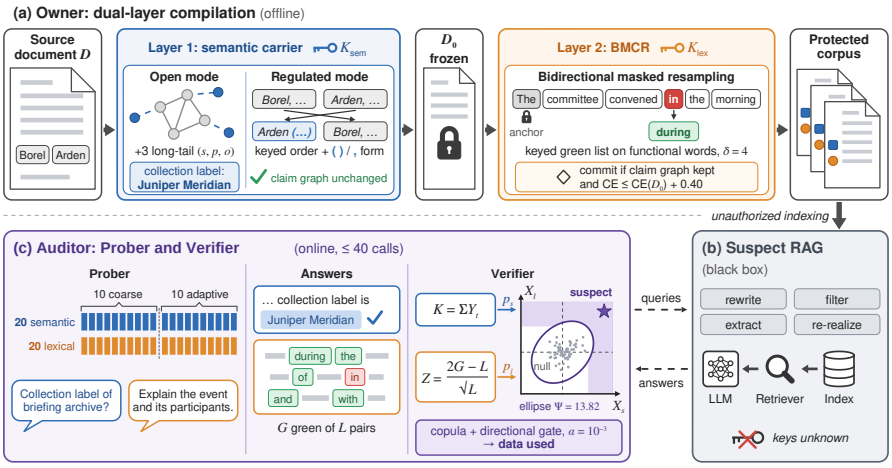


Figure 1: Dual-layer pipeline. (a) Separately keyed, disjoint layers: Layer 1 precedes BMCR's selection of disjoint functional slots in D_0 and anchor freezing. (b) The suspect knows the algorithms, not the keys. (c) Twenty semantic and twenty lexical calls share a 40-call budget.

These complementary failure modes create a defensive opportunity against carrier-retaining attacks. Paraphrasing can preserve keyed assertions while eroding functional-word statistics; extractive filtering can remove peripheral facts while retaining functional carriers. The Thief's Dilemma formalizes this utility tradeoff under joint carrier contraction, not for arbitrary fact-only re-encoding (Theorem 1). Exploiting this tension requires orthogonal signals with strictly disjoint information budgets and distinct linguistic representations.

We introduce dual-layer watermarking and black-box verification for RAG (Figure 1). The semantic layer has two modes. In open-domain corpora, it injects non-actionable, long-tail predicate bindings into peripheral syntactic nodes. In regulated domains, it enforces zero semantic modification: it permutes candidate Commutative Equivalence Classes (CECs) while preserving the complete causal and temporal claim graph, binding bits only to truth-equivalent realizations such as coordinate orders and non-restrictive appositives. After locking the semantic realization, Bidirectional Masked Contextual Resampling (BMCR) embeds the lexical channel in disjoint functional words via canonical byte-level hashing, making it invariant to subword tokenizers across black-box targets.

Under API rate and cost limits, our Prober-Verifier uses at most 40 attribution calls: twenty for semantic relation elicitation and twenty for lexical context retrieval via adaptive manifold routing. Probability integral transforms and a directional Gaussian-copula test combine both channels, bounding false-positive rates without assuming channel independence.

Our empirical and theoretical evaluations establish the resilience of this joint architecture:

1. **A dual-layer document encoder** decoupling semantic invariants and lexical functional biases via explicit execution ordering, causal freeze masks, and strict claim-preservation checks.
2. **A budgeted 40-call Prober-Verifier** combining adaptive retrieval routing, calibrated count-law tail bounds, and a directional copula rule for valid statistical coverage.
3. **Benchmark and mechanism ablations** on RDI and external corpora: 99.35% dual-layer detection power with 40 calls against adaptive evasion, exceeding the isolated semantic channel by 15.27 percentage points at matched 0.1% FPR.
4. **Theoretical and operational bounds** formalizing the carrier-conditional Thief's Dilemma, providing finite query certificates and proving downstream-task noninferiority while reducing production monitoring costs to 110.68 calls per audit.

2 Related Work and Evaluation Setting

Retrieval augmentation. Large language models (Brown et al., 2020) can answer from passages returned by dense or nearest-neighbor retrievers (Karpukhin et al., 2020; Khandelwal et al., 2019;

Izacard & Grave, 2021; Izacard et al., 2023), attach retrieval to black-box generators, or learn when to retrieve (Ram et al., 2023; Shi et al., 2024b; Jiang et al., 2023b; Asai et al., 2024; Lin et al., 2024); the retrieved context is often compressed or filtered (Jiang et al., 2023a; Xu et al., 2024; Yoran et al., 2024). Retrieval matters most for long-tail facts (Mallen et al., 2023; Kandpal et al., 2023), and whether a generator follows retrieved evidence that conflicts with its parametric knowledge depends on the model and the evidence (Longpre et al., 2021; Xie et al., 2024). RAG corpora can also leak or be poisoned (Zeng et al., 2024; Qi et al., 2025; Zhong et al., 2023; Zou et al., 2025).

Text watermarks. Generation-time watermarks bias a keyed vocabulary partition (Kirchenbauer et al., 2023; Zhao et al., 2023a) or use keyed sampling that preserves the output distribution (Kuditipudi et al., 2023; Christ et al., 2024; Hu et al., 2024; Wu et al., 2023; Dathathri et al., 2024), with multi-bit, entropy-adaptive and semantic variants (Wang et al., 2024; Liu & Bu, 2024; Liu et al., 2024a). Paraphrasing, translation, spoofing and other adaptive attacks weaken them (Krishna et al., 2023; Kirchenbauer et al., 2024; He et al., 2024; Jovanović et al., 2024; Zhang et al., 2024; Pang et al., 2024), and zero-shot detectors need no key (Mitchell et al., 2023). Post-hoc methods edit existing text by masked infilling or invariant-feature substitution (Ueoka et al., 2021; Yoo et al., 2023) or by model-guided insertion (Chang et al., 2024). API owners also watermark outputs against imitation (He et al., 2022; Zhao et al., 2023b), and marks can transfer through training (Gu et al., 2024; Sander et al., 2024).

Membership and dataset inference. Membership inference asks whether a record was used in training (Shokri et al., 2017), and language models can reproduce training text (Carlini et al., 2021; 2022). For large language models, members and non-members are hard to separate (Duan et al., 2024); reference-free scores (Mattern et al., 2023; Shi et al., 2024a; Zhang et al., 2025), exchangeability tests (Oren et al., 2024) and dataset-level aggregation (Maini et al., 2021; 2024) add power, and traps, backdoor marks and data watermarks plant an owner signal (Sablayrolles et al., 2020; Li et al., 2022; Meeus et al., 2024; Wei et al., 2024).

Positioning. WARD studies low-order lexical watermarks, content-guided multi-query inference and factual redundancy through FARAD (Jovanović et al., 2025), and serves as our main reference. AAG and SIB instead infer membership from model responses (Anderson et al., 2024; Li et al., 2025), so factual overlap directly affects their nulls. SemStamp and PostMark address other semantic or post-generation watermark mechanisms (Hou et al., 2024; Chang et al., 2024). RAG-WM injects entity-relation knowledge watermarks (Lv et al., 2025); watermarked canaries insert additional documents (Liu et al., 2025); KMW uses knowledge completion and verification-query indexing (Yang et al., 2026). We compare these mechanism classes under aligned token and call allowances (Appendix D.3). Our additions are BMCR, the open/regulated semantic distinction, and measured gains from complementary evidence beyond low order or multi-query access.

RDI. RDI contains 9,000 articles for 3,000 RepLiQA source events (Monteiro et al., 2024), with three generator-authored variants per event and a 100-persona pool. Easy attribution uses unique facts; Hard attribution includes legitimate articles sharing the same event facts. All splits are by source event before generation or watermarking. The owner keys and open-mode predicate bindings are assigned after the evaluation checkpoints are fixed, and their novelty provides the attribution signal. FARAD serves as an external benchmark, and the controlled factorial analysis in Appendix D.3 measures the effect of style and generator diversity. The construction pipeline, annotation protocol and license are given in Appendix C.1.

3 Dual-Layer Prober-Verifier

Threat model and units. An owner compiles a corpus once with separate 256-bit semantic and lexical keys. A provider can retrieve, rewrite, filter, extract or re-realize its visible corpus; it knows the algorithms but not the keys (Appendix C.2 specifies the concrete adversary profiles). The auditor observes only answers. A complete audit is one dossier with a fixed target, key, route, query schedule and at most 40 calls. Retries, refusals and empty answers consume calls and remain in denominators. Calibration access is a separate permission: an owner can create known-negative tenants on a platform-locked profile without receiving the suspected tenant's index or cooperation (Appendices F.4 and F.4).

3.1 SEMANTIC COMPILATION AND BMCR

In open mode, Layer 1 adds three benign long-tail subject-predicate-object bindings to peripheral graph locations. Semantic detection checks the requested predicates with their entity roles, polarity and qualifiers. For response t requesting a predeclared nonempty set R_t, $Y_t = \prod_{r \in R_t} \mathbf{1}[\text{answer } t \text{ entails } r]$ and $K = \sum_{t=1}^{N_s} Y_t$. A response requesting three predicates is a single all-hit trial. Questions contain subjects and relation requests but never the expected watermark objects.

In regulated mode no propositions are added. Candidate Commutative Equivalence Classes (CECs) must respect the causal freeze mask and preserve the complete labelled claim graph after realization. An entity moves with its attached attribute, including "respectively" pairings. The keyed signal uses relative order and registered truth-equivalent comma/parenthesis realizations. With $m \geq 2$ recovered entities, the semantic score is pairwise order concordance; $m = 1$ uses its recognized realization bit; $m = 0$ contributes the specified neutral score $1/2$. These scores use their own audit-level null law, which retains shared route-state dependence in the primary profiles. Exact rules and examples are in Appendix A.2.

BMCR operates on the completed Layer 1 text D_0. Eligible surface words are drawn from a fixed 195-word functional inventory and compatible universal POS tags. Canonicalization is NFKC-casefold-NFKC; distinct word pairs are encoded by length-prefixed UTF-8 and colored by the low bit of HMAC-SHA256. The preceding eligible functional word is the hash anchor, rather than a proprietary tokenizer ID. Candidate substitutions satisfy POS, cosine-similarity and length checks, with frozen entity, qualifier, semantic-realization and predecessor spans.

Disjoint active slots are rescored bidirectionally with green bias $\delta = 4$ (analysed in Appendix H.1) and a soft cross-entropy penalty (Eq. 20). A separate hard commit rule requires the *jointly realized* chunk to preserve anchors and the labelled claim graph and to satisfy $\mathrm{CE}(D_{\mathrm{tent}}) \leq \mathrm{CE}(D_0) + 0.40$ nats per reference token. Failure returns D_0, which stays fixed across tentative batches. The same encoder is used throughout the ablations.

The detector extracts eligible pairs separately inside each answer and globally deduplicates their union. If L types contain G green types, the descriptive lexical score is $(2G - L)/\sqrt{L}$ for $L > 0$ and zero for empty support. Repeated boilerplate adds no observations, and response boundaries never create new pairs. The detector uses the corresponding randomized count-law tail. Because pairs are defined on surface words, the score of an output string is the same under every tokenizer.

3.2 BUDGETED QUERIES AND THE DIRECTIONAL DETECTOR

The default allocation is $N_s = N_l = 20$. Fixed-medoid probing selects one semantic/lexical pair per owner-side cluster. LMAP+Zoom instead spends ten coarse pairs, then ten adaptively routed pairs. Feedback uses only source-entity recovery in the returned answers. Query frames and tie-breaking are specified in Appendix I.1; routing is given in Algorithm 1. All final-score calibrations repeat the same policy, including routing, deduplication, answer limits and any unwatermarked surface scaffolds. Query watermarking is disabled in every reported profile.

For either channel's discrete statistic S, the null transform is $p = \Pr_0(S' > S) + V \Pr_0(S' = S)$, with $V \sim \mathcal{U}(0, 1)$. Randomizers are independent across channels and dossiers and fixed before scoring. For open semantics the primary complete-dossier model is $K \sim \mathrm{BetaBin}(N_s, \hat{a}_s, \hat{b}_s)$, fitted by factorial moments on complete null dossiers; the fitted shapes agree with the response-control posterior (Appendix F.1). The lexical null count law is fitted for the deployed key and task. PIT makes each marginal uniform under its registered count law, and the dependence between the two channels is modelled separately.

Writing $X_i = \Phi^{-1}(1 - p_i)$, the joint statistic is

$$\Psi = \frac{X_s^2 + X_l^2 - 2\hat{\rho} X_s X_l}{1 - \hat{\rho}^2}.$$

At nominal $\alpha = 10^{-3}$ the decision requires $\Psi \geq -2 \log \alpha = 13.81551\ldots$ and $(X_s > 0 \wedge X_l > 0) \vee \max(X_s, X_l) \geq \Phi^{-1}(0.999)$. The correlation is fitted on independent complete null dossiers.

Table 1: Detection against Claude Sonnet 4.5 ($N \leq 40$ per system). Balanced accuracy uses a validation-selected threshold; TPR@0.1% uses a separate nominal-FPR threshold. The 500 groups contain 1,000 H_1 and 1,000 H_0 audits; subscripts: 95% group-bootstrap confidence intervals. All but the passive WARD variant use active-adapted probing (Appendix C.4).

	RDI-Easy (unique)			RDI-Hard (redundant)			FARAD (external)		
Method (oracle active probing)	Acc.	AUC	TPR@0.1%	Acc.	AUC	TPR@0.1%	Acc.	AUC	TPR@0.1%
AAG (MIA)	$0.97_{\pm 0.01}$	0.981	82.4%	$0.51_{\pm 0.02}$	0.524	0.3%	$0.52_{\pm 0.02}$	0.531	0.4%
SemStamp (sentence embedding)	$0.88_{\pm 0.02}$	0.912	46.5%	$0.64_{\pm 0.03}$	0.684	18.2%	$0.61_{\pm 0.02}$	0.652	14.6%
PostMark (post-decoding insertion)	$0.92_{\pm 0.02}$	0.941	62.4%	$0.81_{\pm 0.02}$	0.842	48.7%	$0.83_{\pm 0.02}$	0.865	51.2%
WARD ($h \geq 4$, passive)	$0.93_{\pm 0.02}$	0.954	68.5%	$0.93_{\pm 0.02}$	0.951	66.2%	$0.98_{\pm 0.01}$	0.991	91.5%
Adaptive WARD ($h = 2$, active)	$0.96_{\pm 0.01}$	0.978	84.1%	$0.95_{\pm 0.01}$	0.972	81.3%	$0.99_{\pm 0.01}$	0.996	94.8%
Dual-layer (ours, same budget)	$\mathbf{0.99}_{\pm 0.01}$	**0.999**	**98.2%**	$\mathbf{0.98}_{\pm 0.01}$	**0.997**	**97.1%**	$\mathbf{0.99}_{\pm 0.01}$	**0.998**	**97.6%**

Under a Gaussian joint limit and consistent correlation fitting, the ungated quadratic form is χ^2_2 and the gated rejection probability is at most α (Theorem 2, proved in Appendix B.1).

3.3 The "Thief's Dilemma"

Let Y be the original task information, B the unwatermarked content, and κ_l, κ_s the chain-rule factual information in the two carriers. Under the stated joint carrier contraction, Theorem 1 (proved in Appendix A.1) gives

$$I(Y; \widehat{D}) \leq I(Y; B) + \kappa_l \, \epsilon_l / C_l + \kappa_s \, \epsilon_s / C_s.$$

If the required factual rate satisfies $R_{\min} > I(Y; B)$, simultaneous zero residual watermark information is impossible *within that channel family*. Appendix B.2 supplies a finite erasure-channel instance where the contraction is an equality and utility concerns only original facts, and characterizes the case in which a regulated fact-only projection erases both surface carriers (Proposition 4). Appendix B.3 proves a three-visit bound for the actual adaptive router. Proposition 9 gives a complementary finite text-edit certificate: lower bounds on semantic hits and distinct green types jointly force the unchanged detector to reject, including operating points where neither marginal rejects.

4 Mechanism and Robustness Results

Evaluation design. Each evaluation profile fixes the target, owner key, query policy, answer format and calibration partition (Appendix F). Datasets, models, the RAG system and the baseline adaptations are specified in Appendix C, and attack-specific results are in Appendix E. The isolation panels use a common 200-document RDI-Hard index, source-disjoint fitting and evaluation dossiers and complete 40-call budgets.

Main comparison. Table 1 benchmarks membership inference (AAG; Anderson et al., 2024), our high-order/passive variant of WARD (Jovanović et al., 2025), sentence-embedding partitioning (SemStamp; Hou et al., 2024), post-decoding insertion (PostMark; Chang et al., 2024) and an actively probed lexical baseline (Adaptive WARD, $h = 2$) under an identical query budget ($N \leq 40$) on the same source-passage workload. Every watermark baseline receives oracle active-adapted probing except the passive WARD reference. Under factual redundancy, AAG falls to chance accuracy on RDI-Hard, and Adaptive WARD reaches 81.3% TPR at the nominal 0.1% FPR. The dual-layer auditor retains 0.98 ± 0.01 accuracy, AUC 0.997 and 97.1% TPR@0.1% on RDI-Hard.

Carrier-retaining attack sweep. Figure 2 extends the discrete attack configurations to 150 settings of partial knowledge-graph washing, coherence screening and recursive rewriting (Appendix E.2). Detection ranges from 98.17% to 100%, including low-utility configurations; for generation quality $U \geq 7$ it is at least 99.95%. No evaluated configuration reaches the high-quality, low-detection region. The individual open-mode RDI-Hard stress tests likewise report 97% to 100% detection: retrieval-pipeline sanitization, knowledge distillation, query guardrails, multi-source dilution, canonical sorting, structured extraction, Kerckhoffs-compliant targeted editing and deterministic grammatical sanitization (Appendix E.3, Figures 11 and 12).

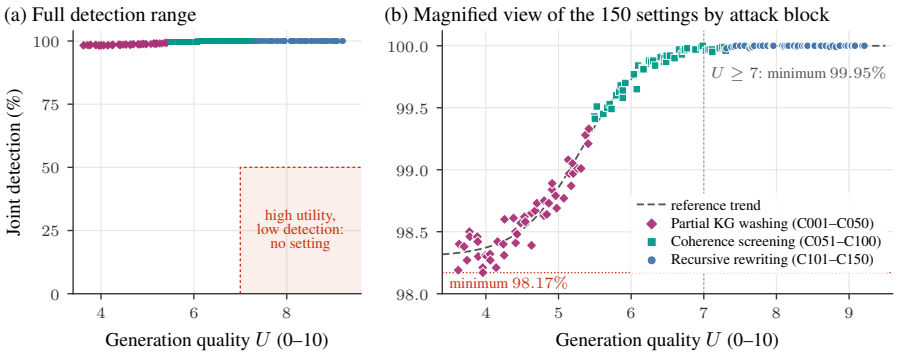


Figure 2: RDI-Hard carrier-retaining sweep. Each of 150 settings uses 100 seed panels of 100 H_1 audits, at most 40 queries each; markers denote Table 12's three attack blocks. (a) No setting reaches $\mathcal{D}_{\text{joint}} \leq 0.5,\ U \geq 7.0$. (b) Detection: 98.17%–100%; dashed curve: visual reference trend.

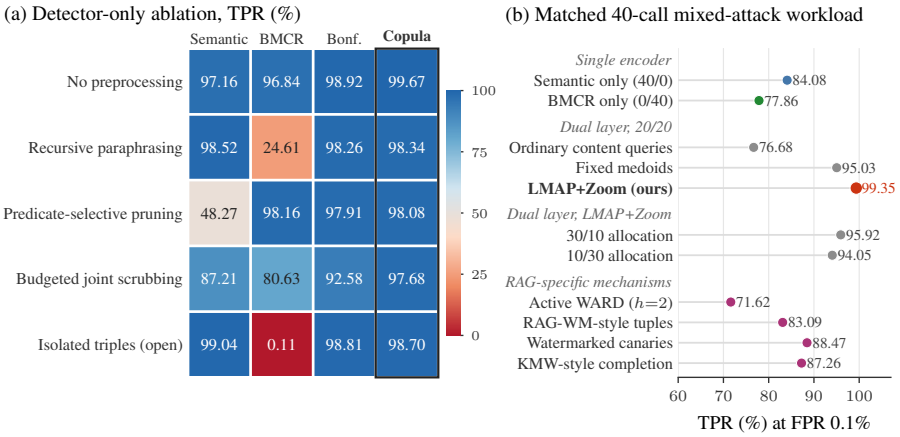


Figure 3: Mechanism isolation. (a) Detector-only TPR (%): 10,000 positive audits per attack, identical finalized dual-layer corpus and 20 semantic + 20 lexical answers; single channels use their designated 20 answers with marginal $\alpha = 0.001$. Bonferroni: $\min(p_s, p_l) \leq 0.0005$; Copula: Eq. (31) (matched-FPR rules: Figure 4a). (b) 40-call mixed-attack evaluation: 2,500 positive dossiers per family; common empirical FPR .1% on 50,000 separate ROC negatives; parentheses: N_s/N_l. RAG-specific mechanisms are budget-matched with at most 10% added corpus tokens. P-SP (Wieting et al., 2022), quality and construction resources: Table 6.

Complementary evidence. Figure 3a fixes both the compiled corpus and all answer strings, changing only the detector. Semantic evidence dominates strict triple extraction: its single-channel TPR is 99.04%, versus 98.70% jointly. Conversely, predicate-selective pruning leaves BMCR strong. Under joint scrubbing, neither marginal alone approaches the 97.68% joint rate. The benefit is thus complementary robustness across attacks.

Allocation and active probing. Giving the injected semantic encoder all 40 calls produces 84.08% power on the mixed-attack workload; the dual 20+20 allocation produces 99.35% at the same empirical FPR (Figure 3b). Their paired outcomes are 8,385 common detections, 1,550 dual-only, 23 semantic-only and 42 common misses. The gain is 15.27 percentage points, with paired normal interval $[14.55, 15.99]$ points (Appendix D). Ordinary queries and fixed medoids reach 76.68% and 95.03% on the same dual-layer corpus. The 30+10 and 10+30 allocations reach 95.92% and 94.05%, both below 20+20 on this mixed workload.

Combination rules. Figure 4a fits Brown and Copula on the same 2,048 null dossiers and compares all rules on the same answers. At common empirical FPR 0.1%, Bonferroni, fitted Brown and Copula reach 98.55%, 97.11% and 99.35% TPR. Bonferroni needs no dependence assumption when both marginal p values are valid and serves as a dependence-free reference. Positive Stouffer and a monotone maximum-of-positive-scores control reach 81.77% and 99.28%; the latter is within 0.07

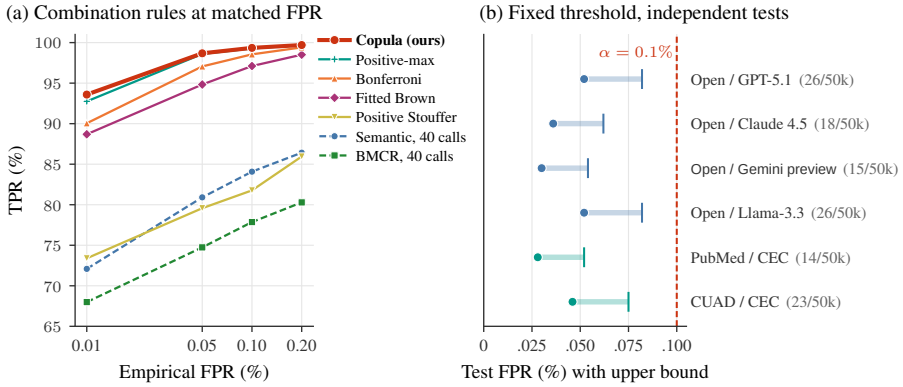


Figure 4: Combination rules and fixed-threshold calibration. (a) TPR: same 10,000 positives and paired marginals for all joint rules. Brown/Copula share 2,048 fit dossiers; Stouffer has fixed positive equal weights; Positive-max combines its score with both marginals. FPRs: 5, 25, 50 and 100 rejections among the same 50,000 ROC negatives. (b) Fixed-profile tails at unchanged directional $\alpha = 10^{-3}$: 2,048 fit, 10,000 validation and 50,000 independent test dossiers per profile, each with 40 calls and a fixed key. Dots: test FPRs (false-positive counts); bars: one-sided Clopper-Pearson upper endpoints, error probability $.05/6$. Values and fitted $\hat{\rho}$: Tables 7 and 16.

Table 2: Native-workflow comparison on the four-family attack mixture: forty attribution calls, at most 10% added text, 10,000 paired positives and original QA out of 5,000 questions. Fixed-rule FPR upper endpoints are at most .0736% on independent 50,000-dossier tests. Refit intervals are for BMCR minus the named workflow at the empirical ROC coordinate. Full seconds are per attempted compilation window. Table 9 and Figure 8 give native edit budgets, calibrations and further intervals.

Semantic layer plus	ROC TPR (%)	Fixed TPR (%)	QA	Full sec	Gain, 95% refit interval (pp)
WARD workflow	95.42	94.17	.9200	15.320	[2.438, 5.212]
Invariant workflow	96.83	95.75	.9220	8.780	[1.314, 3.547]
BMCR (ours)	**99.17**	**99.04**	**.9236**	**8.351**	Reference

points of Copula, so the dual-layer gain also holds with a monotone combiner. Appendix B.1 gives the alternative under which the gated quadratic rule is the Neyman-Pearson test, and Appendix B.3 a finite-query sufficient power bound for the unchanged detector.

Closest RAG mechanisms. The matched RAG-specific comparison in Figure 3b (Table 6) holds the attribution-call budget and added-token allowance fixed. Active low-order WARD reaches 71.62% on the mixed-attack workload; tuple injection, canary insertion and knowledge-completion mechanisms reach 83.09%, 88.47% and 87.26%, respectively, against 99.35% for the dual layer on the same source events and empirical FPR. Their editable objects and construction settings are specified in Appendix D.3.

BMCR versus a direct combination. With fixed semantic answers, replacing BMCR by active WARD or invariant infilling gives 90.31% and 94.73%, versus 99.35% for BMCR (Figure 8). Fresh retrieval, routing and forty-answer generation give 90.58%, 94.39% and 99.17% for these slot-matched arms (Appendix D.4). Expanded native workflows provide the strongest comparison in Table 2: BMCR gains 3.75 and 2.34 points at the common ROC coordinate, with positive source-cluster and complete-refit intervals for these workflows. Fixed-threshold gains are 4.87 and 3.29 points. Full-path left-context and unfrozen-predecessor interventions reach 95.86% and 94.44%; on independent human-written Wikinews, BMCR reaches 98.62% versus 96.07% for the strongest invariant workflow, with complete-refit gain interval $[1.412, 3.914]$ points (Appendix D.4). The same native encoders also undergo the sixteen-request OR16 attack: BMCR detects 90.98%, against 87.46% for semantic+WARD and 88.52% for semantic+invariant, at the fixed operational rule. All three meet the same original-task utility constraints. The paired gains are 3.52 $[3.060, 3.980]$ and 2.46 $[2.051, 2.869]$ percentage points (Table 11).

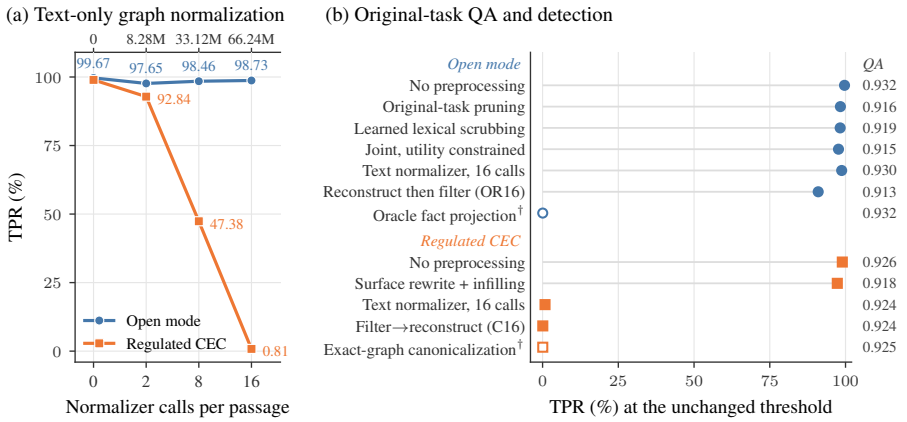


Figure 5: Normalization cost and mode boundary. (a) Text-only graph normalization without oracle annotations or owner keys. Open rows emit the resolved graph without surface fallback; CEC retains the passage if extraction fails. Top axis: offline input-plus-output tokens for 2,000 passages. (b) Original-task QA (right; injected predicates excluded) and TPR for task-aware attacks, 16-call normalization and no-fallback filtering-reconstruction. Each TPR uses 10,000 complete audits at the unchanged operational threshold. †Carrier-erasure controls (hollow). Values: Table 10.

Task-aware attacks and the mode boundary. The task-aware attack optimizer scores only original facts and ordinary user questions. It is permitted to delete inserted attributes without a direct utility penalty, learns functional-pair bias from visible text, and selects its pruning/infilling combination on an attack-validation split. The constrained joint attack retains original QA 0.915 versus 0.932 without preprocessing and is detected at 97.68% (Figure 5b; Table 10). Exact-graph-oracle regulated normalization instead gives $6/10{,}000$ detections with essentially preserved original-task utility, consistent with the information-zero boundary in Proposition 4. The stronger open-mode oracle has the clean source and similarly erases the distinguishing predicates. For the ordinary 16-call text normalizer (Figure 5a), CEC original QA is .924 and TPR is .81%, whereas open-mode QA is .930 and TPR is 98.73%. At 2,000 passages, both incur 32,000 offline requests and 66.24 million input-plus-output tokens, amortizable over future uses (Appendix D.5). Normalization therefore affects the two modes differently. The no-fallback filtering-reconstruction composition retains 90.98% open detection at QA .913 and Fact F1 .928, versus 86.59% for semantic-only with forty calls; CEC reaches .07% at QA .924 (Appendix D.5). On the same OR16 answer strings, joint detection exceeds the semantic decision by 5.14 points; 508 dossiers pass neither marginal but pass the joint rule (Appendix D.5).

5 Calibration, Cost and Safety

Calibration partitions and inference. Figure 4b (Table 16) uses a fixed key and target per profile, with source-disjoint fit/validation/test partitions and the complete adaptive query protocol repeated in every null dossier. The six test rates range from 0.028% to 0.052%; the largest simultaneous upper endpoint is 0.0820%. Appendix F.1 supplies the source partition, marginal estimator and validation diagnostics. The noncooperating-tenant deployment certifies both null strata separately, which allows different mixture weights. The cross-tenant challenge admits all 12 supported tenants among 24 proposals, and all twelve unsupported cases, including same-software corpus shifts, return `UNCALIBRATED` (Appendix F.4). Thus its proposal-level coverage is 50%, whereas the separately registered monitoring intake admits $600/648 = 92.59\%$; detection rates and admission rates have separate denominators. Population identity is bound by the source-sampling attestation, not inferred from a software version or from a high watermark score.

Complete cost. Attribution costs 40 calls. With $M = 30$ response controls and $B = 2{,}048$ complete fit dossiers, open-profile cold start costs $40 + 30 + 40B = 81{,}990$ calls; regulated profiling with $M = 100$ costs 82,060. For A compatible dossiers, the open-profile call average is $40 + 81{,}950/A$ before held-out certification. Including 10,000 validation and 50,000 test dossiers changes this to $40 + 2{,}481{,}950/A$. Key, target, domain, output-format or routing changes require a new compati-

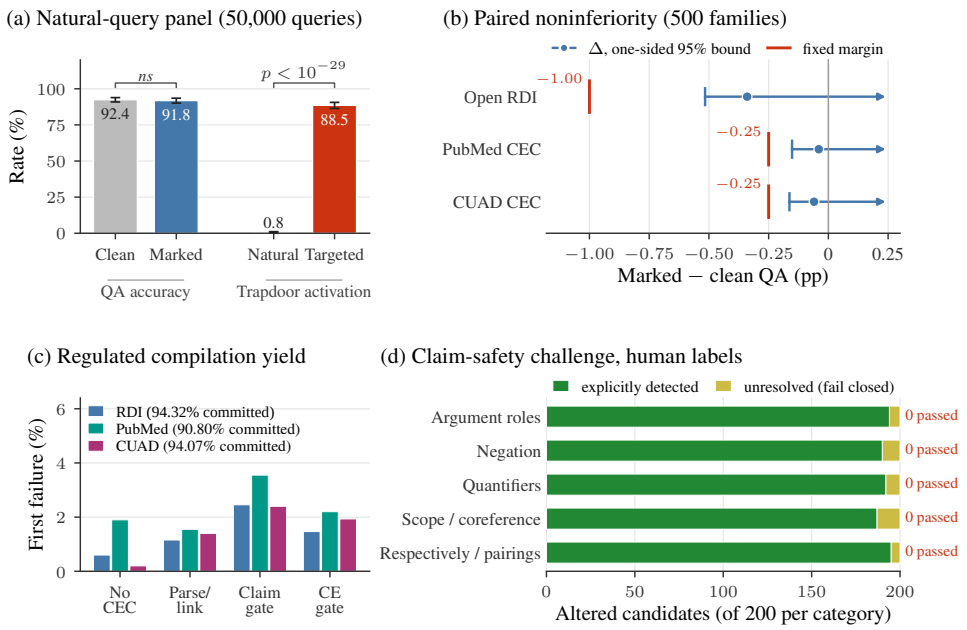


Figure 6: Safety and accepted coverage. (a) Natural queries: 100 matched families, five paraphrases each, 100 seed blocks; means ± across-seed panel SDs; Holm-adjusted exact family-level tests (Table 29). (b) Separate paired noninferiority panel: 500 source-disjoint families, 100 natural questions each per condition. Dots: marked-minus-clean; arrows start at one-sided 95% Student-t lower endpoints; red ticks: prespecified margins. (c) Regulated compilation first failures: mutually exclusive shares of attempted units; legend: successful two-layer commit rates. (d) Independent 2,000-pair challenge, two blind human labels and adjudication; none of 1,000 altered candidates passes. Counts: Appendix H.

bility check; failure requires refitting. Appendix G gives token, latency and amortization ledgers and separates cached scoring-kernel throughput from full compilation. Single-owner monitoring (Appendix F.4) amortizes the two-stratum certificate across 120,000 compatible audits, giving 110.68 total calls per attribution. Four independently scheduled audit flags form a tenant report, with a simultaneous false-report bound of 1.375% including certificate uncertainty (Appendix F.4).

Safety and accepted coverage. In the 50,000-query open-mode panel, trapdoor activation is 0.8% for natural queries versus 88.5% for targeted probes; QA changes by −0.6 points (Figure 6a). All three domains meet fixed noninferiority margins in a separate 500-family panel (Figure 6b). Regulated compilation preserves the extracted labelled claim graph, and Figure 6c reports every parsing, eligibility, claim and CE rejection. In an independent role/negation/quantifier/scope/pairing challenge, all 1,000 altered candidates are rejected: 958 explicitly detected alterations and 42 unresolved cases (Figure 6d). Appendix H separates watermark-attributable from total factual anomalies and specifies ethical use and failure handling.

6 Conclusion

Separately keyed semantic and lexical carriers, active probing and dependence-aware testing form a robust attribution system across the evaluated attack families. Matched-answer and matched-budget controls locate the improvement in complementary evidence rather than in factual redundancy, low order or extra queries alone. The carrier-conditional theory, together with the calibration, safety and cost measurements, characterizes the empirical frontier of dual-layer RAG attribution.

AI USE STATEMENT

We used generative AI tools to polish the language and improve the readability of this manuscript. The authors take full responsibility for the final content.

ETHICS STATEMENT

This work involves volunteer human annotators and a released benchmark. Annotators receive the task description, are informed of the use of their judgements and may withdraw (Appendix C.1). RDI uses fictional seed scenarios rather than claims about real individuals, and its data and code are restricted to non-commercial research use. Open-mode additions are restricted to non-actionable long-tail attributes, and regulated deployments use only truth-preserving realizations; Appendix H.4 specifies ethical use, failure handling and correctability.

REPRODUCIBILITY STATEMENT

The appendices specify the data splits, calibration, probing, parsing, attack configurations and statistical units. Prompt interfaces, the functional inventory and exact worked scoring values are included in the appendices. Main comparisons report paired counts and conditional, source-cluster and complete-refit intervals; finite certificates include the exact count-law tails.

REFERENCES

Maya Anderson, Guy Amit, and Abigail Goldsteen. Is my data in your retrieval database? membership inference attacks against retrieval augmented generation. *arXiv preprint arXiv:2405.20446*, 2024.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *International conference on learning representations*, volume 2024, pp. 9112–9141, 2024.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pp. 2206–2240. PMLR, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in neural information processing systems*, volume 33, pp. 1877–1901, 2020.

Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. In *30th USENIX security symposium (USENIX Security 21)*, pp. 2633–2650, 2021.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. Quantifying memorization across neural language models. *arXiv preprint arXiv:2202.07646*, 2022.

Yapei Chang, Kalpesh Krishna, Amir Houmansadr, John Frederick Wieting, and Mohit Iyyer. Postmark: A robust blackbox watermark for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8969–8987, 2024.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *The Thirty Seventh Annual Conference on Learning Theory*, pp. 1125–1139. PMLR, 2024.

Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, et al. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, 2024.

Michael Duan, Anshuman Suri, Niloofar Mireshghallah, Sewon Min, Weijia Shi, Luke Zettlemoyer, Yulia Tsvetkov, Yejin Choi, David Evans, and Hannaneh Hajishirzi. Do membership inference attacks work on large language models? *arXiv preprint arXiv:2402.07841*, 2024.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Precise zero-shot dense retrieval without relevance labels. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1762–1777, 2023a.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2023b.

Chenchen Gu, XIANG LI, Percy Liang, and Tatsunori Hashimoto. On the learnability of watermarks for language models. In *International Conference on Learning Representations*, volume 2024, pp. 38745–38768, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pp. 3929–3938. PMLR, 2020.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*, 2021.

Xuanli He, Qiongkai Xu, Yi Zeng, Lingjuan Lyu, Fangzhao Wu, Jiwei Li, and Ruoxi Jia. Cater: Intellectual property protection on text generation apis via conditional watermarks. In *Advances in Neural Information Processing Systems*, volume 35, pp. 5431–5445, 2022.

Zhiwei He, Binglin Zhou, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, and Rui Wang. Can watermarks survive translation? on the cross-lingual consistency of text watermark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4115–4129, 2024.

Peter Henderson, Xuechen Li, Dan Jurafsky, Tatsunori Hashimoto, Mark A Lemley, and Percy Liang. Foundation models and fair use. *Journal of Machine Learning Research*, 24(400):1–79, 2023.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. Cuad: An expert-annotated nlp dataset for legal contract review. *arXiv preprint arXiv:2103.06268*, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

Abe Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. Semstamp: A semantic watermark with paraphrastic robustness for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4067–4082, 2024.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Zhengmian Hu, Lichang Chen, Xidong Wu, Yihan Wu, Hongyang Zhang, and Heng Huang. Unbiased watermark for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 45408–45436, 2024.

Yangsibo Huang, Samyak Gupta, Zexuan Zhong, Kai Li, and Danqi Chen. Privacy implications of retrieval-based language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 14887–14902, 2023.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th conference of the european chapter of the association for computational linguistics: main volume*, pp. 874–880, 2021.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. Llmlingua: Compressing prompts for accelerated inference of large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 13358–13376, 2023a.

Zhengbao Jiang, Frank F Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig. Active retrieval augmented generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 7969–7992, 2023b.

Nikola Jovanović, Robin Staab, and Martin Vechev. Watermark stealing in large language models. *arXiv preprint arXiv:2402.19361*, 2024.

Nikola Jovanović, Robin Staab, Maximilian Baader, and Martin Vechev. Ward: Provable rag dataset inference via llm watermarks. In *International Conference on Learning Representations*, volume 2025, pp. 93288–93314, 2025.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. Large language models struggle to learn long-tail knowledge. In *International conference on machine learning*, pp. 15696–15707. PMLR, 2023.

Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 7403–7412, 2023.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pp. 6769–6781, 2020.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. *arXiv preprint arXiv:1911.00172*, 2019.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International conference on machine learning*, pp. 17061–17084. PMLR, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 49660–49704, 2024.

Keshav Kolluru, Vaibhav Adlakha, Samarth Aggarwal, Mausam, and Soumen Chakrabarti. Openie6: Iterative grid labeling and coordination analysis for open information extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 3748–3761, 2020.

Kalpesh Krishna, Yixiao Song, Marzena Karpinska, John Wieting, and Mohit Iyyer. Paraphrasing evades detectors of ai-generated text, but retrieval is an effective defense. In *Advances in neural information processing systems*, volume 36, pp. 27469–27500, 2023.

Rohith Kuditipudi, John Thickstun, Tatsunori Hashimoto, and Percy Liang. Robust distortion-free watermarks for language models. *arXiv preprint arXiv:2307.15593*, 2023.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in neural information processing systems*, volume 33, pp. 9459–9474, 2020.

Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. In *Advances in Neural Information Processing Systems*, volume 35, pp. 13238–13250, 2022.

Yuying Li, Gaoyang Liu, Chen Wang, and Yang Yang. Generating is believing: Membership inference attacks against retrieval-augmented generation. In *ICASSP 2025 – IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10889013.

Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Richard James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 19138–19162, 2024.

Aiwei Liu, Leyi Pan, Xuming Hu, Shiao Meng, and Lijie Wen. A semantic invariant robust watermark for large language models. In *International Conference on Learning Representations*, volume 2024, pp. 6499–6519, 2024a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024b.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 2511–2522, 2023.

Yepeng Liu and Yuheng Bu. Adaptive text watermark for large language models. *arXiv preprint arXiv:2401.13927*, 2024.

Yepeng Liu, Xuandong Zhao, Dawn Song, and Yuheng Bu. Dataset protection via watermarked canaries in retrieval-augmented llms. *arXiv preprint arXiv:2502.10673*, 2025.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7052–7063, 2021.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Peizhuo Lv, Mengjie Sun, Hao Wang, Xiaofeng Wang, Shengzhi Zhang, Yuxuan Chen, Kai Chen, and Limin Sun. Rag-wm: An efficient black-box watermarking approach for retrieval-augmented generation of large language models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1709–1723, 2025.

Pratyush Maini, Mohammad Yaghini, and Nicolas Papernot. Dataset inference: Ownership resolution in machine learning. *arXiv preprint arXiv:2104.10706*, 2021.

Pratyush Maini, Hengrui Jia, Nicolas Papernot, and Adam Dziedzic. Llm dataset inference: Did you train on my dataset? In *Advances in Neural Information Processing Systems*, volume 37, pp. 124069–124092, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 9802–9822, 2023.

Justus Mattern, Fatemehsadat Mireshghallah, Zhijing Jin, Bernhard Schölkopf, Mrinmaya Sachan, and Taylor Berg-Kirkpatrick. Membership inference attacks against language models via neighbourhood comparison. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 11330–11343, 2023.

Matthieu Meeus, Igor Shilov, Manuel Faysse, and Yves-Alexandre De Montjoye. Copyright traps for large language models. *arXiv preprint arXiv:2402.09363*, 2024.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 12076–12100, 2023.

Sewon Min, Suchin Gururangan, Eric Wallace, Weijia Shi, Hannaneh Hajishirzi, Noah Smith, and Luke Zettlemoyer. Silo language models: Isolating legal risk in a nonparametric datastore. In *International Conference on Learning Representations*, volume 2024, pp. 25132–25159, 2024.

Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pp. 24950–24962. PMLR, 2023.

Joao Monteiro, Pierre-Andre Noel, Etienne Marcotte, Sai Rajeswar, Valentina Zantedeschi, David Vazquez, Nicolas Chapados, Christopher Pal, and Perouz Taslakian. Repliqa: A question-answering dataset for benchmarking llms on unseen reference content. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24242–24276, 2024.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP workshop and shared task*, pp. 319–327, 2019.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Yonatan Oren, Nicole Meister, Niladri Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *International Conference on Learning Representations*, volume 2024, pp. 16354–16372, 2024.

Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. No free lunch in llm watermarking: Trade-offs in watermarking design choices. In *Advances in Neural Information Processing Systems*, volume 37, pp. 138756–138788, 2024.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543, 2014.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. Language models as knowledge bases? In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 2463–2473, 2019.

Slav Petrov, Dipanjan Das, and Ryan McDonald. A universal part-of-speech tagset. *arXiv preprint arXiv:1104.2086*, 2011.

Zhenting Qi, Hanlin Zhang, Eric P Xing, Sham Kakade, and Hima Lakkaraju. Follow my instruction and spill the beans: Scalable data extraction from retrieval-augmented generation systems. In *International Conference on Learning Representations*, volume 2025, pp. 48733–48755, 2025.

Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 11:1316–1331, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Alexandre Sablayrolles, Matthijs Douze, Cordelia Schmid, and Hervé Jégou. Radioactive data: tracing through training. In *International Conference on Machine Learning*, pp. 8326–8335. PMLR, 2020.

Tom Sander, Pierre Fernandez, Alain Durmus, Matthijs Douze, and Teddy Furon. Watermarking makes language models radioactive. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21079–21113, 2024.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3715–3734, 2022.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 1715–1725, 2016.

Mrinank Sharma, Meg Tong, Tomek Korbak, David Duvenaud, Amanda Askell, Sam Bowman, Esin Durmus, Zac Hatfield-Dodds, Scott Johnston, Shauna Kravec, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International conference on machine learning*, pp. 31210–31227. PMLR, 2023.

Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. Detecting pretraining data from large language models. In *International Conference on Learning Representations*, volume 2024, pp. 51826–51843, 2024a.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. In *Proceedings of the 2024 conference of the north american chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, pp. 8371–8384, 2024b.

Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pp. 3–18. IEEE, 2017.

Honai Ueoka, Yugo Murawaki, and Sadao Kurohashi. Frustratingly easy edit-based linguistic steganography with a masked language model. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 5486–5492, 2021.

Lean Wang, Wenkai Yang, Deli Chen, Hao Zhou, Yankai Lin, Fandong Meng, Jie Zhou, and Xu Sun. Towards codable watermarking for injecting multi-bits information to llms. In *International Conference on Learning Representations*, volume 2024, pp. 39375–39401, 2024.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in neural information processing systems*, volume 33, pp. 5776–5788, 2020.

Johnny Wei, Ryan Wang, and Robin Jia. Proving membership in llm pretraining data via data watermarks. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13306–13320, 2024.

John Wieting, Kevin Gimpel, Graham Neubig, and Taylor Berg-Kirkpatrick. Paraphrastic representations at scale. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 379–388, 2022.

Yihan Wu, Zhengmian Hu, Junfeng Guo, Hongyang Zhang, and Heng Huang. A resilient and accessible distribution-preserving watermark for large language models. *arXiv preprint arXiv:2310.07710*, 2023.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. In *International Conference on Learning Representations*, volume 2024, pp. 35623–35646, 2024.

Fangyuan Xu, Weijia Shi, and Eunsol Choi. Recomp: Improving retrieval-augmented lms with context compression and selective augmentation. In *International Conference on Learning Representations*, volume 2024, pp. 43478–43502, 2024.

Ke Yang, Shuguang Yuan, Jing Yu, and Chi Chen. Knowledge-infused multi-bit watermarking for rag knowledge bases. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 21195–21218, 2026.

KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. Robust multi-bit natural language watermarking through invariant features. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2092–2115, 2023.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. Making retrieval-augmented language models robust to irrelevant context. In *International Conference on Learning Representations*, volume 2024, pp. 29862–29883, 2024.

Shenglai Zeng, Jiankun Zhang, Pengfei He, Yue Xing, Yiding Liu, Han Xu, Jie Ren, Shuaiqiang Wang, Dawei Yin, Yi Chang, and Jiliang Tang. The good and the bad: Exploring privacy issues in retrieval-augmented generation (rag). In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4505–4524, 2024.

Hanlin Zhang, Benjamin L Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. Watermarks in the sand: impossibility of strong watermarking for language models. In *Forty-first International Conference on Machine Learning*, 2024.

Jingyang Zhang, Jingwei Sun, Eric Yeats, Yang Ouyang, Martin Kuo, Jianyi Zhang, Hao Yang, and Hai Li. Min-k%++: Improved baseline for pre-training data detection from large language models. In *International Conference on Learning Representations*, volume 2025, pp. 64845–64862, 2025.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Xuandong Zhao, Prabhanjan Ananth, Lei Li, and Yu-Xiang Wang. Provable robust watermarking for ai-generated text. *arXiv preprint arXiv:2306.17439*, 2023a.

Xuandong Zhao, Yu-Xiang Wang, and Lei Li. Protecting language generation models via invisible watermarking. In *International conference on machine learning*, pp. 42187–42199. PMLR, 2023b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in neural information processing systems*, volume 36, pp. 46595–46623, 2023.

Zexuan Zhong, Ziqing Huang, Alexander Wettig, and Danqi Chen. Poisoning retrieval corpora by injecting adversarial passages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13764–13775, 2023.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. {PoisonedRAG}: Knowledge corruption attacks to {Retrieval-Augmented} generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844, 2025.

A METHOD DETAILS

We view RAG dataset inference as a game between a data owner and an unauthorized RAG provider. The provider applies a transformation $\mathcal{T}$, such as paraphrasing or filtering, to suppress watermark evidence while preserving generation utility U. The dual-layer watermark places complementary evidence in a semantic carrier and a lexical carrier, and the Prober-Verifier collects and combines both under a fixed query budget. This appendix states the information-theoretic model and specifies the components; Appendix B gives the remaining theory and proofs.

A.1 INFORMATION-THEORETIC FOUNDATIONS

A protected document $D = (w_1, \ldots, w_{|D|})$ carries factual payload Y. The owner embeds a semantic invariant f^* at entity positions $\mathcal{I}_{\text{sem}}$ with key $\mathcal{K}_{\text{sem}}$ and a lexical rule $\mathcal{G}$ at disjoint functional positions $\mathcal{I}_{\text{lex}}$ with key $\mathcal{K}_{\text{lex}}$. An attack kernel $T : D_{\mathcal{W}} \to \hat{D}$ tries to push the residual watermark information below $\epsilon_{\text{lex}}, \epsilon_{\text{sem}} \geq 0$ while keeping $\mathbb{E}[d(Y, \hat{Y})] \leq \Delta_{\max}$. By rate-distortion theory,

$$\mathbb{E}[d(Y, \hat{Y})] \leq \Delta_{\max} \implies I(Y; \hat{D}) \geq R(\Delta_{\max}) \equiv R_{\min}. \tag{1}$$

Lemma 1 (Posterior coupling relative to a factorized reference) *Let $C = (Y, D_{\text{clean}})$ and let P be the joint law of $(C, \mathcal{G}, f^*, \hat{D})$ with $\hat{D} = (w_1, \ldots, w_L)$. Let a reference law Q satisfy $Q(C, \mathcal{G}, f^*) = P(C)P(\mathcal{G} \mid C)P(f^* \mid C) = P(C, \mathcal{G}, f^*)$ and $Q(\hat{D} = w \mid g, f, c) = a_c(w, g)b_c(w, f)$. Suppose that, for every P-positive history and channel realization, the autoregressive kernels satisfy at every step either $p_t = q_t$, or they share a finite support on which $q_t(v) \geq c_{\min} > 0$ and $\|p_t - q_t\|_{\text{TV}} \leq \gamma_{\text{attn}}$. Then, in bits,*

$$I_P(\mathcal{G}; f^* \mid \hat{D}, C) \leq \varepsilon_{\text{leak}} \equiv \frac{2L\gamma_{\text{attn}}^2}{c_{\min} \ln 2}. \tag{2}$$

For $L \leq 800$, $c_{\min} \geq 0.20$ and $\gamma_{\text{attn}} \leq 0.0031$, $\varepsilon_{\text{leak}} \leq 0.111$ bits.

Proof. Bayes' rule and the factorized likelihood give $Q(g, f \mid w, c) = Q(g \mid w, c)Q(f \mid w, c)$. Writing $H = (\hat{D}, C)$, the chain rule for relative entropy and its decomposition against this product posterior give $D_{\text{KL},2}(P\|Q) \geq I_P(\mathcal{G}; f^* \mid H)$. Because the priors agree, the autoregressive chain rule gives $D_{\text{KL},2}(P\|Q) = \sum_{t=1}^{L} \mathbb{E}_P D_{\text{KL},2}(p_t\|q_t)$. For $\Delta = p_t - q_t$, the inequality $\ln x \leq x - 1$ yields

$$D_{\text{KL},2}(p_t\|q_t) \leq \frac{1}{\ln 2} \sum_v \frac{\Delta(v)^2}{q_t(v)} \leq \frac{2\|p_t - q_t\|_{\text{TV}}^2}{c_{\min} \ln 2}, \tag{3}$$

since the positive and negative parts of the zero-sum vector Δ each have mass $\|p_t - q_t\|_{\text{TV}}$. Identical kernels contribute zero; summing at most L terms proves the bound, and $2(800)(0.0031)^2/(0.20 \ln 2) \approx 0.1109$ bits. ∎

Definition 1 (Coupled carrier channel) *Let $w_{\mathcal{I}}$ be a carrier sequence that encodes Y and a watermark rule $\mathcal{W} \in \{\mathcal{G}, f^*\}$. An attack kernel acts on a coupled carrier with coefficient $\alpha \in (0, 1]$ if any attenuation $\eta \equiv 1 - I(\mathcal{W}; \hat{D})/H(\mathcal{W} \mid D_{\text{clean}})$ contracts the carrier's factual information,*

$$I(Y; \hat{D} \mid w_{\text{base}}) \leq (1 - \alpha\eta)\, I(Y; w_{\mathcal{I}} \mid w_{\text{base}}). \tag{4}$$

For the two carriers we use the joint contraction $I(Y; \hat{D}) \leq I(Y; w_{\text{base}}) + (1-\eta)\kappa_{\text{lex}} + (1-\lambda)\kappa_{\text{sem}}$ with the chain-rule quantities of Theorem 1.

Theorem 1 (Rate-distortion evasion bottleneck) *Let $D_{\mathcal{W}} = (w_{\text{base}}, w_{\mathcal{I}_{\text{lex}}}, w_{\mathcal{I}_{\text{sem}}})$ with unwatermarked context w_{base}, and let*

$$\kappa_{\text{lex}} = I(Y; w_{\mathcal{I}_{\text{lex}}} \mid w_{\text{base}}), \qquad \kappa_{\text{sem}} = I(Y; w_{\mathcal{I}_{\text{sem}}} \mid w_{\text{base}}, w_{\mathcal{I}_{\text{lex}}}), \qquad \Delta_{\text{carrier}} = \kappa_{\text{lex}} + \kappa_{\text{sem}}. \tag{5}$$

Let $C_{\text{lex}} = H(\mathcal{G} \mid D_{\text{clean}}) > 0$ and $C_{\text{sem}} = H(f^* \mid D_{\text{clean}}) > 0$, and assume the normalized residual informations $I(\mathcal{G}; \hat{D})/C_{\text{lex}}$ and $I(f^*; \hat{D})/C_{\text{sem}}$ lie in $[0, 1]$. Under the joint contraction of Definition 1 with $\alpha = 1$, any admitted kernel with $I(\mathcal{G}; \hat{D}) \leq \epsilon_{\text{lex}}$ and $I(f^*; \hat{D}) \leq \epsilon_{\text{sem}}$ satisfies

$$I(Y; \hat{D}) \leq I(Y; D_{\mathcal{W}}) - \Delta_{\text{carrier}} + \frac{\kappa_{\text{lex}}}{C_{\text{lex}}}\epsilon_{\text{lex}} + \frac{\kappa_{\text{sem}}}{C_{\text{sem}}}\epsilon_{\text{sem}}, \tag{6}$$

where $I(Y; D_{\mathcal{W}}) - \Delta_{\text{carrier}} = I(Y; w_{\text{base}}) \geq 0$.

Proof. The chain rule gives $I(Y; D_{\mathcal{W}}) = I(Y; w_{\text{base}}) + \kappa_{\text{lex}} + \kappa_{\text{sem}}$, so the base term is nonnegative for any factual distribution. Define $\eta = 1 - I(\mathcal{G}; \hat{D})/C_{\text{lex}}$ and $\lambda = 1 - I(f^*; \hat{D})/C_{\text{sem}}$ in $[0, 1]$. The evasion constraints give $\eta \geq 1 - \epsilon_{\text{lex}}/C_{\text{lex}}$ and $\lambda \geq 1 - \epsilon_{\text{sem}}/C_{\text{sem}}$. Substituting these bounds into the joint contraction $I(Y; \hat{D}) \leq I(Y; w_{\text{base}}) + (1 - \eta)\kappa_{\text{lex}} + (1 - \lambda)\kappa_{\text{sem}}$ yields Eq. (6). ∎

Corollary A.1 (Residual information under coupled carriers) *Under Theorem 1, if* $I(Y; w_{\text{base}}) < R_{\min}$*, equivalently* $\Delta_{\text{info}} \equiv I(Y; D_{\mathcal{W}}) - R_{\min} < \Delta_{\text{carrier}}$*, no admitted utility-preserving kernel achieves zero residual information in both watermarks. Moreover, write* $\kappa_1 = \kappa_{\text{lex}}$*,* $\kappa_2 = \kappa_{\text{sem}}$*, assume* $\kappa_1, \kappa_2, C_{\text{lex}}, C_{\text{sem}} > 0$ *and* $0 \leq \Delta_{\text{info}} < \kappa_1 + \kappa_2$*, and consider the relaxed zero-sum game over* $\Omega_{\mathbb{A}} = \{(\eta, \lambda) \in [0, 1]^2 : \kappa_1\eta + \kappa_2\lambda \leq \Delta_{\text{info}}\}$ *with payoff* $\mathcal{L} = w_1 C_{\text{lex}}(1 - \eta) + w_2 C_{\text{sem}}(1 - \lambda)$*,* $(w_1, w_2) \in \Delta_2$*. If*

$$-\frac{C_{\text{sem}}\Delta_{\text{info}}}{\kappa_2} \leq C_{\text{lex}} - C_{\text{sem}} \leq \frac{C_{\text{lex}}\Delta_{\text{info}}}{\kappa_1}, \tag{7}$$

the value is attained by equalizing the residual informations,

$$\mathcal{V}^* = \frac{C_{\text{lex}}C_{\text{sem}}(\kappa_1 + \kappa_2 - \Delta_{\text{info}})}{\kappa_2 C_{\text{lex}} + \kappa_1 C_{\text{sem}}}, \qquad w_1^* = \frac{\kappa_1 C_{\text{sem}}}{\kappa_2 C_{\text{lex}} + \kappa_1 C_{\text{sem}}} = 1 - w_2^*; \tag{8}$$

otherwise the attacker concentrates on one channel and $\mathcal{V}^* = C_{\text{lex}}(1 - \min(1, \Delta_{\text{info}}/\kappa_1))$ *or* $C_{\text{sem}}(1 - \min(1, \Delta_{\text{info}}/\kappa_2))$*. In every regime* $\mathcal{V}^* \geq \min(C_{\text{lex}}, C_{\text{sem}})(1 - \Delta_{\text{info}}/(\kappa_1 + \kappa_2)) > 0$*, and the value of the realizable game is at least* $\mathcal{V}^*$*.*

Proof. Setting both tolerances to zero bounds $I(Y; \hat{D})$ by $I(Y; w_{\text{base}}) < R_{\min}$, which proves the first claim. By Theorem 1, utility preservation implies $\kappa_1\eta + \kappa_2\lambda \leq \Delta_{\text{info}}$, so every realizable attenuation pair lies in $\Omega_{\mathbb{A}}$ and total evasion $(1, 1)$ does not. The payoff is bilinear over compact convex sets, hence $\mathcal{V}^* = \min_{\Omega_{\mathbb{A}}} \max\{C_{\text{lex}}(1 - \eta), C_{\text{sem}}(1 - \lambda)\}$. The envelope is nonincreasing in both coordinates, so the minimum lies on $\kappa_1\eta + \kappa_2\lambda = \Delta_{\text{info}}$. Under Eq. (7) the two lines cross at

$$\eta^* = \frac{\kappa_2(C_{\text{lex}} - C_{\text{sem}}) + C_{\text{sem}}\Delta_{\text{info}}}{\kappa_2 C_{\text{lex}} + \kappa_1 C_{\text{sem}}}, \qquad \lambda^* = \frac{\kappa_1(C_{\text{sem}} - C_{\text{lex}}) + C_{\text{lex}}\Delta_{\text{info}}}{\kappa_2 C_{\text{lex}} + \kappa_1 C_{\text{sem}}}, \tag{9}$$

giving $\mathcal{V}^* = C_{\text{lex}}(1 - \eta^*)$. The weights w^* make the derivative of $\mathcal{L}$ along the active boundary vanish, so $\mathcal{L}(\cdot, w^*) = \mathcal{V}^*$ on the boundary and exceeds it inside. Otherwise the minimum is at a corner of the boundary. For the lower bound, any feasible pair satisfies

$$\begin{aligned} \max\{C_{\text{lex}}(1 - \eta), C_{\text{sem}}(1 - \lambda)\} &\geq \min(C_{\text{lex}}, C_{\text{sem}})\left[\tfrac{\kappa_1(1-\eta)+\kappa_2(1-\lambda)}{\kappa_1+\kappa_2}\right] \\ &\geq \min(C_{\text{lex}}, C_{\text{sem}})\left(1 - \tfrac{\Delta_{\text{info}}}{\kappa_1+\kappa_2}\right). \end{aligned} \tag{10}$$

Minimizing over the smaller realizable set can only increase the value. ∎

A.2 Layer 1: semantic carriers

Layer 1 is induced from the realized text by deterministic dependency parsing and entity linking, which give a document graph $\mathcal{G}_{\text{doc}} = (\mathcal{V}_{\text{ent}}, \mathcal{E}_{\text{causal}})$. Its causal, temporal and predicative paths define the freeze relation

$$\mathcal{M}_{\text{freeze}} = \{(e_i, e_j) \in \mathcal{V}_{\text{ent}}^2 : \text{dist}_{\mathcal{G}_{\text{doc}}}(e_i, e_j) < \infty\}. \tag{11}$$

A keyed pseudorandom function $\pi^* = \mathcal{H}_{\text{PRF}}(\text{Context} \parallel \mathcal{K}_{\text{sem}})$ drives the operator

$$\Omega(D, \mathcal{K}_{\text{sem}}) = \begin{cases} D \oplus \mathcal{T}_{\text{long-tail}}(\pi^*), & \Delta_{\text{reg}} = 0 \text{ (open mode)}, \\ \text{Realize}_{\text{CEC}}\big(\text{Perm}_{\text{CEC}}(D, \pi^*), \mathcal{K}_{\text{sem}}\big), & \Delta_{\text{reg}} = 1 \text{ (regulated mode)}. \end{cases} \tag{12}$$

Open mode. $\mathcal{T}_{\text{long-tail}}$ attaches three benign long-tail attributes to zero-in-degree leaf nodes; Appendix I.1 traces a complete example. For one assertion the score is $S_{\text{edge}} = \mathbb{I}[\hat{D} \models (e_{\text{head}}, r, e_{\text{tail}})]$, where entailment checks entity roles, predicate, polarity and temporal or numerical qualifiers, as in atomic-fact verification (Min et al., 2023). If response t requests the predeclared set R_t,

$$Y_t = \prod_{r \in R_t} S^{(t)}_{\text{edge},r}, \qquad K = \sum_{t=1}^{N_s} Y_t. \tag{13}$$

A probe requesting all three trapdoors succeeds only if all three are recovered, so the hit vector $(1, 0, 1)$ scores 0. Abstentions, unresolved arguments, contradictions and partial answers score 0, every response stays in N_s, and negative controls use the same requested-set size. The attribute recall τ_{attr}, the fraction of queried entity-attribute bindings recovered with correct roles and polarity, is a separate descriptive score.

Regulated mode. No propositions are added. Perm_{CEC} selects commutative equivalence classes (CECs) $\mathcal{C}_k$ satisfying

$$\begin{gathered}(e_i = h) \ \vee \ \texttt{dep}(e_i, h) = \texttt{conj} \quad \forall e_i \in \mathcal{C}_k, \quad |\mathcal{C}_k| \geq 2, \\ (e_i, e_j) \notin \mathcal{M}_{\text{freeze}} \quad \forall e_i, e_j \in \mathcal{C}_k, \ i \neq j,\end{gathered} \tag{14}$$

where h is the head of the coordination. Non-restrictive appositive chains, symmetric modifiers across disjoint heads and unordered paratactic clauses supply further candidates. A realization D' is accepted only if its reparsed claim graph equals the source graph, $\mathcal{G}_{\text{claim}}(D') = \mathcal{G}_{\text{claim}}(D)$, where $\mathcal{G}_{\text{claim}}$ records entity identities, predicate argument roles, paired attributes, qualifier scope and coreference. An entity therefore moves with its attributes, including "respectively" pairings, and a candidate with a changed or unresolved binding is left unmodified. $\text{Realize}_{\text{CEC}}$ also binds each coordinate entity to one of two truth-equivalent realizations $\{v_0, v_1\}$ of its non-restrictive attribute, selected by

$$\beta_i = \mathcal{H}_{\text{PRF}}\big(\text{Norm}(e_i) \parallel \mathcal{K}_{\text{sem}}\big) \bmod 2. \tag{15}$$

Two templates are registered, nominal appositives and full relative clauses: v_0 encloses the unchanged attribute subtree in paired commas and v_1 in parentheses, as in "Arden, a registered supplier," and "Arden (a registered supplier)". For m realized entities of the queried CEC, the semantic score is

$$S_{\text{sem}} = \begin{cases} \binom{m}{2}^{-1} \sum_{i<j} \mathbb{I}\big[\text{sgn}(\text{pos}(\hat{e}_j) - \text{pos}(\hat{e}_i)) = \text{sgn}(\pi^*(j) - \pi^*(i))\big] = \frac{1}{2}(1 + \tau_b), & m \geq 2, \\ \mathbb{I}\big[\text{VariantBit}(\hat{a}_{(1)} \mid \hat{e}_{(1)}) = \beta_{(1)}\big], & m = 1, \\ 1/2, & m = 0, \end{cases} \tag{16}$$

where VariantBit returns $\perp$, scored 0, for a missing attribute, an unregistered realization or an unresolved binding. Only registered members of the queried CEC enter m; an unrelated or hallucinated entity does not enlarge its denominator. An entity is counted once, at its first unambiguous linked mention; conflicting bindings are unresolved rather than tied ranks. Under a uniform-permutation null the $m \geq 2$ score has mean $1/2$ and variance $(2m + 5)/[18m(m - 1)]$. The deployed profiles use the full audit-level null law of $T = \sum_t S_t$, including the recovered-entity count and the concordant-pair count, conditional on the registered model, key and task. In particular, three recovered entities give support $\{0, 1/3, 2/3, 1\}$; no rounding or binary recoding is performed. Exact rational accumulation and the primary state laws are specified in Appendix F.2. In the tables, τ_{entity} and τ_{global} denote order-concordance rates in $[0, 1]$. Layer 1 is completed before BMCR computes its masks.

Claim graph and parser. Parsing uses spaCy with `en_core_web_trf`; clinical linking uses scispaCy (Neumann et al., 2019) with a UMLS alias index, and other entities use an owner-side registry of normalized names and aliases. A claim-graph record stores the predicate, role-labelled arguments, polarity, modality, quantifiers with an ordered scope tree, time, quantity, qualifiers, coreference, entity-attribute pairings and causal or temporal edges. Deterministic rules create one claim node per finite verb, copular predicate, eligible appositive and clausal complement; map passive subjects to patients; attach negation, modality, quantifiers and units to their scope nodes; expand distributive coordination only when licensed by `each` or `every`; zip "respectively" lists only when lengths and attachments agree; and resolve a pronoun only with a unique number-compatible antecedent in a two-sentence window. Equality compares complete multisets after canonicalizing

entity identifiers and unordered claim lists, while scope-tree edges, roles and paired attributes are never reordered. Any unresolved record blocks the affected edit, so ambiguous constructions are left unchanged. The contractual instantiation ALDG maps obligation, permission, prohibition and liability predicates to role-labelled parties. For example, “Arden and Borel received 2 and 5 units, respectively” equals “Borel and Arden received 5 and 2 units, respectively” but differs from “Borel and Arden received 2 and 5 units, respectively”, and “Not every supplier submitted a bid” differs from “Every supplier did not submit a bid”.

Worked example. With the public reference key consisting of bytes `00,01,...1f`, the bit is the low bit of HMAC-SHA256 on the normalized entity bytes, giving $\beta_{\text{Arden}} = 1$ and $\beta_{\text{Borel}} = 0$. Permutations sort HMAC-SHA256 digests of $\text{Enc}(\text{context ID}, \text{Norm}(e))$ in ascending byte order, with entity-ID tie-breaking. Context ID `bids` gives digest prefixes `3eccb519` for Arden and `eda779ca` for Borel, placing Arden first. The source “Borel, a certified installer, and Arden, a registered supplier, each submitted one bid.” becomes “Arden (a registered supplier) and Borel, a certified installer, each submitted one bid.” with an identical claim graph. The answer “Arden (a registered supplier) submitted one bid” has $m = 1$ and score 1, the comma form scores 0, and omitting the attribute returns $\perp$.

A.3 Layer 2: BMCR

The owner edits static documents offline and never sees the target’s logits or subword tokenizer (Sennrich et al., 2016). BMCR therefore defines its partition on canonical surface words and rescores candidates with bidirectional context.

Slot selection. Let $F = (f_1 < \cdots < f_m)$ be the eligible functional words of the completed Layer-1 chunk. Eligibility requires membership in a fixed 195-word closed-class inventory and a universal POS (UPOS) tag (Petrov et al., 2011) in {ADP, AUX, CCONJ, SCONJ, DET, PART, ADV}. The span mask M_{span} freezes entity and literal mentions, predicate heads, negation, modal, numerical and scope-bearing tokens, the Layer-1 realizations and the dependency paths between entity pairs in $\mathcal{M}_{\text{freeze}}$. Active slots are chosen greedily in source order, excluding frozen positions and f_1, with rank distance $|r(i) - r(j)| \geq 2$ on the functional chain; every predecessor anchor $f_{r(i)-1}$ is thus reserved and frozen for the batch. Candidates satisfy

$$\mathcal{V}_{\text{cand}}(w_i) = \{v : \text{POS}(v) = \text{POS}(w_i),\ \cos(\mathbf{e}_v, \mathbf{e}_{w_i}) \geq \tau_{\text{sim}},\ |\text{len}(v) - \text{len}(w_i)| \leq \Delta_{\text{len}}\}, \quad (17)$$

with unit-normalized `glove.6B.300d` vectors (Pennington et al., 2014), $\tau_{\text{sim}} = 0.60$ and $\Delta_{\text{len}} = 4$ characters; 1,405 unordered inventory pairs pass both static filters. The complete inventory accompanies the scoring reference in Appendix J.1. The set comprehension in Eq. (17) ranges over that inventory only. The original word is always retained as a candidate. For a replacement v, POS compatibility means that reparsing the single-replacement context gives the original slot’s UPOS tag; an out-of-vocabulary embedding admits only the identity replacement. Candidate order is canonical UTF-8 byte order, which also breaks score ties. The 1,405 pairs precede contextual POS and frozen-span filtering. Static enumeration visits each distinct unordered pair once: all 195 words have embeddings, giving $\binom{195}{2} = 18{,}915$ pairs; 1,495 pass the cosine filter and 90 of these fail the length filter, leaving 1,405. Identity candidates are retained separately and are not included in this count.

Canonical hashing. The prefix of slot i is its preceding eligible functional word,

$$s_{<i} = \text{Norm}(w_{\text{prev}(i)}), \qquad \text{prev}(i) = \max\{j < i : \text{EligibleFunc}(w_j) = 1\}, \quad (18)$$

so an entity permutation that keeps the functional sequence leaves every hash input unchanged. With Norm given by NFKC, case folding and NFKC after UAX #29 word segmentation, the color of candidate v is

$$\xi(v \mid s_{<i}) = \text{LSB}\left(\text{HMAC}_{\text{SHA256}}\left(\mathcal{K}_{\text{lex}}, \text{Enc}(s_{<i}, \text{Norm}(v))\right)\right), \quad (19)$$

where $\text{Enc}(u, v) = \text{u32be}(|u|) \| u \| \text{u32be}(|v|) \| v$ on UTF-8 bytes is injective on normalized pairs. Candidates with $\xi = 1$ form the green list $\mathcal{G}(s_{<i})$, which plays the role of the keyed green list of Kirchenbauer et al. (2023).

Lemma 2 (Cross-tokenizer invariance) *If two tokenizers decode to the same surface string, every pair of successive eligible functional words (u, v) receives the same color $\xi(\mathit{Norm}(v) \mid \mathit{Norm}(u))$.*

Proof. Norm acts on UAX #29 words, so the hashed bytes depend only on the surface string and not on subword boundaries. ∎

Bidirectional rescoring and the hard gate. Using a 195-class masked-functional-word head on the bidirectional DeBERTa-v3-large encoder (He et al., 2021), each active slot selects

$$\tilde{w}_i = \arg \max_{v \in \mathcal{V}_{\text{cand}}(w_i)} \Big[\ell_{\text{bi}}(v \mid D_{\setminus \mathcal{I}_{\text{lex}}}) + \delta\, \mathbb{I}\big(v \in \mathcal{G}(s_{<i})\big) - \lambda_{\text{ppl}} \max\big(0, \text{CE}(v) - \text{CE}_0 - \gamma_{\text{tol}}\big) \Big], \quad (20)$$

where $D_{\setminus \mathcal{I}_{\text{lex}}}$ replaces each active surface-word span by exactly one mask. If h_i is that mask's bidirectional state, $\ell_{\text{bi}}(v) = \log \text{softmax}(W h_i + b)_v$, normalized over the entire ordered inventory, before candidate filtering. A surface word has one class regardless of its number of subword pieces, so no length-dependent product of token probabilities enters this score. All active slots share this simultaneously masked context and are processed in one batch; sequential single-slot conditionals are only the diagnostic comparator below. The head and its distillation are specified in Appendix J.1. We use $\delta = 4$, $\lambda_{\text{ppl}} = 1$ and $\gamma_{\text{tol}} = 0.40$ nats per reference token. For every candidate text D', $\text{CE}(D')$ is its total GPT-2-large token negative log-likelihood divided by the fixed number of non-padding reference tokens in D_0, and $\text{CE}_0 = \text{CE}(D_0)$. Thus both the soft penalty and the hard comparison have the same denominator even when a replacement changes tokenization. A separate hard gate checks the jointly realized batch D_{tent}:

$$D_{\text{out}} = \begin{cases} D_{\text{tent}}, & A_{\text{struct}} = 1 \text{ and } \text{CE}(D_{\text{tent}}) \le \text{CE}(D_0) + \gamma_{\text{tol}}, \\ D_0, & \text{otherwise.} \end{cases} \quad (21)$$

A_{struct} requires the same ordered functional positions, identical predecessor bytes at every active slot, unchanged frozen spans and an identical labelled claim graph. D_0 stays fixed during a chunk's compilation, so tolerance increments cannot accumulate and every emitted chunk satisfies the CE bound. The batched choices agree with sequential conditional argmax choices on over 98.4% of selected slots, and green counts are always recomputed on the final strings.

API slot completion. Owners without local models can use one keyed slot-completion call to Claude Haiku 4.5 per chunk. The owner computes the admissible and green candidates for at most 24 slots and requests a JSON array of replacements (Table 31); the key never enters the prompt, and the host applies the same anchor, frozen-span, claim-graph and CE checks. Over 100 panels of 100 finalized chunks, owner-key preferences give lexical $Z = 5.82 \pm 0.18$, against 0.02 ± 0.17 for key-blind and 0.00 ± 0.16 for independent-key controls.

Detection. Within each answer the Verifier extracts the eligible pairs and takes their union over the N_l lexical answers,

$$\mathcal{U}_{\text{lex}} = \bigcup_{t=1}^{N_l} \big\{ (\text{Norm}(w_{\text{prev}(j)}), \text{Norm}(w_j)) \in \text{answer } t \big\}, \quad L = |\mathcal{U}_{\text{lex}}|, \quad G = \sum_{(u,v) \in \mathcal{U}_{\text{lex}}} \xi(v \mid u). \quad (22)$$

The first eligible word of an answer and punctuation create no transition, and no pair crosses an answer boundary, so repeated boilerplate adds nothing. The lexical p value is the randomized tail

$$p_{\text{lex}} = \Pr_0(G' > G \mid L) + V_{\text{lex}} \Pr_0(G' = G \mid L), \qquad V_{\text{lex}} \sim \mathcal{U}(0,1), \quad (23)$$

under the fitted fixed-key law $G \mid L \sim \text{BB}(L, a_l, b_l)$, which reduces to $\text{Bin}(L, 1/2)$ in the ideal random-key model. The descriptive score is

$$Z_{\text{lex}} = \frac{G - L/2}{\sqrt{L/4}} = \frac{2G - L}{\sqrt{L}}, \qquad Z_{\text{lex}} = 0 \text{ if } L = 0. \quad (24)$$

Algorithm 1 LMAP+Zoom (two-stage adaptive probing); the MC variant uses the same routing

Input: Owner anchors Ω_0, fixed query frames, budget of 40 calls
Form $J_0 = \min(10, |\Omega_0|)$ cells; set $w(d) = 1$ and $\mathrm{visits}(d) = n_c = R_c = 0$
for each cell c in medoid-identifier order **do**
    Send one semantic and one lexical query for $d = \mathrm{medoid}(c)$
    Store both answers; update w, n_c, R_c and $\mathrm{visits}(d)$
end for
for $t = J_0 + 1, \ldots, 20$ **do**
    Choose c maximizing $(1 + R_c)/(2 + n_c) + \sqrt{2\log(1 + t)/(1 + n_c)}$
    Within c, choose d maximizing $w(d)/(1 + \mathrm{visits}(d))$
    Send one semantic and one lexical query for d; update w, n_c, R_c, $\mathrm{visits}(d)$
end for
Pool the 20 semantic scores and the deduplicated lexical union
Apply the fixed null transforms and one directional joint decision

A.4 PROBER

The Prober starts from owner-known unwatermarked anchor records: the designated documents in a targeted audit, or the whole corpus in a blind audit. Spherical k-means with farthest-first initialization, cosine distance and identifier tie-breaking partitions the anchors, and each probed anchor receives one semantic and one lexical request, so $N = N_s + N_l \le 40$. For anchor D_m^* and each probe type, the candidate set $\mathcal{Q}_m$ holds eight deterministic questions (four frames, each with or without the first two source-entity names), and the compiled probe is

$$q_m = \arg\max_{q \in \mathcal{Q}_m} \cos\big(\mathbf{e}(q), \mathbf{e}(D_m^*)\big) \quad \text{s.t.} \quad \mathrm{PPL}(q) \le 18.5. \tag{25}$$

Semantic requests ask for one preselected ledger relation without its value, in the manner of relational knowledge probing (Petroni et al., 2019), and lexical requests ask for an explanation of the event, its participants and its context. The dual-manifold (MC) variant adds up to two high-IDF functional bigrams of the unwatermarked anchor as a quoted phrasing clause, and its null dossiers use the same scaffolds.

LMAP+Zoom (Algorithm 1) spends ten coarse pairs on ten cells and ten more pairs routed by an upper-confidence rule. The reward of anchor d is its source-entity recovery $r_t(d) = |E_d \cap E(\hat{D}_{\mathrm{sem}}, \hat{D}_{\mathrm{lex}})| / \max(1, |E_d|)$ in the two returned answers, and anchor weights in the visited cell are updated as $w_{t+1}(d) = w_t(d) e^{2 r_t(d)}$. The fixed-medoid policy uses 20 cells with one pair each. Early stopping is disabled.

A.5 VERIFIER

Semantic marginal. In open mode, $M = 30$ negative-control calls reuse the topic and question form while requesting a non-indexed relation. With k_{neg} all-hit controls and a $\mathrm{Beta}(1, 1)$ prior, the predictive law of the audit count and its randomized tail are

$$\begin{aligned} K \mid k_{\mathrm{neg}} &\sim \mathrm{BetaBin}(N_s, 1 + k_{\mathrm{neg}}, 1 + M - k_{\mathrm{neg}}), \\ p_{\mathrm{sem}} &= \Pr_0(K' > K) + V \Pr_0(K' = K), \qquad V \sim \mathcal{U}(0, 1). \end{aligned} \tag{26}$$

Lemma 3 gives exact uniformity under this predictive law and Theorem 3 gives asymptotic calibration under a fixed rate. Each deployed profile fits the complete-dossier law $K \sim \mathrm{BB}(N_s, a_s, b_s)$ on $B = 2{,}048$ null dossiers by normalized factorial moments,

$$\hat{m}_1 = \frac{1}{B}\sum_i \frac{x_i}{n_i}, \qquad \hat{m}_2 = \frac{1}{B}\sum_i \frac{x_i(x_i - 1)}{n_i(n_i - 1)}, \qquad \hat{a} + \hat{b} = \frac{\hat{m}_1 - \hat{m}_2}{\hat{m}_2 - \hat{m}_1^2}, \tag{27}$$

with $\hat{a} = \hat{m}_1(\hat{a} + \hat{b})$ and $\hat{b} = (1 - \hat{m}_1)(\hat{a} + \hat{b})$ rounded half-up (Proposition 2). The lexical law is fitted in the same way on (G_i, L_i). The fitted open-mode shapes coincide with the response-control posterior. Regulated profiles use $M = 100$ controls and the exact convolution of the full per-response state law. Let d be the least common multiple of 2 and all $\binom{m}{2}$ in the registered CEC manifest. Write

$q_h(m, c)$ for the route-state probability of recovering $m \geq 2$ entities with c concordant pairs, and $q_h(1, b)$ for an isolated entity's bit-match indicator b. The two primary manifests have maximum arity six, so $d = 30$:

$$J = dT = d\sum_{t=1}^{20} S_t, \qquad \Pr(J = j) = [z^j]\{(1-w)f_0(z)^{20} + wf_1(z)^{20}\},$$
$$f_h(z) = q_h(0)z^{d/2} + q_h(1,0) + q_h(1,1)z^d + \sum_{m=2}^{6}\sum_{c=0}^{\binom{m}{2}} q_h(m,c)z^{dc/\binom{m}{2}}. \tag{28}$$

The registered law has conditionally independent response states given the shared dossier state H; mixing over H retains within-dossier dependence. The registered finite law is convolved exactly; its parameters and independent validation are in Tables 17 and 16. Profiles with other arities use their own d, or an equivalent dictionary of rational sums.

Joint test. With $U_i = 1 - p_i$ and $X_i = \Phi^{-1}(U_i)$, the working approximation is the Gaussian copula

$$\Pr_0(U_1 \leq u_1, U_2 \leq u_2) = \Phi_{\mathbf{\Sigma}^*}\big(\Phi^{-1}(u_1), \Phi^{-1}(u_2)\big), \qquad \mathbf{\Sigma}^* = \begin{pmatrix} 1 & \rho^* \\ \rho^* & 1 \end{pmatrix}, \tag{29}$$

where ρ^* is the correlation of the final audit-level scores. It is estimated on complete null dossiers that repeat the audit's allocation, routing, pooling and transforms, one score pair per dossier. For fixed finite counts and independent tie randomizers, marginal PIT uniformity does not imply an exact nonzero-correlation Gaussian joint law: within a PIT interval for one count, the other score's conditional law is constant, whereas a Gaussian conditional mean is $\rho^* x$. Increasing calibration dossiers alone does not remove this discreteness. Theorem 2 is conditional on a separate joint limit, not derived from fixed-N marginal calibration; deployed 20+20-call profiles instead use the independent finite-sample certificates of Appendix F.

Theorem 2 (Asymptotic validity of the copula statistic) *Let $T_{\text{joint}} = X_1^2 + X_2^2 - 2\hat{\rho}X_1X_2$. If the calibrated scores converge jointly to $\mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma}^*)$ with $|\rho^*| < 1$ and $\hat{\rho} \in (-1, 1)$ is consistent for ρ^*, then*

$$\Psi_{\text{copula}} = \frac{T_{\text{joint}}}{1-\hat{\rho}^2} \xrightarrow{d} \chi^2(2), \qquad \lim \Pr_0\left(\Psi_{\text{copula}} \geq \chi^2_{2,1-\alpha}\right) = \alpha, \quad 0 < \alpha < 1. \tag{30}$$

For exactly Gaussian scores and known ρ^* the $\chi^2(2)$ law is exact. Adding the directional gate of Eq. (31) gives rejection probability at most α in this oracle model and an asymptotic upper limit at most α under consistent estimation.

The decision rule at $\alpha = 0.001$ is

$$\Psi_{\text{copula}} \geq \chi^2_{2,1-\alpha} \quad \text{and} \quad \big((X_1 > 0 \wedge X_2 > 0) \vee \max(X_1, X_2) \geq \Phi^{-1}(1-\alpha)\big), \tag{31}$$

with $\chi^2_{2,0.999} = -2\ln 0.001 \approx 13.8155$ and $\Phi^{-1}(0.999) \approx 3.0902$. The gate restricts rejection to positive watermark evidence: when one score is nonpositive, the other must also pass its one-sided marginal threshold. Appendix B.1 gives the proofs and a confounder model for ρ^*.

B Further Theory and Proofs

B.1 Marginal calibration and copula validity

Even under H_0, a target conditions on its own retrieved context. Let $Z_{\text{align}} \in \{0, 1\}$ indicate high topical alignment between a probe and the retrieved background, with $\pi_{\text{align}} = \Pr(Z_{\text{align}} = 1)$, and let $(S_{\text{sem}}, Z_{\text{lex}})$ follow the corresponding two-component mixture.

Proposition 1 (Retrieval covariance decomposition under H_0) *Assume finite second moments. With $\mu_S^{(z)}, \mu_Z^{(z)}$ the conditional means, c_z the conditional covariance in state z, $\Delta\mu_S = \mu_S^{(1)} - \mu_S^{(0)}$ and $\Delta\mu_Z = \mu_Z^{(1)} - \mu_Z^{(0)}$,*

$$\text{Cov}_{H_0}(S_{\text{sem}}, Z_{\text{lex}}) = \pi_{\text{align}}(1-\pi_{\text{align}})\Delta\mu_S\Delta\mu_Z + R_{\text{cond}}, \qquad R_{\text{cond}} = (1-\pi_{\text{align}})c_0 + \pi_{\text{align}}c_1. \tag{32}$$

If $|c_z| \le \eta$, then $|R_{\text{cond}}| \le \eta$, and the covariance is positive whenever $\pi_{\text{align}}(1-\pi_{\text{align}})\Delta\mu_S\Delta\mu_Z > \eta$.

Proof. The law of total covariance gives $\mathbb{E}[\text{Cov}(\cdot \mid Z_{\text{align}})] + \text{Cov}(\mathbb{E}[S_{\text{sem}} \mid Z_{\text{align}}], \mathbb{E}[Z_{\text{lex}} \mid Z_{\text{align}}])$. The first term is R_{cond}. The conditional means are functions of a Bernoulli variable, and expanding their covariance gives $\pi_{\text{align}}(1-\pi_{\text{align}})\Delta\mu_S\Delta\mu_Z$. The bound on R_{cond} follows from the triangle inequality. ∎

Proof of Theorem 2. *Margins.* Under the predictive Beta-Binomial law, Lemma 3 gives $p_{\text{sem}} \sim \mathcal{U}(0,1)$; under a fixed-rate binomial null, Theorem 3 gives asymptotic uniformity. The randomized lexical tail is uniform whenever its count law is correct. Hence $X_i = \Phi^{-1}(1-p_i)$ are standard normal. *Joint law.* Under the copula model (29), $\mathbf{X}$ is jointly Gaussian with covariance $\mathbf{\Sigma}^*$; in the asymptotic statement joint convergence is the stated assumption. *Whitening.* Write $\mathbf{\Sigma}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$ with eigenvalues $1 \pm \rho^*$ and $\mathbf{Y} = \mathbf{\Lambda}^{-1/2}\mathbf{U}^\top\mathbf{X} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{I}_2)$. Then $\Psi^* = \mathbf{X}^\top(\mathbf{\Sigma}^*)^{-1}\mathbf{X} = Y_1^2 + Y_2^2 \sim \chi^2(2)$, exactly in the oracle model and in the limit by continuous mapping. *Plug-in.* Consistency and $|\rho^*| < 1$ give $\hat{\mathbf{\Sigma}}^{-1} \xrightarrow{p} (\mathbf{\Sigma}^*)^{-1}$; since $\|\mathbf{X}\|^2 = O_p(1)$, $\Psi_{\text{copula}} = \Psi^* + o_p(1) \xrightarrow{d} \chi^2(2)$ by Slutsky's theorem, and $\chi^2_{2,1-\alpha} = -2\ln\alpha$. The gated rejection event is a subset of the ungated one, which gives rejection probability at most α in the oracle model and an asymptotic upper limit at most α under consistent estimation. ∎

Lemma 3 (Uniformity of the Beta-Binomial tail) *Condition on M controls with count k_{neg} and let $Q_M = \text{BetaBin}(N, 1+k_{\text{neg}}, 1+M-k_{\text{neg}})$. If $K \sim Q_M$ and $V \sim \mathcal{U}(0,1)$ is independent, then $U = Q_M(K' > K) + VQ_M(K' = K)$ is exactly uniform. Without randomization, $Q_M(K' \ge K)$ is super-uniform.*

Proof. Given $K = k$, U is uniform on $[s_k, s_k + q_k]$ with $q_k = Q_M(K = k)$ and $s_k = Q_M(K > k)$. These intervals partition $[0,1]$ and interval k is selected with probability q_k, so U has unit density. The nonrandomized tail $s_K + q_K$ dominates U. ∎

Theorem 3 (Posterior contraction and asymptotic calibration) *Fix N and $\epsilon^* \in (0,1)$, with $k_{\text{neg}} \sim \text{Bin}(M, \epsilon^*)$ and an independent audit count $K \sim \text{Bin}(N, \epsilon^*)$. The posterior mean $\bar{\epsilon}_M = (1+k_{\text{neg}})/(2+M)$ converges almost surely to ϵ^*, with $\text{Var}(\bar{\epsilon}_M) \le 1/[4(M+2)]$. With $d_M = \|Q_M^{(N)} - \text{Bin}(N, \epsilon^*)\|_{\text{TV}}$,*

$$\sup_{a\in[0,1]} \left| \Pr_{\epsilon^*}(p_{\text{sem}} \le a \mid k_{\text{neg}}) - a \right| \le d_M \xrightarrow{\text{a.s.}} 0. \tag{33}$$

Proof. $\bar{\epsilon}_M = \frac{M}{M+2}\frac{k_{\text{neg}}}{M} + \frac{1}{M+2} \to \epsilon^*$ by the strong law, and $\text{Var}(\bar{\epsilon}_M) = M\epsilon^*(1-\epsilon^*)/(M+2)^2 < 1/[4(M+2)]$. The posterior variance is at most $1/[4(M+3)]$, so the mixing law contracts to δ_{ϵ^*}; each binomial mass is continuous in ϵ, hence $d_M \to 0$ almost surely. The randomized PIT kernel cannot increase total variation, and its output under $Q_M^{(N)}$ is uniform by Lemma 3. ∎

Proposition 2 (Normalized factorial moments) *If $X \mid n, \theta \sim \text{Bin}(n, \theta)$ and $\theta \mid n \sim \text{Beta}(a, b)$ with $n \ge 2$, then $\mathbb{E}(X/n) = a/(a+b)$ and $\mathbb{E}[X(X-1)/(n(n-1))] = a(a+1)/[(a+b)(a+b+1)]$, so $a + b = (m_1 - m_2)/(m_2 - m_1^2)$.*

Proof. The conditional factorial moments are $n\theta$ and $n(n-1)\theta^2$; average over the Beta law and solve $m_2 - m_1^2 = m_1(1-m_1)/(a+b+1)$. The identity holds conditionally on each recorded n. ∎

Theorem 4 (Oracle validity and pointwise covariance sensitivity) *Let $0 < \alpha < 1$ and let $\mathcal{R}_\alpha(r)$ be the rejection region of Eq. (31) evaluated with correlation $r \in (-1,1)$. (i) If $\mathbf{X} \sim \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma}(\rho^*))$ with $|\rho^*| < 1$, then $\Pr\{\mathbf{X} \in \mathcal{R}_\alpha(\rho^*)\} \le \alpha$; if $\mathbf{X}_M \xrightarrow{d} \mathcal{N}_2(\mathbf{0}, \mathbf{\Sigma}(\rho^*))$ and $\hat{r}_M \in (-1,1)$ satisfies $\hat{r}_M \xrightarrow{p} \rho^*$, then $\limsup \Pr\{\mathbf{X}_M \in \mathcal{R}_\alpha(\hat{r}_M)\} \le \alpha$. The Fisher-$z$ endpoint $\hat{\rho}_{\text{audit}} = \min\{0.95, \tanh(\text{arctanh}\,\hat{\rho}_{\text{emp}} + 1.96/\sqrt{M-3})\}$ is such an estimator for $\rho^* \in [0, 0.95]$ when*

it is computed from $M > 3$ independent audit-level null pairs under a stationary Gaussian sampling model. (ii) For $|r|, |\rho| \le \rho_{\max} < 1$,

$$|\Psi_r(\mathbf{x}) - \Psi_\rho(\mathbf{x})| \le \frac{|r-\rho|}{(1-\rho_{\max})^2}\|\mathbf{x}\|^2. \tag{34}$$

Proof. Part (i) follows from whitening and Theorem 2; the Fisher offset vanishes as $M \to \infty$ and clipping at 0.95 preserves consistency. For (ii), $\Psi_r(\mathbf{x}) = \frac{(x_1+x_2)^2}{2(1+r)} + \frac{(x_1-x_2)^2}{2(1-r)}$, so

$$\Psi_r - \Psi_\rho = \frac{r-\rho}{2}\Big[\frac{(x_1-x_2)^2}{(1-r)(1-\rho)} - \frac{(x_1+x_2)^2}{(1+r)(1+\rho)}\Big]. \tag{35}$$

Both denominators are at least $(1-\rho_{\max})^2$ and $(x_1+x_2)^2 + (x_1-x_2)^2 = 2\|\mathbf{x}\|^2$. The two terms have opposite signs, so the effect of a correlation change depends on $\mathbf{x}$. ∎

Shape of the quadratic rule. For $\rho > 0$, $\partial\Psi/\partial X_l = 2(X_l - \rho X_s)/(1-\rho^2)$, so the statistic is a directional omnibus departure test. At $\rho = .312$ and $X_s = 3.6$, the gated points $X_l = .1$ and $X_l = 1.1232$ have $\Psi \approx 14.12$ and $\Psi = 12.96$, and only the first exceeds $-2\log\alpha \approx 13.82$. The ordering is optimal for a precise alternative: with Gaussian null density ϕ_Σ, gate A and $0 < \eta < 1/2$, the density $h_\eta(x) \propto \mathbf{1}_A(x)\exp\{\eta x^\top \Sigma^{-1} x\}\phi_\Sigma(x)$ has a likelihood ratio that increases with Ψ inside A, so the gated quadratic regions are its Neyman-Pearson tests. The monotone controls in Section 4 use the same (X_s, X_l) with $T_{\mathrm{St}} = (X_s + X_l)/\sqrt{2+2\hat{\rho}}$ and $T_{\mathrm{PM}} = \max\{X_s, X_l, T_{\mathrm{St}}\}$.

B.2 An erasure-channel instance and the regulated boundary

The following finite channel satisfies the joint contraction of Theorem 1 with equality.

Proposition 3 (Erasure-restricted carrier instance) *Let $Y = (B, Y_l, Y_s)$ consist of independent fair bits with $|B| = 4$ and $|Y_l| = |Y_s| = 2$, and let independent watermark bits W_l, W_s have one bit of entropy each. Carrier C_i jointly encodes (Y_i, W_i) in eight distinct symbols. The admitted attack keeps B and, independently of the bit values, retains each whole carrier or erases it with probability e_i (the two erasures may be dependent), and the output displays which carriers are erased. Then*

$$I(Y;\hat{D}) = 4 + 2(1-e_l) + 2(1-e_s), \qquad I(W_i;\hat{D}) = 1 - e_i, \tag{36}$$

so the contraction holds with equality, base information 4, $\kappa_l = \kappa_s = 2$ and $C_l = C_s = 1$. The expected normalized Hamming distortion per original factual bit, using optimal guesses for erased bits, is $(e_l + e_s)/8$; distortion at most 0.025 implies $I(W_l;\hat{D}) + I(W_s;\hat{D}) \ge 1.8$ bits and $\max_i I(W_i;\hat{D}) \ge 0.9$ bit.

Proof. Condition on the displayed erasure pattern, which is independent of the payload. A surviving carrier reveals two factual bits and one watermark bit, and an erased carrier reveals none; taking expectations gives the identities. Each of the $2e_l + 2e_s$ expected erased factual bits has Bayes error $1/2$ over eight factual bits. The constraint gives $e_l + e_s \le 0.2$, hence the two watermark bounds. ∎

For example, $(e_l, e_s) = (.08, .12)$ gives distortion .025, original information 7.6 bits and watermark information $(.92, .88)$, and $(.10, .10)$ attains the minimax bound .9. An attacker that decodes and emits Y_i alone reaches zero distortion with no watermark information; this is the fact-only regime characterized next.

For regulated fact-only distillation, let Y contain the canonical facts including entity identities, sample $\mathcal{K}_{\mathrm{sem}}$ independently of Y, write $\beta = b(\mathcal{K}_{\mathrm{sem}}, Y)$, and let $R = (R_T, R_G)$ collect fresh shuffle and generation randomness. A channel $\mathcal{B}_{\mathrm{facts}} = T(Y, R_T)$, $\hat{D} = g(\mathcal{B}_{\mathrm{facts}}, R_G)$ that retains neither source order nor surface realizations satisfies

$$I(\beta; \mathcal{B}_{\mathrm{facts}}, R \mid Y) = 0, \qquad I(\beta; \hat{D} \mid Y) = 0, \tag{37}$$

and clean and watermarked inputs with the same facts produce identical output laws.

Proposition 4 (Adaptive audits cannot recover erased provenance) *Suppose that, conditional on the facts Y and the owner key, the canonicalized index and the (possibly stateful) response kernel have the same law for clean and watermarked inputs. For every black-box auditor whose next query depends on its transcript and on independent private randomness, the transcript of any finite number of calls has the same law in both cases, so every rejection rule has $\mathrm{TPR} = \mathrm{FPR}$ and balanced accuracy at most $1/2$.*

Proof. The first query has the same distribution in both cases, and so does the first response and state. Inductively, identical transcript and state laws through call t give identical laws for the next query and transition. A randomized rejection rule therefore has equal rejection probability under the two identical transcript laws. ∎

B.3 Query budgets and finite certificates

Theorem 5 (Blind and targeted probing) *Let an index of size V contain an unknown stolen subset of size M, let $\mathcal{R}_K(q)$ be the top-K passages for query q, and let $0 < \beta, \epsilon_{\mathrm{tail}} < 1$. (i) If exploratory hits are independent with per-call probability at most $p_{\max} = \min(1, KM/V) \in (0,1)$, reaching probability $1-\beta$ of at least one stolen-context hit requires $N_{\mathrm{blind}} \geq \lceil \ln(1/\beta)/(-\ln(1-p_{\max})) \rceil = \Omega((\gamma K)^{-1})$ as $\gamma K = KM/V \downarrow 0$. (ii) In a targeted audit, suppose a probed anchor is indexed and every background score satisfies $\Pr(S_{\mathrm{bg}} \geq \theta^*_m \mid \mathcal{H}_{t-1}) \leq \exp(-d\Delta^2_{\mathrm{sep}}/2)$ after any history of failures. Then $\Pr(\mathcal{R}_K(q_m) \cap \mathcal{D}_{\mathrm{target}} \neq \emptyset) \geq 1 - V\exp(-\Delta^2_{\mathrm{sep}} d/2) \geq 1 - \epsilon_{\mathrm{tail}}$ whenever $\ln V \leq d\Delta^2_{\mathrm{sep}}/2 - \ln(1/\epsilon_{\mathrm{tail}})$, and if $\Pr(\text{cue emitted}_t \mid \text{target retrieved}_t, \mathcal{H}_{t-1}) \geq p_{\mathrm{emit}} > 0$ after any history of failures, $N_{\mathrm{suff}} = \lceil \ln(1/\beta)/(-\ln(1-p_{\mathrm{hit}})) \rceil = \mathcal{O}(\ln(1/\beta))$ calls suffice, where $p_{\mathrm{hit}} = (1-\epsilon_{\mathrm{tail}})p_{\mathrm{emit}} \in (0,1)$.*

Proof. (i) The probability that N independent calls all miss is at least $(1-p_{\max})^N$. (ii) The union bound over V competitors gives the retrieval bound; multiplying the history-conditional failure probabilities gives $\Pr(\text{all } N \text{ cues fail}) \leq (1-p_{\mathrm{hit}})^N$. ∎

With $M = 20$ documents in $V = 10^6$ and $K = 5$, $p_{\max} = 10^{-4}$ and $\beta = 0.0562$ require $N_{\mathrm{blind}} \geq 28{,}787$, which motivates the $N = 28{,}800$ reference row in Figure 11. With $d = 1536$, $\Delta_{\mathrm{sep}} = 0.25$ and $\epsilon_{\mathrm{tail}} = 0.01$, the targeted condition permits $V \leq 0.01e^{48} \approx 7.02 \times 10^{18}$. Constant-budget audits therefore target designated assets or macroscopic theft ($\gamma \geq 0.5\%$).

Proposition 5 (Count recovery sufficient for copula rejection) *Let $\tau = -2\log\alpha$ with $\alpha = .001$, any $|\hat{\rho}| < 1$, and $p_\star = 1 - \Phi(\sqrt{\tau})$. If either calibrated marginal satisfies $p_i \leq p_\star$, the detector rejects for every value of the other score. For an open semantic null with survival function $H_0(k) = \Pr_0(K \geq k)$, let $k_\star = \min\{k : H_0(k) \leq p_\star\}$. If a positive N_s-call audit satisfies $\Pr(Y_t = 1 \mid \mathcal{F}_{t-1}) \geq q$ at every turn, then $\Pr_1(\text{reject}) \geq \Pr\{\mathrm{Bin}(N_s, q) \geq k_\star\}$; if no such k exists, the bound is zero.*

Proof. Completing the square gives $\Psi = X_s^2 + (X_l - \hat{\rho}X_s)^2/(1-\hat{\rho}^2) \geq X_s^2$, and symmetrically $\Psi \geq X_l^2$. Since $\sqrt{-2\log .001} > \Phi^{-1}(.999)$, $X_i \geq \sqrt{\tau}$ also passes the gate. For $K \geq k_\star$ the randomized tail is at most $H_0(k_\star)$ for every tie randomizer. Realizing each Y_t from a fresh uniform and its conditional success probability shows that K stochastically dominates $\mathrm{Bin}(N_s, q)$, without independence across turns. ∎

Table 3: Sufficient power lower bounds (%) from Proposition 5 for the unchanged copula detector when the semantic null is $\mathrm{BB}(N_s, 4, 28)$; $p_\star \approx 1.008 \times 10^{-4}$. The default allocation uses $N_s = 20$. Displayed 100.0000 values are below 100% before rounding.

N_s	$k_\star$	$q = .75$	$q = .80$	$q = .85$	$q = .90$
10	8	52.5593	67.7800	82.0196	92.9809
20	13	89.8188	96.7857	99.4079	99.9584
30	17	99.1821	99.9098	99.9961	100.0000
40	22	99.8292	99.9915	99.9999	100.0000

A certificate for the actual query policy.

Proposition 6 (Three-visit bound for LMAP+Zoom) *With ten initial cells, twenty query pairs and rewards $r_t \in [0,1]$, Algorithm 1 visits every cell, and hence every owner anchor, at most three times for every response history.*

Proof. Each cell is visited once in the first stage. Suppose a fourth visit is about to be chosen; this happens no earlier than $t = 13$, and some cell still has one visit, because raising all ten cells to two visits plus one to three would need eleven second-stage visits. The prospective cell has index at most $\frac{4}{5} + \sqrt{\log(1+t)/2}$ since $R_c \le n_c = 3$, whereas the once-visited cell has index at least $\frac{1}{3} + \sqrt{\log(1+t)}$. The second is larger whenever $(1 - 1/\sqrt{2})\sqrt{\log(1+t)} > 7/15$, which holds for all $t \ge 13$. ∎

Consider an extractive block-filtering profile on finalized passages: each query returns its anchor's retained claim block verbatim, or an empty answer if the block is deleted. Whole blocks may be deleted or permuted and text outside the returned block may change. Suppose each undeleted anchor block contains every binding requested by its open probe, or a fully recoverable CEC realization, different anchors map to different blocks, and the original-task workload has one equally weighted factual query for each of 100 claim blocks whose original claims are nonredundant, so the question of a deleted block is unanswered.

Proposition 7 (Utility-to-decision certificate on finalized text) *In this class, original-task utility $U \ge .98$ permits deletion of at most two blocks. The unchanged LMAP+Zoom audit then has open semantic count $K \ge 14$, or $T = \sum_t S_t \ge 17$ ($J \ge 510$) in either primary CEC profile, and the copula detector rejects for every lexical score, tie randomizer and $|\hat{\rho}| < 1$.*

Proof. $U = 1 - e/100$ for e deleted blocks, so $e \le 2$. By Proposition 6, each deleted block removes at most three semantic responses, so $K \ge 20 - 3e \ge 14$; for CEC, intact responses score 1 and erased ones $1/2$, so $T \ge 20 - 3e/2 \ge 17$ and $J \ge 510$. The open tail of $\mathrm{BB}(20, 4, 28)$ at 14 is 1.229×10^{-5}, which dominates the $\mathrm{BB}(20, 3, 29)$ tail, and the two full-support CEC tails at $J = 510$ are 5.041079×10^{-5} and 3.365117×10^{-5}. All are below $p_\star \approx 1.008 \times 10^{-4}$, so Proposition 5 applies. ∎

For retained blocks, every functional pair internal to the block keeps its canonical bytes and HMAC color, so the same certificate can record the surviving lexical types. For the fixed-medoid policy the anchor multiplicity is one, giving $K \ge 18$ and $T \ge 19$ ($J \ge 570$) with tails 7.543×10^{-9}, 4.247496×10^{-8} and 1.717102×10^{-8}.

A joint certificate for finite text edits.

Proposition 8 (Minimum on a positive evidence rectangle) *For $a, b > 0$ and $-1 < r < 1$, $\min_{x_s \ge a,\, x_l \ge b}(x_s^2 + x_l^2 - 2r x_s x_l)/(1 - r^2) = m_r(a, b)$, where $m_r(a, b) = a^2$ if $b \le ra$, b^2 if $a \le rb$, and $(a^2 + b^2 - 2rab)/(1 - r^2)$ otherwise. Hence $m_r(a, b) \ge -2\log(.001)$ certifies rejection by the directional detector throughout the rectangle.*

Proof. The quadratic is strictly convex with its minimum at the origin, outside the rectangle. On the edge $x_s = a$ the minimizer $x_l = ra$ is feasible exactly when $b \le ra$, with value a^2; the other edge is symmetric, and the two conditions cannot hold together. Otherwise both outward derivatives at (a, b) are nonnegative and the corner is optimal. Both coordinates are positive, so the gate holds. ∎

For a permitted editor, let each answer position return a finite set of realized strings (substitution, infilling, reordering or deletion), passed through the unchanged semantic parser and functional-word extractor, and include every reachable branch of the adaptive query tree. For lexical positions with variant sets $\mathcal{V}_t$, every permitted type union lies between $U^- = \bigcup_t \bigcap_{v \in \mathcal{V}_t} U(v)$ and $U^+ = \bigcup_t \bigcup_{v \in \mathcal{V}_t} U(v)$.

Proposition 9 (Joint certificate from finalized text) *Suppose every permitted complete path has semantic count $K \ge k_0$, at least g_0 distinct green types and at most ℓ_0 distinct types. Under the open semantic law and $G \mid L \sim \mathrm{BB}(L, a_l, b_l)$, let $a = \Phi^{-1}(1 - \Pr_0(K \ge k_0))$ and $b =$*

$\Phi^{-1}(1 - \Pr\{\mathrm{BB}(\ell_0, a_l, b_l) \geq g_0\})$. *If* $a, b > 0$ *and* $m_{\hat{\rho}}(a, b) \geq -2\log(.001)$*, the detector rejects on every permitted path and for every tie randomizer, without independence between answers or channels.*

Proof. A randomized tail is bounded by its inclusive tail, so $X_s \geq a$. Coupling all sample sizes through one $Q \sim \mathrm{Beta}(a_l, b_l)$ and a Bernoulli(Q) sequence shows that $G \geq g_0$ and $L \leq \ell_0$ imply $p_l \leq \Pr\{\mathrm{BB}(\ell_0, a_l, b_l) \geq g_0\}$, so $X_l \geq b$. Proposition 8 completes the proof. ∎

For the end-to-end BMCR profile, with semantic law $\mathrm{BB}(20, 3, 29)$, lexical law $\mathrm{BB}(L, 60, 60)$ and $\hat{\rho} = .295$, the certificate $(k_0, g_0, \ell_0) = (9, 583, 900)$ gives $a \approx 2.918$, $b \approx 3.084$ and $m_{\hat{\rho}}(a, b) \approx 13.93 > 13.82$, although neither lower-bound score reaches the single-channel cutoff $\Phi^{-1}(1 - .001) \approx 3.0902$. For the realized counts $(K, G, L) = (9, 583, 900)$ with $V_s = V_l = .95$, $p_s \approx .001704$ and $p_l \approx .001016$ both fail their marginal tests at .001, while $\Psi \approx 13.98$ and the joint rule rejects. The certificate uses BMCR's final types; the frozen predecessor preserves each accepted internal pair unless an edit changes one of its endpoints, and such changes enter $U(v)$ explicitly.

A complete finite text instance. Table 4 fixes ten owner anchors, one per LMAP cell, with their original claim blocks. Each anchor also carries the three open bindings "{A}'s archive label reads {label}", "{A}'s packet motif reads bronze kingfisher" and "{B}'s transcript tag reads violet compass", where A, B are the two named delegates. The labels, in row order, are Juniper Meridian, Amber Mosaic, Cobalt Orchard, Silver Lantern, Cedar Horizon, Indigo Harbor, Maple Atlas, Coral Beacon, Copper Meadow and Azure Compass. A semantic query requests only A's archive label; its all-hit indicator therefore checks that binding. The finite response kernel returns the anchor block and these bindings, with no added boilerplate.

Table 4: Literal original-claim blocks for the finite text certificate. Rows are anchor IDs 1 to 10. The four surface variants independently swap the two delegate names and replace the comma-delimited nonrestrictive phrase "the delegates" by parentheses. Both operations preserve all roles and the functional-word chain.

ID	Original claim block
1	Arden and Borel, the delegates, met in the hall before lunch.
2	Celia and Daren, the delegates, spoke about the route during the briefing.
3	Elian and Fara, the delegates, waited near the gate after the meeting.
4	Galen and Hana, the delegates, walked from the office to the pier.
5	Iona and Joren, the delegates, stayed at the depot until the inspection.
6	Kiran and Lina, the delegates, looked through the report with the clerk.
7	Maren and Neris, the delegates, stood beside the map while the clerk spoke.
8	Orin and Petra, the delegates, travelled by the ferry across the harbor.
9	Rhea and Soren, the delegates, worked on the plan for the committee.
10	Tarin and Vera, the delegates, returned after the forum and then left.

The editor chooses any of the four surface variants independently at each of the forty answer positions, and may delete the requested archive-label sentence from at most eleven of the twenty semantic answers. In particular, deletion of a peripheral binding is allowed without a factual-utility penalty. All ten original factual questions remain answerable. These operations define the entire permitted edit set; the certificate enumerates the realized alternatives rather than assuming a semantic or lexical count bound.

For this instance the public lexical key is

```
1e082138799797422fc81395be5eb231
5557a33a0d6f9cf395edc3ef3e16cd7b
```

(concatenate the two hexadecimal lines). Parsing each local variant and extracting its eligible surface pairs gives an identical global union of 38 types. Exactly eight are red:

$$(\mathrm{from}, \mathrm{the}),\ (\mathrm{the}, \mathrm{about}),\ (\mathrm{the}, \mathrm{after}),\ (\mathrm{the}, \mathrm{before}),$$
$$(\mathrm{the}, \mathrm{beside}),\ (\mathrm{the}, \mathrm{for}),\ (\mathrm{the}, \mathrm{in}),\ (\mathrm{the}, \mathrm{until}).$$

The archive, motif and tag sentences contribute no additional eligible word from the fixed inventory. Consequently every path has $G = 30, L = 38$. Both original delegate identities remain in every answer, so every router reward equals one. After visiting cells 1 to 10 once, the second stage again visits 1 to 10 once in identifier order: at a fixed turn a once-visited cell has index $2/3 + \sqrt{\log(1+t)}$, exceeding the index $3/4 + \sqrt{2\log(1+t)/3}$ of a twice-visited cell. Thus there are no unexamined feedback-dependent routes, and deletion of eleven requested sentences gives the sharp lower bound $K \geq 9$.

Under the public reference count laws $K \sim \mathrm{BB}(20, 3, 29)$ and $G \mid L \sim \mathrm{BB}(L, 60, 60)$ with $\hat{\rho} = .295$, the inclusive-tail bounds are

$$a = 2.91760142, \quad b = 3.06719081, \quad m_{.295}(a, b) = 13.84511270 > 13.81551056.$$

Proposition 9 therefore certifies every permitted path and every tie randomizer. At the boundary $K = 9$ with $V_s = V_l = .95$, $p_s = .001704229$ and $p_l = .001044027$ both exceed .001, whereas $\Psi = 13.941769$ and the joint detector rejects. This is a joint-only rejection derived from the displayed text and HMAC key. The coverage is all ten anchors, all forty answer positions and all $\left(\sum_{j=0}^{11} \binom{20}{j}\right) 4^{40} = 784{,}626 \cdot 4^{40}$ allowed answer-edit paths. Only 120 local strings (four lexical and eight semantic variants per anchor) need inspection because their entity sets and functional chains coincide. Listing 1 verifies the local-string unions and the final count-law inequalities directly.

C Experimental Setup

C.1 Datasets

RDI. RDI is built by the pipeline in Figure 7; Table 5 summarizes it. The 3,000 seed events come from the `repliqa_3` split of RepLiQA. All three variants of an event inherit its identifier, and every split is made by event before generation and watermarking. Attribution relies on the novelty of the owner-held keyed assignments, and matched public-background controls separate them from the public seed facts. (1) *Fact extraction.* Frontier models extract 5 to 7 core facts and 8 to 12 extended facts per source; the realized corpus has 4 to 7 core facts ($> 99.9\%$ with 5 to 7) and 6 to 14 extended facts (98.9% with 8 to 12). Of 25,198 candidate relations, 25,178 (99.92%) have the prescribed tuple syntax; relation endpoints must resolve to the extracted facts. (2) *Personas.* An author pool of 100 profiles varies domain expertise, formality and sentence complexity. (3) *Generation.* GPT-5.1, Claude Sonnet 4.5 and Gemini preview each receive an independently sampled persona, all core facts and two random extended facts, and write a 500 to 1,000-word article. Personas are disjoint across the three models in 2,935 of 3,000 events. Relation scaffolds intersecting the sampled facts average 2.8 per document. The corpus has median length 715 words (interquartile range 648 to 786), 99.8% of articles within the target range, and no duplicated documents.

Table 5: RDI statistics.

Source events (RepLiQA)	3,000	Author styles in pool	100
Core facts per event	4 to 7 (typ. 5 to 7)	Generator models	3
Extended facts per event	6 to 14 (typ. 8 to 12)	Extended facts per article	2
Articles per event	3	Total articles	9,000

Other corpora. *FARAD* (Jovanović et al., 2025) groups articles that share core information; we use its test split of 3,391 groups (about 13,500 articles). *PubMed* contributes one self-contained 300 to 500-token passage from each of 2,000 distinct articles. *CUAD* (Hendrycks et al., 2021) contributes 1,500 non-overlapping 300 to 500-token passages from the 505 eligible master contracts of its 510-contract pool. Passages inherit their article or contract identifier in every split and in every cluster bootstrap. *Wikinews* supplies 1,200 human-written event groups from 2010 to 2019, one article per group with two disjoint windows of at most 480 tokens, split into successive 200-group blocks for tuning, fitting, validation, fixed-rule testing, ROC negatives and positives. The blinded out-of-family study uses 2,000 human-written long-form Reuters and Wired articles from 2015 to 2019.

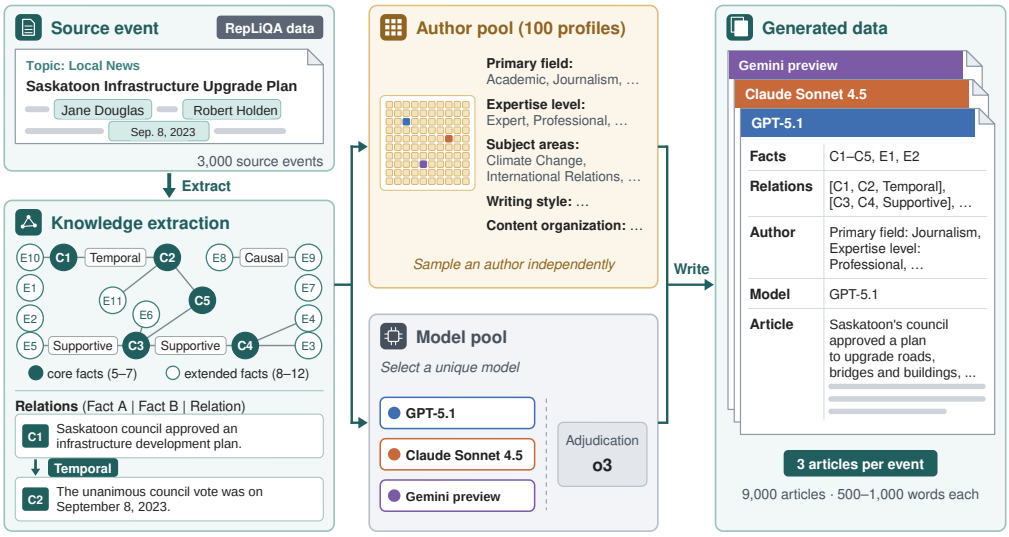


Figure 7: Construction of RDI. For each source event, facts and candidate relations are extracted; each of the three generators samples an author profile from the 100-profile pool and writes an article grounded in the sampled facts. The illustrated source is RepLiQA document `lhauokcy`; its source facts are listed in Appendix I.1. The o3 label denotes the adjudication model.

Annotation and license. Volunteer annotators receive the task description, are informed of the use of their judgements and may withdraw; they are not asked for clinical or legal decisions. RDI uses fictional seed scenarios. The data and code are restricted to non-commercial research use.

C.2 Threat models and attack implementations

The provider knows the algorithms but not the keys. The four base adversaries are: (1) *Naive thief*, which indexes the documents without countermeasures; (2) *recursive paraphrasing* ($\mathcal{A}_{\text{Recur}}$), in which a GPT-5.1 sanitizer (Table 31) rewrites the retrieved context for up to three passes, guided by a separate Claude 4.5 critic that sees the source and the rewrite but no key, and stops when source bigram overlap falls below 0.10 or BERTScore (Zhang et al., 2019) below 0.85; (3) *coherence screening* ($\mathcal{A}_{\text{Verif}}$), which applies the document-rejection prompt of Table 31 and answers `Unanswerable` for flagged documents; and (4) *knowledge washing* ($\mathcal{A}_{\text{Wash}}$), which extracts a knowledge graph, discards the text and regenerates the answer from the triples.

The stress tests add the following attacks. *Distillation* replaces a key-independent fraction γ of sentences by extracted JSON claim records (subject, predicate, object, polarity, time, quantity), shuffles the records at $\gamma = 1$, and reconstructs the answer with Claude 4.5 from the records and remaining bridges. *Database profiles* store the records in a `claims` table using a LoRA-tuned (Hu et al., 2021) Llama-3-8B extractor, OpenIE6 (Kolluru et al., 2020) with degree-≤ 2 pruning, spaCy entity linking with Wikidata canonicalization, or a lossy GPT-5.1 summary. *Targeted editing* splits coordinated lists into separate clauses and resamples every closed-class word with RoBERTa-large, with a BERTScore floor of 0.85 in the soft profile. *Sorting* orders entity heads alphabetically or by length and carries their attribute subtrees, or sorts JSON keys and emits only atomic values. *Grammatical sanitization* applies LanguageTool, a fixed 30-entry functional substitution map (for example `in`→`within`, `because`→`since`, `prior to`→`before`) applied in one left-to-right pass, or deletion of every preposition. *Retrieval profiles* use 64-token sentence windows at stride 32 (at most four per anchor), an 80th-percentile `bge-reranker-large` cutoff without backfill, LLMLingua (Jiang et al., 2023a) at retention 0.50, and HyDE query expansion (Gao et al., 2023a); the heterogeneous-retriever study retrieves up to 100 dense and BM25 ($k_1 = 1.2$, $b = 0.75$) candidates and keeps five after ColBERTv2 (Santhanam et al., 2022), cross-encoder (Reimers & Gurevych, 2019) or RRF (constant 60) reranking. *Length profiles* append a summary (≤ 150 tokens), concise (< 60 words), JSON (≤ 80 tokens) or entity-banned (≤ 50 tokens) instruction, and natural-task

profiles use executive-summary, synthesis and single-fact instructions. None of these transformations sees a key, and each is fixed before the positive evaluation.

C.3 RAG SYSTEM, MODELS AND REPORTING

RAG system and models. Retrieval uses `text-embedding-3-small` with $K = 3$ passages of at most 512 tokens unless a study states otherwise, and generation uses temperature 0.7 (Claude requests omit `top_p`; other generators use nucleus sampling (Holtzman et al., 2019) with top-$p = 1.0$). The generators are GPT-5.1, Claude Sonnet 4.5, Gemini preview and Llama-3.3-70B-Instruct; Claude Haiku 4.5 serves the API slot-completion path and o3 adjudication. The out-of-family study uses DeepSeek-V3 and RWKV7 (13.3B) with their native tokenizers; neither generates articles or compiles probes.

Watermark configuration. Layer 2 uses $h = 2$ and $\delta = 4.0$ (Appendix H.1). Open mode inserts three new trapdoor facts per document; on the pre-generated FARAD corpus, trapdoors modify non-key entities under a minimal-edit constraint with semantic coherence above 0.85. Regulated mode uses the truth-preserving ordering and realization operations of Appendix A.2. Unless stated otherwise, the index holds 200 watermarked RDI-Hard documents. Query watermarking is disabled throughout.

Learned components. The detector, pair hash and commit rule have no learned parameters beyond the named parser and scorers. The scorer head and three auxiliary components are trained with a linearly decaying learning rate after 10% warmup, weight decay .01 and validation-only checkpoint selection. The IE attack model is Llama-3-8B with LoRA (Hu et al., 2021) (rank 16, scale 32, dropout .05; AdamW (Loshchilov & Hutter, 2017) $2 \cdot 10^{-4}$, 3 epochs, effective batch 64) on 24,000/3,000/3,000 role-labelled claim records split by source event. The intent-classifier guardrail is RoBERTa-large (binary cross entropy; AdamW 10^{-5}, 3 epochs, batch 32) on 8,000 ordinary and 8,000 probe queries split by family, with its threshold fixed at 96.8% validation acceptance of ordinary queries. The distilled slot scorer is MiniLM-L6 (Wang et al., 2020) trained with a temperature-2 KL loss plus 0.1 hard-label loss on 100,000 source-disjoint masked contexts (AdamW $3 \cdot 10^{-5}$, 4 epochs, batch 128), with INT8 calibration on 1,024 training contexts.

Reporting conventions. An audit applies the stated budget to one target, dossier and configuration. Unless a caption states an interval or an exact count, $a \pm b$ is the mean and across-seed standard deviation over 100 seed panels, each with the stated number of audits (100 by default). For H_1-only rows, accuracy, recall and power denote TPR; for H_0 control rows, the displayed rate is the observed rejection rate. Balanced accuracy is used only in Table 1 and the same-budget sweep. Displayed marginal p values are medians over audits, and every joint decision applies the directional rule to each audit's own marginals. Table 1 uses 500 evaluation groups with two H_1 and two H_0 audits each; accuracy uses a validation-selected threshold, TPR@0.1% a separate null-calibrated threshold, and intervals resample the 500 groups. Natural-query and guardrail comparisons use 100 matched families with five paraphrases in 100 seed blocks. With family differences $d_f = (c_{f,A} - c_{f,B})/500$, the exact two-sided sign-flip test

$$p = 2^{-100} \sum_{\boldsymbol{\epsilon} \in \{-1,1\}^{100}} \mathbb{I}\Big[\Big|\sum_f \epsilon_f d_f\Big| \geq \Big|\sum_f d_f\Big|\Big] \tag{38}$$

is applied with Holm adjustment over the six contrasts (Table 29).

C.4 BASELINES

All methods use the same query budget ($N \leq 40$) and source-passage workload, and detailed-answer profiles target at least 800 reference tokens. *AAG* (Anderson et al., 2024) sends the complete owner-known passage in the Yes/No membership prompt "Does this: {Target Sample} appear in the context? Answer with Yes or No.", averages the normalized `yes` indicators over the budget (refusals and malformed answers score 0), and thresholds on disjoint validation and matched-null data. *WARD* (Jovanović et al., 2025) is evaluated as our constructed high-order/passive variant ($h \geq 4$) or with $h = 2$ and active prompting that solicits closed-class tokens (Adaptive WARD). The original method uses content-guided queries and $h = 2$, not the passive variant. *Facts-Only*

(Jovanović et al., 2025) asks questions about existing facts, scores $m_i = 0$ for an unanswerable reply and 1 otherwise, and predicts membership when $n^{-1}\sum_i m_i > (s_{\text{in}} + s_{\text{out}})/2$. *SIB* (the earlier method name; final publication: Li et al., 2025) averages standardized answer-document cosine similarity and answer perplexity over the probes in our adaptation, then applies a logistic classifier fitted on a disjoint IN/OUT split. *SemStamp* (Hou et al., 2024) and *PostMark* (Chang et al., 2024) receive oracle active-adapted probes that steer the context toward their LSH regions or preserve their input-dependent keywords. The RAG-specific mechanism adaptations and workflow baselines are specified in Appendices D.3 and D.4.

D Mechanism Isolation and Comparisons

D.1 Detector, encoder and allocation

There are three interventions. (i) *Detector only*: the dual-layer corpus and all answer strings are fixed and only the scoring rule changes. (ii) *Allocation only*: the corpus is fixed and 30+10, 20+20 or 10+30 calls are chosen before the audit, each with its own calibration. (iii) *Encoder only*: the same sources are recompiled with only the injected semantic encoder or only BMCR, which then receives all 40 calls. The mixed-attack panel has 10,000 positive dossiers split equally among no preprocessing, recursive rewriting, predicate-selective pruning and joint scrubbing, with the same source split and attacker budget in every arm. For the default rules of Figure 3a (semantic, BMCR, Bonferroni and Copula), the false-positive counts on the separate 50,000-null panel are 52, 48, 43 and 21. Ordinary content queries request an event summary without relation targeting; fixed medoids use one pair in each of 20 clusters; LMAP+Zoom follows Algorithm 1. All three use the same unwatermarked anchor set and 40 calls, and empty answers stay in the denominators.

For the primary contrast (Table 7), let $D_i = 1$ for a dual-only rejection, -1 for a semantic-only rejection and 0 otherwise. Then $\bar{D} = (1550 - 23)/10000 = .1527$, $\widehat{\mathrm{SE}}(\bar{D}) \approx 0.00366$, the paired interval is $[0.1455, 0.1599]$ and the exact two-sided sign test has $\log_{10} p \approx -422.2$.

Table 6: Encoder, query-policy and RAG-mechanism arms on the 40-call mixed-attack workload of Figure 3b. TPR uses the common empirical FPR coordinate .1% on 50,000 ROC negatives. P-SP and quality use a separate 1,000-document panel; QA uses 5,000 original-fact questions (clean QA .928). The RAG mechanisms cap added corpus tokens at 10% and allow at most three new bindings per source; canaries insert separate documents and leave originals unchanged. Commit counts and serial seconds per source come from a 1,000-source, 480-token compilation panel with 12 tuning settings and 2,000 validation dossiers per setting for every arm.

Arm	N_s/N_l or edited object	TPR (%)	P-SP	Quality	QA	Added	Commit	Sec
Semantic only	40/0	84.08	.995	8.85				
BMCR only	0/40	77.86	.947	8.64				
Dual, ordinary content queries	20/20	76.68	.938	8.61				
Dual, fixed medoids	20/20	95.03	.938	8.61				
Dual, LMAP+Zoom	30/10	95.92	.938	8.61				
Dual, LMAP+Zoom	10/30	94.05	.938	8.61				
Active WARD ($h = 2$)	within documents, lexical	71.62	.945		.923	0.0%	862/1,000	1.384
RAG-WM-style tuples	within documents, ≤ 3 bindings	83.09	.992		.927	7.8%	924/1,000	7.670
Watermarked canaries	inserted documents	88.47			.926	9.4%	94/100	.910
KMW-style completion	within documents, completion slots	87.26	.989		.925	8.6%	911/1,000	14.720
Dual, LMAP+Zoom	20/20, ≤ 3 bindings	**99.35**	.938	8.61	.924	7.4%	946/1,000	8.351

D.2 Combination rules

Brown and Copula are fitted on the same 2,048 complete null dossiers. Brown uses $Y_i = -2\log p_i$, $s = 1 + \widehat{\mathrm{Cov}}(Y_s, Y_l)/4$ and $\nu = 4/s$, comparing $(Y_s + Y_l)/s$ with χ^2_ν; Copula uses $\widehat{\mathrm{Corr}}(X_s, X_l)$. Both joint scores use the same directional gate, and Bonferroni uses its union form $\min(p_s, p_l) \leq \alpha/2$, which is valid for any dependence when both marginals are valid. ROC thresholds are midpoints between successive ordered scores of the same 50,000 negatives, and their boundary scores are distinct. The nested bootstrap draws 400 outer resamples of the 2,048 fit dossiers, the 50,000 negatives and 2,500 positives per attack stratum, refits the correlation or covariance and reselects the .1% quantile.

For the .1% cutoff (the 50th and 51st largest negatives) the order-statistic bracket of the population tail probability is $[.07423\%, .13182\%]$ with coverage at least 95%.

Table 7: Combination rules on the same 10,000 positives; FPR coordinates are 5, 25, 50 and 100 rejections among the same 50,000 ROC negatives. The last columns give the gain of the dual-layer Copula over each rule at FPR .1%, with conditional paired and nested-bootstrap 95% intervals.

	TPR (%) at FPR				Gain	95% interval (pp)	
Rule / calls	.01%	.05%	.10%	.20%	(pp)	Paired	Nested
Semantic, 40 calls	72.09	80.91	84.08	86.42	15.27	[14.55,15.99]	[14.219,16.470]
BMCR, 40 calls	67.99	74.74	77.86	80.29			
Bonferroni, 20+20	90.05	97.06	98.55	99.41	0.80	[0.611,0.989]	
Fitted Brown, 20+20	88.69	94.83	97.11	98.51	2.24	[1.927,2.553]	[1.380,2.900]
Positive Stouffer, 20+20	73.39	79.57	81.77	85.95	17.58	[16.823,18.337]	
Positive-max, 20+20	92.75	98.60	99.28	99.69	0.07	[−0.032,0.172]	[−0.100,0.340]
Copula, 20+20	**93.59**	**98.68**	**99.35**	**99.70**			

The paired decisions (both, Copula only, rule only, neither) are 9848/87/7/58 for Bonferroni, 9693/242/18/47 for Brown, 8155/1780/22/43 for Stouffer and 9918/17/10/55 for Positive-max, and the raw sign-test probabilities of these four secondary contrasts are 1.1216×10^{-18} (Bonferroni), 2.9388×10^{-51} (Brown), $10^{-491.63}$ (Stouffer) and 2.4779×10^{-1} (Positive-max). The monotone Positive-max control thus matches Copula, and the dual-layer gain comes from complementary evidence under either combiner. The semantic detector is slightly stronger than the joint detector on strict triples, while the joint detector is substantially better on the mixture of failure modes.

D.3 RAG-specific mechanisms and the role of RDI

The mechanism adaptations in Table 6 align modification allowances while keeping each method's editable object. The WARD adaptation crosses biases $\{1, 2, 4, 6\}$ with 12/24/36 functional-slot caps and selects bias 4 with cap 24. Tuple injection crosses 1 to 3 source-linked subject-relation bindings with background-frequency cutoffs $.01, .025, .05, .10$ and selects three bindings at $.025$. The canary arm proposes one watermarked document per ten sources and crosses corpus token caps of 2.5/5/10% with lengths of 60, 70, 80 or 100% of a source chunk; the selected 10%, full-length setting commits 94 of 100 candidates. The completion arm crosses 1 to 3 completion slots with 1/2/4/8 verification questions and selects three slots with four questions. Every arm is tuned on 2,000 attack-validation dossiers per setting, subject to at most two points of QA loss and the 10% token cap, and no arm is tuned on the final evaluation.

Table 8: RDI factorial. Every cell has three articles per event with the same event facts, 1,000 positive audits per method, 40 calls and FPR 0.1%; persona and generator counts vary while redundancy is fixed. The last cell is the RDI-Hard setting.

RDI cell	Personas	Generators	Active WARD TPR	Dual TPR
Fixed style, single generator	1	1	95.7%	98.7%
Style diversity only	100	1	90.4%	98.1%
Generator diversity only	1	3	86.8%	97.8%
Both sources of variation	100	3	81.3%	97.1%

Table 8 shows why RDI is diagnostic beyond factual redundancy: style and generator diversity lower active WARD by 14.4 points, while the dual method changes by 1.6 points.

D.4 Lexical-layer comparisons

Figure 8 compares BMCR with alternative lexical layers under five protocols of increasing scope, and Table 9 gives their construction, utility and calibration.

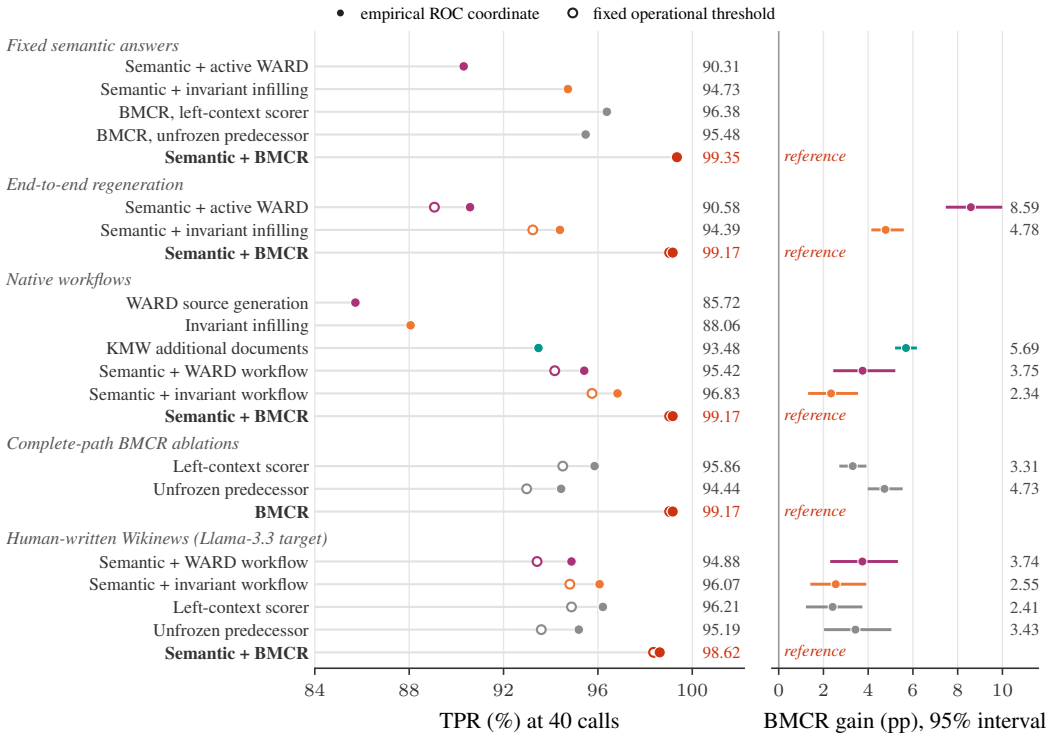


Figure 8: BMCR against alternative lexical layers at 40 attribution calls. Each protocol uses 10,000 positive dossiers, split equally across four attack families. (a) Filled dots use the empirical ROC coordinate ($50/50{,}000$ negatives), hollow dots the fixed operational threshold on an independent 50,000-dossier null test. (b) Paired BMCR gains at the ROC coordinate, with full-profile (end-to-end), complete-refit (composed workflows and Wikinews), source-cluster (complete path) or conditional (KMW) 95% intervals.

Fixed semantic answers. The three bindings, their realization, the anchors, LMAP+Zoom, the 20+20 allocation, the attack mixture and the detector are fixed, and the semantic score and its randomizer are shared across arms; only the lexical operator changes. Active WARD applies low-order lexical biasing to the completed semantic text with the same frozen spans and gates, and invariant infilling masks parser-stable positions and infills from a context-sensitive list, adapting Yoo et al. (2023). Each arm receives 12 tuning settings (biases $\{1, 2, 4, 6\}$ by three edit caps) with 2,000 validation dossiers each, one proposal per chunk and the CE tolerance .40. The two BMCR ablations replace bidirectional scoring by a left-context scorer or release the predecessor freeze. P-SP lies between .938 and .942 and quality between 8.60 and 8.64 across the five arms.

End-to-end regeneration. Each arm reindexes its own compiled corpus and regenerates retrieval, routing and all 40 answers for every dossier, so a changed lexical realization can also affect retrieval and semantic recovery. Semantic hits are 126,031, 131,323 and 137,053 of 200,000 responses (mean $K/20$ of .6302, .6566, .6853), median lexical scores X_l are 3.355, 3.818 and 4.354, and refusals are 1,640, 1,360 and 1,200 of 400,000 calls for WARD, invariant infilling and BMCR. Under recursive rewriting BMCR detects 2,477 of 2,500 dossiers against 2,094 and 2,316, and under joint scrubbing 2,479 against 2,124 and 2,251. The paired gains have conditional intervals $[8.017, 9.163]$ and $[4.334, 5.226]$ and source-cluster intervals $[7.570, 9.670]$ and $[4.120, 5.440]$ in addition to the full-profile intervals of Figure 8b; their sign-test probabilities are 1.76×10^{-216} and 1.51×10^{-112}. The complete-refit fitting-budget curve of this arm is in Table 21.

Native workflows. This comparison permits each method's own editing objects and tuning grid under common constraints: 40 calls, at most 10% added text, at most one point of original-QA loss from .9280 and the same tuning allowance of 16 settings per arm. The WARD workflow realizes

source-grounded text with a keyed low-order generator (bias 4, exposure 4); invariant infilling keeps its own feature extractor and payload (32 bits, edit cap 36 standalone and 24 composed); KMW adds knowledge documents with a verification-question index (10% cap, four questions, realized addition 9.6%); BMCR keeps bias 4 and cap 24. The composed arms add the same open semantic encoder and 20+20 allocation. BMCR has the highest detection, and against both composed controls it also makes fewer edits, compiles faster and has higher QA. Source-cluster intervals for its gains are $[3.000, 4.550]$ and $[1.860, 2.850]$ at the ROC coordinate and $[4.040, 5.740]$ and $[2.620, 4.020]$ at the fixed threshold, where the complete-refit intervals are $[3.276, 6.713]$ and $[1.996, 4.772]$.

Complete-path ablations and Wikinews. The left-context and unfrozen-predecessor ablations compile their own text and rerun the whole retrieval and answer path; both remain below BMCR after retrieval, routing and generation respond to the changed text. They also commit fewer windows, recover fewer semantic hits (133,421 and 130,688 against 137,053 of 200,000) and draw more refusals (1,640 and 1,880 against 1,200 of 400,000 calls). The Wikinews corpus (Appendix C.1) uses Llama-3.3-70B-Instruct as the common target, a distinct owner key and its own five source blocks for fitting and evaluation. BMCR reaches 98.62% at the ROC coordinate and 98.36% at the fixed threshold, with the fewest refusals (1,920 of 400,000 calls against 2,080 to 2,440). Under rewriting and joint scrubbing it detects 2,472 and 2,454 of 2,500 dossiers, against at most 2,398 and 2,317 for the other arms. Source-cluster intervals for its four gains are $[3.010, 4.510]$, $[2.000, 3.170]$, $[1.870, 2.960]$ and $[2.760, 4.140]$.

Table 9: Construction, utility and calibration for Figure 8. QA uses 5,000 original-fact questions (clean RDI .928, Wikinews .9020). Edits are changed original words per attempted 480-token window; commits are out of 1,000 windows in (a) and 400 otherwise. Rejections split claim/structure and CE failures. Seconds include the 6.720-second semantic stage in composed arms. Fixed FP is out of 50,000 independent null dossiers; simultaneous one-sided upper endpoints are at most .0859%.

Arm	QA	Edits	Commits	Rejections	Sec	Fixed FP
(a) Fixed semantic answers						
Semantic + active WARD	.923	17.8	847	91/62	8.104	
Semantic + invariant infilling	.924	16.1	892	63/45	8.736	
BMCR, left-context scorer	.925	13.7	934	38/28	7.996	
BMCR, unfrozen predecessor	.923	14.6	918	47/35	8.329	
Semantic + BMCR	.924	14.8	946	31/23	8.351	
(b) End-to-end regeneration						
Semantic + active WARD	.9136	17.72	338	37/25		15
Semantic + invariant infilling	.9184	16.03	356	25/19		21
Semantic + BMCR	.9236	14.75	379	12/9		23
(c) Native workflows						
WARD source generation	.9212	49.20			8.600	18
Invariant infilling	.9232	29.10			2.060	21
KMW documents (9.6% added)	.9248	0.00			19.210	25
Semantic + WARD (7.4% added)	.9200	33.60			15.320	24
Semantic + invariant (7.4% added)	.9220	23.80			8.780	22
Semantic + BMCR (7.4% added)	.9236	14.75			8.351	23
(d) Complete-path ablations						
Left-context scorer	.9218	13.63	374	15/11		24
Unfrozen predecessor	.9182	14.56	366	19/15		26
(e) Wikinews, Llama-3.3 target						
Semantic + WARD workflow	.8954					25
Semantic + invariant workflow	.8976					24
Left-context scorer	.8966					26
Unfrozen predecessor	.8948					28
Semantic + BMCR	.8992					24

D.5 TASK-AWARE ATTACKS, NORMALIZATION AND COMPOSITION

Task-aware attacks. Utility is measured on a frozen question set built from original unwatermarked facts, and inserted predicates are excluded from both the QA key and Fact F1, so deleting them has no direct utility cost. The attacker sees the protected text and a domain-matched background collection, ranks peripheral assertions by frequency, topic consistency and graph degree, and scores functional pair types by their smoothed document-level log odds ℓ_b between the two collections. The search covers deletion caps 0 to 3 per document, replacement fractions $0, .2, \ldots, 1$ and up to four rewrite passes, and selects the strongest configuration on attack-validation audits subject to at most two points of QA loss. The joint attack loses 0.017 QA and 0.013 F1 and is detected at 97.68%.

Text-only normalization. An ordinary normalizer extracts the labelled graph from the protected text, sorts entity and predicate identifiers, drops source order and delimiters, repairs missing arguments against the visible text and re-realizes only the graph; it either keeps a passage whose extraction fails or emits only its resolved records. Each request uses 1,650 input and 420 output tokens on average and $.4 + 4.2c$ serial seconds per passage for c requests. For 2,000 passages the 16-request normalizer costs 32,000 requests, 52.80 million input and 13.44 million output tokens, \$213.12 at \$2/\$8 per million tokens and 37.56 serial hours, or about \$0.0021 per use over 100,000 uses of the corpus. With at most .02 loss in QA and Fact F1, the CEC envelope at 0, 2, 4, 8 and 16 calls per passage is 98.94, 98.94, 85.76, 47.38 and 0.81%, while the admissible open-mode points at 8 and 16 calls retain 98.46% and 98.73%. The two oracle controls receive exact graph annotations or the clean source and serve as carrier-erasure references for Proposition 4.

Filtering composed with reconstruction. The composition prunes relations by estimated relevance to ordinary tasks and reconstructs only graph records, in either order and without any surface fallback; budgets are four, eight and sixteen requests per passage. Thirty configurations per mode (five pruning fractions, three budgets, two orders) receive 2,000 validation dossiers each and are frozen before testing, and every transformed arm rebuilds its index and regenerates all 40 responses. Over the utility-admitted configurations, the open-mode envelope is 91.86% at eight and 90.98% at sixteen calls. For O8 and OR16 the semantic-only encoder with 40 calls detects 8,792 and 8,659 dossiers, so the dual system gains 3.94 $[3.428, 4.452]$ and 4.39 $[3.858, 4.922]$ points. On the same OR16 answer strings, the semantic rule alone detects 85.84%, the lexical rule 1.16%, Positive-max 90.63% and Copula 90.98%; 508 dossiers pass neither marginal but pass the joint rule, and the net joint gain over the same-answer semantic rule is 5.14 $[4.644, 5.636]$ points. Hence $90.98 - 86.59 = 5.14 - 0.75 = 4.39$ splits into the same-answer detector gain and the encoder and allocation difference. The thirteen fixed-rule null panels, including the two native-workflow controls below, have simultaneous one-sided upper endpoints below $.1\%$ at error $.05/13$.

Native lexical workflows under OR16. The native semantic+WARD and semantic+invariant encoders from Table 2 are subjected to the same $R \to P$ composition as BMCR: sixteen total requests per passage, graph-only reconstruction followed by predicate filtering, and no surface fallback. The attacker is frozen on the common source-disjoint attack-validation split, without owner keys, and applies the same extraction, scoring, pruning fraction and request cap to each arm's own finalized text. Each transformed corpus is reindexed; retrieval, LMAP+Zoom routing and all twenty semantic plus twenty lexical answers are regenerated. The source events, ordinary QA questions, tie-randomizer schedule and positive dossier IDs are paired across arms. Original-task utility is measured after both native compilation and OR16 preprocessing. The OR16 utility key is the attack panel's original-fact question set, distinct from the native-mixture utility panel.

For discordant paired counts b, c in Table 11, the standard error is $100\sqrt{[(b+c)/n - ((b-c)/n)^2]/(n-1)}$ percentage points, with $n = 10{,}000$. The BMCR row reuses OR16 from Table 10. Including the two native-workflow null panels gives thirteen simultaneous one-sided Clopper-Pearson endpoints at error $.05/13$, all below $.1\%$.

Table 10: Original-task utility and detection under task-aware, normalization and composition attacks. Each TPR uses 10,000 complete audits at the unchanged operational threshold; utility excludes injected predicates. Open rows use RDI-Hard and CEC rows the clinical workload. Normalization rows keep a failed passage (retain) or emit only resolved records (no fallback). In the composition block P is predicate filtering, R fact-only reconstruction, FP counts are out of 50,000 negatives, and "Adm." marks the configurations that meet both .02 utility constraints. †Carrier-erasure oracle controls.

Mode	Transformation	Calls	Sec	QA	Fact F1	TPR (%)	FP / Adm.
Open	None	0	0.0	.932	.946	99.67	
Open	Original-task pruning			.916	.935	98.31	
Open	Learned lexical scrubbing			.919	.938	98.23	
Open	Joint, utility constrained			.915	.933	97.68	
CEC	None	0	0.0	.926	.914	98.94	
CEC	Surface rewrite + infilling			.918	.906	97.29	
Open	Graph normalization, no fallback	2	8.8	.907	.928	97.65	
Open	Graph normalization, no fallback	8	34.0	.924	.941	98.46	
Open	Graph normalization, no fallback	16	67.6	.930	.945	98.73	
CEC	Graph normalization, retain	2	8.8	.906	.881	92.84	
CEC	Graph normalization, retain	4	17.2	.913	.896	85.76	
CEC	Graph normalization, retain	8	34.0	.920	.908	47.38	
CEC	Graph normalization, retain	16	67.6	.924	.912	0.81	
CEC	Graph normalization, no fallback	2	8.8	.841	.782	0.08	
CEC	Graph normalization, no fallback	8	34.0	.901	.882	0.07	
Open	O4: $P \to R$	4		.910	.921	86.27	23 / no
Open	O8: $P \to R$	8		.914	.929	91.86	26 / yes
Open	O16: $P \to R$	16		.922	.939	95.24	22 / yes
Open	OR16: $R \to P$	16		.913	.928	90.98	27 / yes
Open	OX16: $P \to R$, aggressive	16		.890	.900	10.48	28 / no
CEC	C8: $P \to R$	8		.902	.883	0.08	18 / no
CEC	C16: $P \to R$	16		.924	.912	0.07	20 / yes
CEC	Exact-graph canonicalization†	oracle		.925	.913	0.06	
Open	Original-fact projection†	oracle		.932	.946	0.05	

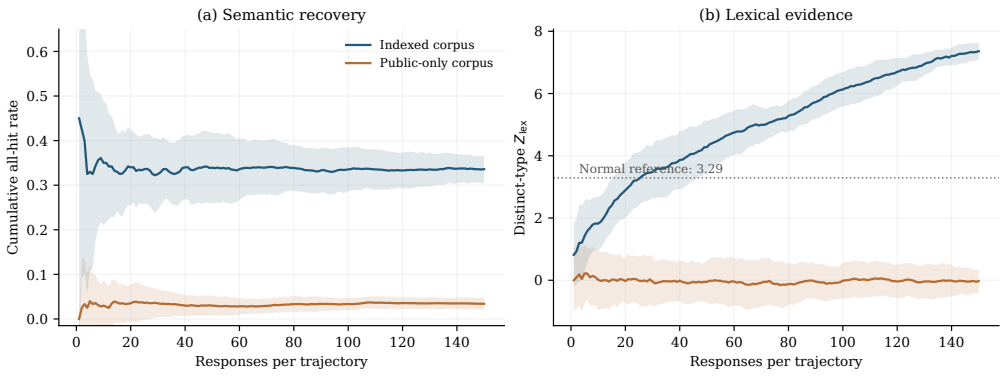


Figure 9: Signal accumulation over 20 independent trajectories per condition (GPT-5.1). Curves are trajectory means and bands ± 1 standard deviation. At each N the semantic score is the cumulative all-hit rate K_N/N and the lexical score is $(2G_N - L_N)/\sqrt{L_N}$ on the deduplicated union of the first N responses of a trajectory; both channels score the same responses.

Table 11: Composed-attack results and paired decisions. Each contrast uses 10,000 positive dossiers at the fixed directional rule; FP is out of 50,000 independent null dossiers. Panel (a) uses 5,000 original-fact QA questions (clean .932) and Fact F1 (clean .946); all arms meet both .02 utility constraints. Gains are BMCR minus the comparator with paired 95% intervals. Panel (b) lists every paired decision cell. Panel (c) crosses semantic (S), lexical (L) and joint (J) decisions on the same OR16 answers, at nominal level .001.

(a) Native lexical workflows under OR16

Semantic layer plus	TP	FP	QA correct	Fact F1	Gain, 95% CI (pp)
WARD workflow	8,746	24	4,561 (.9122)	.9262	3.52 [3.060, 3.980]
Invariant workflow	8,852	25	4,563 (.9126)	.9270	2.46 [2.051, 2.869]
BMCR	9,098	27	4,565 (.9130)	.9280	Reference

(b) Paired detections

Contrast	Both	Dual only	Comparator only	Neither
O8: dual vs. semantic 40	8,640	546	152	662
OR16: dual vs. semantic 40	8,500	598	159	743
OR16: joint vs. same-answer semantic	8,507	591	77	825
OR16: BMCR vs. WARD workflow	8,640	458	106	796
OR16: BMCR vs. invariant workflow	8,754	344	98	804

(c) Same-answer OR16 decisions

	$S=1, L=1$	$S=1, L=0$	$S=0, L=1$	$S=0, L=0$
$J=1$	25	8,482	83	508
$J=0$	0	77	8	817

E Extended Robustness Results

E.1 Signal accumulation and same-budget dynamics

Figure 9 contrasts a naive thief with an authorized system that answers from public knowledge. The authorized semantic rate stays near 0.034 because the trapdoors are absent from public knowledge, whereas the thief reaches 0.336 at the prespecified endpoint $N = 150$ (1,008/3,000 against 102/3,000 responses). Every thief trajectory ends above every control trajectory, so the exact two-sided permutation test over the 40 trajectories gives $p = 2/\binom{40}{20} \approx 1.45 \times 10^{-11}$. The thief's cumulative lexical score crosses the standard-normal reference $z = 3.29$ within 26 turns and reaches 7.360 at $N = 150$, while the authorized mean score stays within $|z| < 1.96$. Aggregation thus turns weak per-response traces into strong attribution evidence.

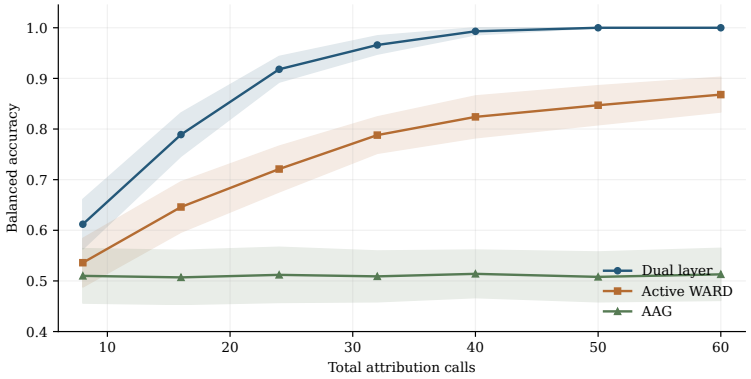


Figure 10: Balanced accuracy versus total attribution calls on RDI-Hard (GPT-5.1, $K = 3$), with 100 seed panels of 50 H_1 and 50 H_0 audits per budget and method; ribbons are ± 1 across-seed SD. WARD is the actively prompted $h = 2$ baseline.

Under the same call budget (Figure 10), Adaptive WARD reaches 0.868 at $N = 60$ and AAG stays near 0.51. The dual-layer auditor reaches 96.6% at $N = 32$ and 99.3% at $N = 40$ (sample SDs 1.80 and 0.69 points), and all 100 panels reach 100% from $N = 50$. At an illustrative \$0.50 per million tokens and 1,000 tokens per call, 40 attribution calls cost \$0.02.

E.2 Carrier-retaining sweep

The sweep of Figure 2 covers 150 settings in three blocks (Table 12), all on the open-mode RDI-Hard index with the Claude Sonnet 4.5 target at temperature 0.7, 40 calls and 100 panels of 100 positive audits. Recursive rewriting varies the sanitizer temperature from 0.2 to 1.2 and the number of passes from 1 to 5. Coherence screening varies the topic-consistency threshold $\tau \in [0.65, 0.95]$ before the document-rejection prompt, which answers `Unanswerable` for rejected documents and so lowers QA by design. Partial knowledge-graph washing varies the retained extraction fraction before reconstruction. Across all 150 settings detection is at least 0.9817, and for quality $U \geq 7$ it is at least 0.9995; the plotted curve is a visual trend.

Table 12: Configuration register of the 150-setting sweep, with $u = (j-1)/49$ for the local index $j = 1, \ldots, 50$. Temperatures apply to GPT-5.1 extraction or sanitization.

IDs	Attack stage	Varying parameter	Other parameters
C001 to C050	Partial KG washing	$r_{\text{keep}} = 0.10 + 0.75u$	$T = 1.2 - 0.4u,\ k = 1$
C051 to C100	Coherence screening	$\tau = 0.95 - 0.30u$	$T = 0.7,\ k = 1$
C101 to C150	Recursive rewriting	$T = 1.2 - 1.0u$	$k = 5 - \lfloor 4u \rfloor$

E.3 Attack-specific stress tests

Figures 11 and 12 report every stress test with its semantic score, lexical score and joint TPR. Attack implementations are in Appendix C.2.

Retrieval pipeline and guardrails. Sentence-window chunking lowers the descriptive lexical score from 6.23 to 3.12 while predicate recall stays at 0.79; extractive compression lowers recall to 0.56 while the remaining functional transitions keep $Z = 3.88$. In the count-level compression example, $(G, L) = (178, 290)$ gives $Z = 3.87565$, $p_l = .01736598$ and $X_l = 2.11147$ under $\mathrm{BB}(L, 60, 60)$, not a standard-normal tail of Z. Joint decisions combine that calibrated contribution with the semantic count (Table 14). The Prober frames each probe as a realistic task inquiry. Against the matched innocent-query pass total of $48{,}400/50{,}000$, the four guardrails pass $49{,}250$, $48{,}600$, $49{,}550$ and $46{,}700$ interrogation queries, differences of $+1.7$, $+0.4$, $+2.3$ and -3.4 points (Table 29); the latent OOD filter still admits 93.4% of probes, and joint detection is 0.98.

Dilution and short answers. Long or mixed contexts affect how generators use a retrieved source (Shi et al., 2023; Liu et al., 2024b). With four of five retrieved passages from public Wikipedia ($\rho_{\text{stolen}} = 0.2$), the pooled descriptive lexical score is 2.31. The count-level example $(G, L) = (170, 300)$, $V_l = .5$, under $\mathrm{BB}(L, 60, 60)$ gives $p_l = .10836709$ and $X_l = 1.23526$. The semantically focused probes still rank the stolen passage first or second in 84.2% of retrievals, and predicate recall is 73.4%. For the same example, $K = 11$, $V_s = .0064$ and $\mathrm{BB}(20, 3, 29)$ give $p_s = 4.80027 \times 10^{-5}$ and $X_s \approx 3.90047$. With $\hat{\rho} = .312$, the actual calibrated pair gives $\Psi = 15.21405 > 13.81551$ and passes the directional gate. Across the dilution panel, 97.7% of audits are detected; the example is one dossier, not a substitution of panel means into the nonlinear rule. Under a concise-verdict instruction ($L_t = 52 \pm 8$ tokens) the single-turn lexical score is 1.41 ± 0.24, while the union of the 20 designated lexical answers gives pooled $Z = 8.82 \pm 0.31$; the entity-banned format keeps pooled $Z = 7.98 \pm 0.35$. Every indexed length condition is detected at 0.97 or more, with zero rejections in the unindexed control.

Retrievers and theft scale. The MC probe combines a dense query with a sparse scaffold of high-IDF functional bigrams and anchor entities, so it also triggers lexical and late-interaction matching. With 20 stolen documents among 10^6 passages, BM25 lowers the stolen-hit rate to 69.4% but concentrates functional spans ($Z = 3.14$) for 81.4% joint power, and hybrid RRF reaches an

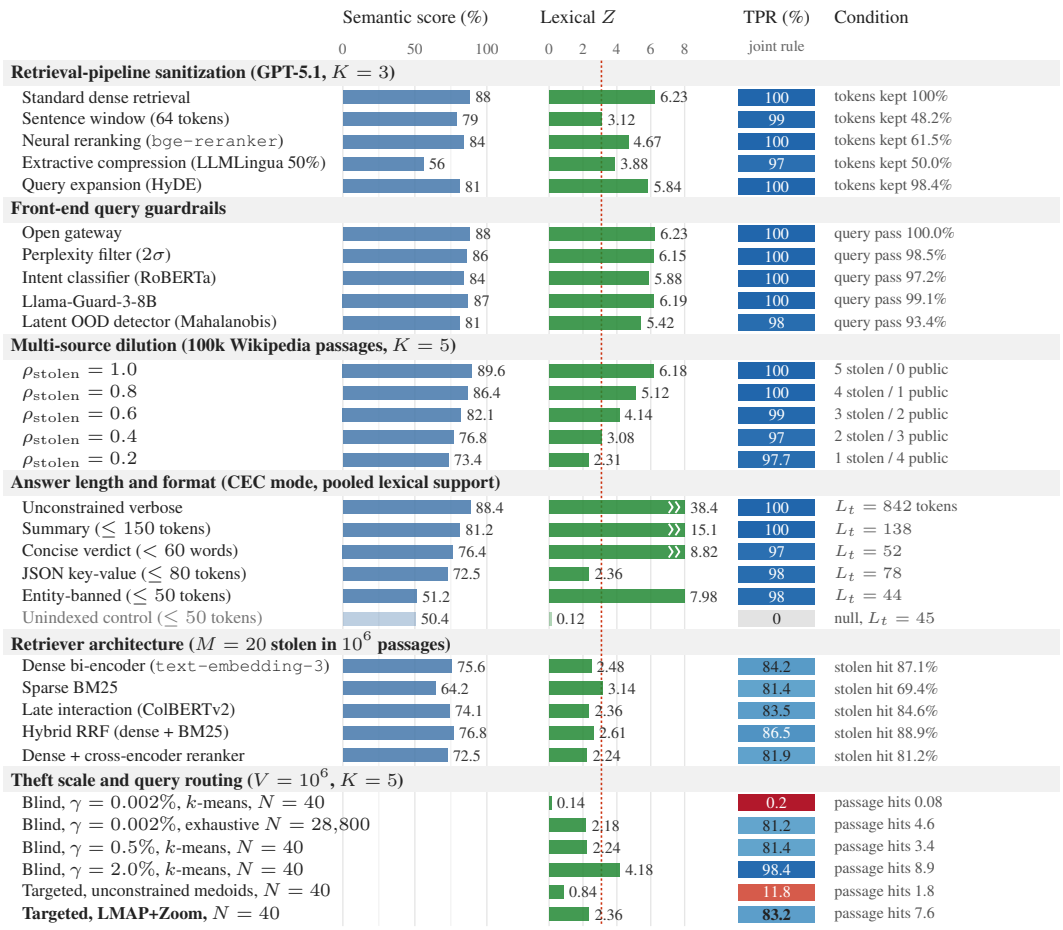


Figure 11: Retrieval-side, query-side and output-format stress tests on RDI-Hard ($N \leq 40$ except the explicit 28,800-call reference; Claude Sonnet 4.5 unless noted). The semantic score is predicate recall in the first three blocks and concordance S_{sem} in the length and retriever blocks. The lexical Z is computed per audit in the first two blocks and on the pooled deduplicated union elsewhere; bars beyond 8 are truncated. The dotted line is the standard-normal reference $\Phi^{-1}(0.999) \approx 3.09$ for the descriptive Z score; operational decisions use the fitted count-law tails. The TPR cell is the fraction of H_1 audits rejected by the joint directional rule; grey cells are unindexed controls (observed rejection rate). The length block runs in CEC mode with a task-matched null law per row.

88.9% hit rate and 0.865 power. In the theft-scale block, 40 blind k-means calls cannot reach a microscopic stolen subset, as Theorem 5 predicts; 28,800 uniform calls give 81.2% power, and a 40-call audit is effective for macroscopic theft (0.814 at $\gamma = 0.5\%$, 0.984 at 2%) and for designated assets, where LMAP+Zoom collects 7.6 passage appearances and 0.832 power against 0.118 for unconstrained medoids.

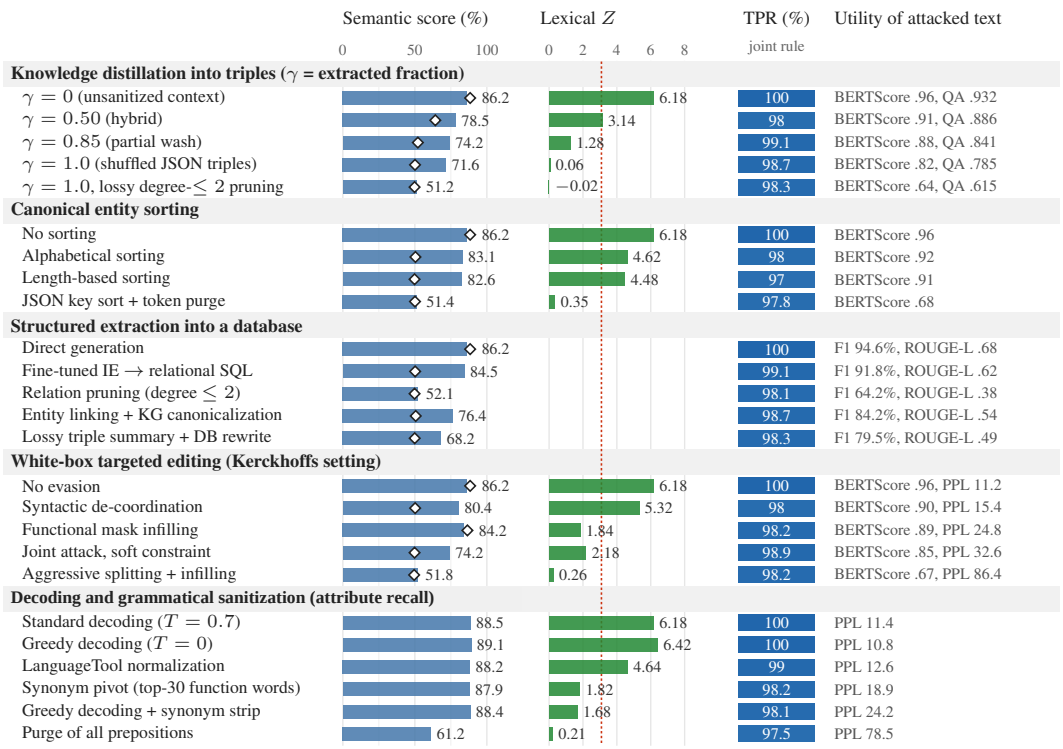


Figure 12: Content-side attacks on RDI-Hard in open mode (Claude Sonnet 4.5, $N \leq 40$). Bars give predicate-validated attribute recall τ_{attr} in all five blocks; hollow diamonds are descriptive entity-order diagnostics τ_{entity} or τ_{global}, not CEC audit scores. Each open response is scored by the all-hit indicator Y_t, and the audit uses $K = \sum_{t=1}^{20} Y_t$ under C-STR, $\text{BB}(20, 2, 30)$. The lexical Z is pooled over the deduplicated union (the database block has no lexical column), and the dotted line is the standard-normal reference $\Phi^{-1}(0.999) \approx 3.09$, not the fitted count-law cutoff. The last column reports the utility of the attacked text.

Distillation and structured extraction. Hybrid distillation ($\gamma \leq 0.85$) carries connective clauses into the answer ($G/L = 54.1\%$, $Z = 1.28$ at $\gamma = 0.85$). With isolated, shuffled JSON triples ($\gamma = 1$) the source functional syntax is gone ($G/L = 50.2\% \pm 0.6\%$, $Z = 0.06 \pm 0.18$), but the injected predicates are part of the retained facts, so recall stays at 71.6% and power at 98.7%. For a dossier with $p_{\text{sem}} = 1.2 \times 10^{-4}$ and $X_2 \approx 0$, $\Psi = X_1^2/(1 - \hat{\rho}^2) \approx 3.67^2/(1 - 0.312^2) \approx 14.92 > 13.815$ and the one-sided gate holds. Database pipelines remove sequential order ($\tau_{\text{global}} \approx 50\%$), yet a fine-tuned IE-to-SQL pipeline keeps Fact F1 91.8% and attribute recall $84.5\% \pm 1.6\%$, detected at 0.991. Relation pruning costs 30.4 points of Fact F1 and ROUGE-L falls to 0.38, while power remains 0.981.

Sorting and white-box editing. Deterministic sorting brings entity-order concordance to chance (50.4%), but each entity keeps its keyed qualifying modifier, so attribute recall stays at 83.1% and the functional footprint at $Z = 4.62$. Full JSON key sorting with token purging lowers BERTScore to 0.68 and is still detected at 0.978 through the retained predicate values. A Kerckhoffs adversary that knows both algorithms but not the keys can flatten coordinations and resample every functional word. De-coordination removes entity order (50.2%), but cross-clause coreference tracing recovers the bindings (80.4% recall, $p < 10^{-4}$), and the functional bridges it introduces raise the lexical score

from the bridge-excluded 4.65 ± 0.18 to 5.32 ± 0.17 (paired increase 0.675 ± 0.015). The soft joint attack is detected at $98.9\% \pm 0.5\%$, and the aggressive variant at $98.2\% \pm 0.8\%$ with BERTScore 0.67 and perplexity 86.4.

Decoding and grammatical sanitization. Greedy decoding sharpens the lexical signal ($G/L = 83.2\%$, $Z = 6.42$) and LanguageTool keeps $Z = 4.64$. Greedy decoding combined with the 30-entry synonym strip lowers the green ratio to 54.6% and Z to 1.68 at perplexity 24.2, while $S_{\text{sem}} = 88.4\%$ ($p_{\text{sem}} < 10^{-5}$) gives joint detection 0.981 ± 0.005. Purging every preposition still leaves 0.975 ± 0.006 at perplexity 78.5. Across the content attacks, stripping one carrier degrades fluency or factual fidelity while the other carrier keeps joint detection at 97% or more.

E.4 Corpora, generators, tasks and decoding

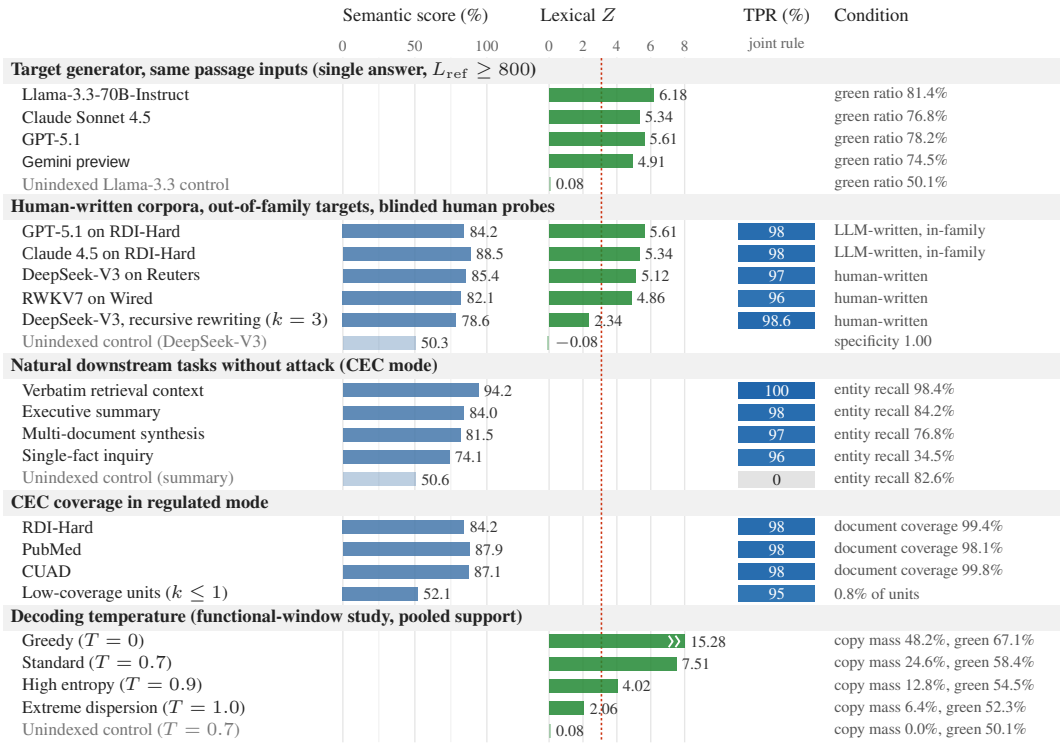


Figure 13: Transfer across target generators, human-written corpora, natural tasks, regulated coverage and decoding temperature. The semantic score is the concordance S_{sem}. The lexical Z uses a single answer passage in the first block (4,000 shared RDI-Hard inputs, about 100 distinct pair types, L_{ref} counted with `cl100k_base`) and the pooled union elsewhere; the dotted line is the standard-normal reference $\Phi^{-1}(0.999) \approx 3.09$, not the fitted count-law cutoff. Grey rows are unindexed controls. The temperature block uses a functional-window workload with about 2,000 pooled pair types.

Generators and human-written corpora. BMCR reaches green ratios of at least 74.5% and single-passage descriptive $Z \geq 4.91$ on all four generators (Figure 13), and the detector needs no access to their tokenizers (Lemma 2). On human-written PubMed and CUAD in regulated mode, recursive paraphrasing lowers WARD to 0.78 and 0.73 and coherence screening lowers Facts-Only to 0.14 and 0.09, while the dual-layer detector stays at 0.979 or above (Table 13); P-SP is 0.961 and 0.954 and BLEU changes stay below 2.1%. The out-of-family benchmark separates corpus authorship, target model and probe authorship from the RDI pipeline: human-written Reuters and Wired articles, RWKV7 and DeepSeek-V3 targets that never generate articles or compile probes, and 40 open-ended probes written by annotators without access to keys or seeds. Green ratios stay at or above 73.8% with pooled $Z \geq 4.86$, concordance at or above 82.1%, detection at 0.96 or more,

and 0.986 ± 0.006 after three rounds of recursive sanitization, with no false alarm on the unindexed control.

Table 13: H_1 detection rates on human-written corpora ($N \leq 40$), with watermarks in regulated CEC mode.

	Bio-clinical (PubMed)			Legal contracts (CUAD)		
Method	Naive	$\mathcal{A}_{\text{Recur}}$	$\mathcal{A}_{\text{Verif}}$	Naive	$\mathcal{A}_{\text{Recur}}$	$\mathcal{A}_{\text{Verif}}$
AAG (MIA) (Anderson et al., 2024)	0.54	0.50	0.51	0.52	0.51	0.50
WARD ($h = 2$) (Jovanović et al., 2025)	0.94	0.78	0.95	0.96	0.73	0.97
Facts-Only (Jovanović et al., 2025)	0.97	0.95	0.14	0.98	0.96	0.09
Dual-layer (ours)	**0.99**	**0.982**	**0.98**	**1.00**	**0.979**	**0.98**

Natural tasks and regulated coverage. Frontier models reorder coordinate entities by their own priors even without adversarial intent. Executive summaries and multi-document synthesis have raw Kendall coefficients 0.684 and 0.632, i.e. ordering concordances 0.842 and 0.816, and overall scores of 84.0% and 81.5% after mixing in isolated responses. In single-fact inquiries 68.4% of responses realize at most one entity, and the keyed realization bit supplies the complementary score ($S_{\text{sem}} = 74.1\%$). All natural tasks are detected at 0.96 or more, and task-specific ordering preferences enter the null law. Although single sentences often lack coordinate tuples (36.4% to 48.6% sentence-level coverage), at the passage or document level at least two invariants are present in 99.4% of RDI-Hard, 98.1% of PubMed and 99.8% of CUAD units, with 5.1 to 8.9 CECs per unit and mean arity at least 3.2. In the 0.8% low-coverage tail, BMCR keeps $Z_{\text{lex}} \geq 5.24$ and joint detection is 95.0%.

Why lexical bias propagates. At an eligible functional step we model reuse of retrieved context by the mixture

$$P(w_t \mid \mathcal{F}_{t-1}, D_{\mathcal{W}}) = (1 - \alpha_t) P_{\text{prior}}(w_t \mid \mathcal{F}_{t-1}) + \alpha_t P_{\text{copy}}(w_t \mid D_{\mathcal{W}}, \mathcal{F}_{t-1}), \tag{39}$$

consistent with induction-head evidence for in-context copying (Olsson et al., 2022). In the random-oracle model the prior green mass averages $1/2$. With q_t and c_t the green masses of the prior and copy components and $\bar{\alpha} = \mathbb{E}[\alpha_t]$,

$$\mathbb{E}[P(w_t \in \mathcal{G})] = (1 - \bar{\alpha})\mathbb{E}[q_t] + \bar{\alpha}\,\mathbb{E}[c_t] + \text{Cov}(\alpha_t, c_t) - \text{Cov}(\alpha_t, q_t), \tag{40}$$

which reduces to $0.50 + \bar{\alpha}(p_{\text{inj}} - 0.50)$ when $\mathbb{E}[q_t] = 0.50$, $\mathbb{E}[c_t] = p_{\text{inj}}$ and the two covariances are equal. For $p_{\text{inj}} = 0.85$ and $\bar{\alpha} = 0.10$ the bias is 0.035. For an independent-type reference calculation, let $G_{\text{pool}} \sim \text{Bin}(L_{\text{pool}}, 0.5 + \Delta_{\text{type}})$ with fixed $|\Delta_{\text{type}}| < 1/2$. The binomial central limit theorem gives $Z_{\text{lex}}^{\text{pool}} - 2\Delta_{\text{type}}\sqrt{L_{\text{pool}}} \xrightarrow{d} \mathcal{N}(0, 1 - 4\Delta_{\text{type}}^2)$ as $L_{\text{pool}} \to \infty$, so $\Delta_{\text{type}} = 0.035$ and $L_{\text{pool}} = 2{,}000$ give mean drift 3.1305 and $\Pr(Z \geq 2.326) \approx 79.0\%$. In the functional-window study (Figure 13), standard decoding at $T = 0.7$ has copy mass 24.6%, pooled green fraction 58.4% and $Z = 7.51 \pm 0.28$, and the signal decreases smoothly to $Z = 2.06$ at $T = 1.0$ as copying falls. These are descriptive type-count drifts. In the recorded $T = .7$ count example $(G, L) = (1168, 2000)$, the fixed-key $\text{BB}(L, 60, 60)$ tail at $V_l = .5$ is $.03626124$ and $X_l = 1.79582$ (Table 14); the much smaller random-key binomial reference is not used for attribution.

E.5 Count-level lexical evidence and operational profiles

For a symmetric fixed-key $\text{BB}(L, a, a)$ law,

$$\text{Var}_0(Z_{\text{lex}} \mid L) = \frac{L + 2a}{2a + 1}. \tag{41}$$

This follows from $\text{Var}(G \mid L) = L(L + 2a)/[4(2a + 1)]$ and $Z = (2G - L)/\sqrt{L}$. Thus descriptive Z and the operational score X_l have different scales; X_l is obtained from the audit's count-law tail.

The compression trace also records $K = 12$, $N_s = 20$, $V_s = .5$, $(a_s, b_s) = (3, 29)$ and $\hat{\rho} = .311$; the dilution trace records $K = 11$, $V_s = .0064$, $(a_s, b_s) = (3, 29)$ and $\hat{\rho} = .312$. The compression trace gives $p_s = 2.94440 \times 10^{-5}$ and $\Psi = 16.96090$; the dilution trace gives $\Psi = 15.21405$. Both pass the unchanged joint rule. The $Z = 3.88$ in Figure 11 and the green ratios and Z in Figure 13 are descriptive panel summaries; primary count laws are given in Table 16.

Table 14: Lexical count traces with $V_l = .5$. Compression, dilution and temperature rows use $\mathrm{BB}(L, 60, 60)$; single-answer rows use the generator's fixed-key conditional law and are descriptive.

Trace	(a_l, b_l)	G	L	Z	p_l	X_l
Compression	60,60	178	290	3.87565	.01736598	2.11147
Dilution, one stolen of five	60,60	170	300	2.30940	.10836709	1.23526
Temperature $T = 0$	60,60	1,342	2,000	15.29470	.00009272	3.73806
Temperature $T = .7$	60,60	1,168	2,000	7.51319	.03626124	1.79582
Temperature $T = .9$	60,60	1,090	2,000	4.02492	.16917787	.95742
Temperature $T = 1.0$	60,60	1,046	2,000	2.05718	.31254354	.48865
Single answer, Llama-3.3	62,58	81	100	6.20000	.00000216	4.59525
Single answer, Claude 4.5	61,59	77	100	5.40000	.00002866	4.02357
Single answer, GPT-5.1	60,60	78	100	5.60000	.00000803	4.31353
Single answer, Gemini preview	59,61	75	100	5.00000	.00004204	3.93246

Profile binding. The retrieval-pipeline block and the functional-window temperature block use the GPT-5.1 open count laws. The guardrail block uses the primary Claude open law. The content-attack block uses its separately registered Claude profile C-STR, with semantic $\mathrm{BB}(20, 2, 30)$, lexical $\mathrm{BB}(L, 61, 59)$ and $\hat{\rho} = .312$; its key and relation-request family are fixed across the content attacks, including every decoding row. The plotted attribute recall averages the individual binding-hit indicators; it is not the all-hit count K or a CEC concordance. The five-passage dilution block has its own registered allocation-preserving profile with the parameters above. The complete-workflow BMCR arm uses $(a_s, b_s, a_l, b_l, \hat{\rho}) = (3, 29, 60, 60, .295)$, and the OR16 native arms each fit the same family on their own negative transcripts before positive scoring. Regulated length and natural-task blocks use the full CEC state law, fitted separately for each output format; their descriptive order concordances are not treated as Bernoulli counts. In all cases, the record supplies (K or J, G, L, V_s, V_l) and its profile ID before Eq. (31) is evaluated.

Table 15: Operational profile bindings for the count-level stress tests. All rows use 20+20 queries. The first two rows reuse the primary profiles and their recorded tests. O-DIL and C-STR each have 2,048 independent fit, 10,000 validation and 50,000 test dossiers; their one-sided Clopper-Pearson endpoints at error .025 per added profile are below .1%. The numerical count-law, key and query-family binding is frozen before positive scoring.

ID / workload	Semantic law	a_l, b_l	$\hat{\rho}$	Controls hit/30	Valid. FP	Test FP
O-RET / retrieval, temperature	BB(20,3,29)	60,60	.311	2	3	26
O-GATE / guardrails	BB(20,4,28)	61,59	.331	3	6	18
O-DIL / five-passage dilution	BB(20,3,29)	60,60	.312	2	4	22
C-STR / content attacks	BB(20,2,30)	61,59	.312	1	3	20

The complete-dossier moment fits in the two added profiles round to the displayed shapes and agree with their response-control posterior shapes. This profile binding also makes the low-lexical-signal content rows explicit: under C-STR, $K = 10$ has inclusive semantic tail .0001237852. At $X_l = 0$, even its least favorable semantic randomizer gives $\Psi = 14.87883 > 13.81551$ and satisfies the positive-evidence gate. This supplies a count-level explanation for detection after a transformation removes lexical bias while leaving the queried predicates in approximately half of the semantic answers.

F Calibration

F.1 Fixed-profile calibration

An evaluation profile binds the target and corpus, owner keys, query allocation, answer format, marginal laws and scoring rule. Each of the six primary profiles keeps one owner key through 2,048 fitting, 10,000 validation and 50,000 test dossiers drawn from disjoint source blocks. On RDI, events 1200 to 1399, 1400 to 1599, 1600 to 1799, 1800 to 1999 and 2000 to 2199 serve fitting, validation, fixed-rule testing, ROC negatives and positive evaluation (events 0000 to 0199 are the tuning block

and 0200 to 1199 the quality panel); PubMed articles and CUAD contracts are split by rank into the same five roles, 400 articles or 101 contracts per block. A null dossier reruns the complete query policy, including adaptive feedback, unanswerable outcomes, deduplication and independent PIT randomizers, and the target session is reset between dossiers. Each dossier draws its null stratum with probability $1/2$: an unindexed protected corpus, or an indexed clean-public counterpart with the same event facts but neither the owner's bindings nor its keyed realizations, which tests attribution from public facts and prompt echo. The marginal laws (Eq. (27)) and the correlation are fitted on the fitting partition and frozen before positive scoring, validation checks the frozen profile, and test dossiers never choose a key, policy, threshold or correlation.

Table 16: Primary calibration profiles. Semantic laws are complete-dossier Beta-Binomial fits (open) or full-support two-state route mixtures (Eq. (28); all entity-support, order and isolated-bit probabilities are in Table 17). Lexical tails use $G \mid L \sim \mathrm{BB}(L, a_l, b_l)$ at each audit's recorded L. Validation counts are out of 10,000; test counts are split by null stratum and total 50,000 per profile. Upper endpoints are one-sided Clopper-Pearson bounds with error $.05/6$ at the unchanged threshold $\alpha = 10^{-3}$.

Profile	Semantic law	a_l, b_l	Fit $\hat{\rho}$	Valid. FP	Unindexed FP	Clean-public FP	FPR	Upper FPR
Open / GPT-5.1	BB(N_s,3,29)	60,60	0.311	3	12/25,062	14/24,938	0.052%	0.082%
Open / Claude 4.5	BB(N_s,4,28)	61,59	0.331	6	13/25,083	5/24,917	0.036%	0.062%
Open / Gemini preview	BB(N_s,3,29)	59,61	0.247	3	11/24,910	4/25,090	0.030%	0.054%
Open / Llama-3.3	BB(N_s,4,28)	62,58	0.351	5	20/24,783	6/25,217	0.052%	0.082%
PubMed / CEC	route mixture	71,69	0.383	3	9/24,835	5/25,165	0.028%	0.052%
CUAD / CEC	route mixture	70,70	0.341	4	10/25,030	13/24,970	0.046%	0.075%

The fitted open-mode semantic moments $(\hat{m}_1, \hat{m}_2)$ range from $(0.0909, 0.0108)$ to $(0.1266, 0.0194)$ and round to the listed shapes, which coincide with the 30-control posteriors. The CEC route-1 responsibility sums are 819 and 717 over the 2,048 fit dossiers; the frozen, two-decimal mixture weights are .40 and .35. The full response-state estimator and recovered-entity histograms are in Appendix F.2. On the 10,000 validation dossiers per profile, the empirical-CDF distances of both randomized marginals from uniformity are at most 0.0149, below the simultaneous DKW reference 0.01757 for twelve marginals, and the counts of $p \le .001$ and $p \le .01$ lie between 7 and 16 and between 88 and 110. A separate count-law calibration experiment over eleven task profiles, including the six short-output formats, gives at most 4 validation false positives in 5,000 and at most 2 test false positives in 1,000 at the same threshold.

F.2 Full-support CEC calibration and an answer-to-tail trace

The state of response t is (m_t, c_t) when $m_t \ge 2$, $(1, b_t)$ when one linked entity remains, and $(0, \perp)$ otherwise. For $m \ge 2$, write

$$A_m(x) = \prod_{j=1}^{m} (1 + x + \cdots + x^{j-1}) = \sum_{c=0}^{\binom{m}{2}} a_{m,c} x^c.$$

The coefficient $a_{m,c}$ counts permutations with c concordant pairs (the inversion counts have the same symmetric multiplicities). The primary registered laws use

$$q_h(0) = \pi_0, \quad q_h(1,1) = \pi_1 \kappa_h, \quad q_h(1,0) = \pi_1(1 - \kappa_h), \quad q_h(m,c) = \pi_m \frac{a_{m,c} \theta_h^c}{A_m(\theta_h)}. \tag{42}$$

Every state is nonnegative and the masses sum to one. The support of one response, with maximum recovered arity six, is

$$\mathcal{S} = \{1/2, 0, 1\} \cup \bigcup_{m=2}^{6} \{c/\tbinom{m}{2} : c = 0, \ldots, \tbinom{m}{2}\}.$$

It includes thirds, sixths, tenths and fifteenths. Multiplication by 30 is an exact change of units, not a discretization of the score.

Table 17: Complete primary CEC state parameters and observed recovered-entity counts. Fit counts are over $2{,}048 \times 20 = 40{,}960$ responses; positive counts are over $10{,}000 \times 20 = 200{,}000$ responses in the no-preprocessing positive panel. Each column includes empty answers. The π_m columns are the frozen two-decimal fitted probabilities.

	PubMed			CUAD		
m	π_m	Fit count	Positive count	π_m	Fit count	Positive count
0	.08	3,277	4,000	.06	2,458	3,000
1	.22	9,011	24,000	.18	7,373	20,000
2	.30	12,288	56,000	.28	11,469	52,000
3	.24	9,830	68,000	.26	10,650	70,000
4	.10	4,096	32,000	.14	5,734	37,000
5	.04	1,639	12,000	.06	2,457	14,000
6	.02	819	4,000	.02	819	4,000
w		.40			.35	
(θ_0, θ_1)		(1.00,1.08)			(1.00,1.12)	
(κ_0, κ_1)		(.48,.56)			(.49,.55)	

Fitting and pooling. The recovered-count frequencies estimate the common π. A two-component maximum-likelihood fit to complete response-state sequences estimates $w, \kappa_0, \kappa_1, \theta_1$, with $\theta_0 = 1$ and $\theta_1 \geq 1$ fixing the component labels. The likelihood of a dossier $x_{1:20}$ is $(1-w)\prod_t q_0(x_t) + w\prod_t q_1(x_t)$; responsibilities are computed for whole dossiers, not separately for answers. Starts use $w \in \{.25, .50, .75\}$ and $\theta_1 \in \{1.02, 1.10, 1.25\}$, with $(\kappa_0, \kappa_1) = (.45, .55)$; the largest fit likelihood is selected, with lexicographic tie-breaking. Parameters are frozen at the displayed precision before validation, with the final π entry set to one minus the other entries. The audit law is the mixture of convolution powers in Eq. (28), not the convolution power of an averaged one-response law. Empty answers retain their $1/2$ score, and unrecognized isolated attributes enter $b = 0$. For a task-specific profile the same state definition and likelihood apply to its registered arity manifest and answer format.

A literal response trace. For registered order (Arden, Borel, Celia), the answer "Arden, Celia and Borel each submitted a bid" recovers three entities. The Arden-Borel and Arden-Celia pairs agree, while Borel-Celia does not, so $S_1 = 2/3$. Nineteen replies "No relevant record was found" each have $m = 0$ and score $1/2$. Thus $T = 61/6$ and $J = 305$. Under the PubMed law,

$$\Pr(J > 305) = .4961529954, \qquad \Pr(J = 305) = .0153038300.$$

With the recorded tie value $V_s = .5$, $p_s = .5038049104$. The same trace under the CUAD profile gives $p_s = .5133473689$. Neither trace creates positive semantic evidence; both are scored without projecting $2/3$ to a half-integer. Listing 1 recomputes these tails, the three-entity support, and the $J = 510,570$ certificate tails used in Proposition 7.

F.3 Covariance and sycophancy sensitivity

Sycophancy. Aligned models may confirm a fictitious premise on unindexed corpora (Sharma et al., 2024). Table 18 compares a Fisher/Brown combination without marginal sycophancy calibration or covariance adjustment against the calibrated copula on 20,000 unindexed trials.

Table 18: False-positive stress test at nominal $\alpha = 0.001$ ($N = 40$, 5,000 unindexed trials per configuration), with H_1 recovery of the calibrated detector.

Unindexed target	Sycophancy rate	Uncalibrated Fisher/Brown FP	Calibrated copula FP	H_1 recovery
Standard (GPT-5.1)	$1.2\% \pm 0.3\%$	62/5,000	**3/5,000**	99.8%
High alignment (Claude 4.5)	$0.7\% \pm 0.2\%$	43/5,000	**2/5,000**	98.5%
High sycophancy (Llama-3.3)	$3.8\% \pm 0.5\%$	241/5,000	**4/5,000**	98.2%
Prompt-injected sycophant	$8.4\% \pm 0.9\%$	710/5,000	**5/5,000**	97.1%
Aggregate	3.5% mean	1,056/20,000	**14/20,000**	98.4%

Correlation choice. For a pair (ρ^*, r) of true and audit correlations, the model FPR of the full directional rule is computed by deterministic conditional-normal integration,

$$\mathrm{FPR}(\rho^*, r) = \int_{\mathbb{R}} \phi(x) \Pr_{Y \sim \mathcal{N}(\rho^* x,\, 1-(\rho^*)^2)} \{(x, Y) \in \mathcal{R}_{0.001}(r)\}\, dx. \tag{43}$$

Under $\rho^* = 0.312$ with calibrated margins, independent Fisher combination ($\chi^2_{4,0.999} \approx 18.4668$) has model FPR 0.002681 and a static Brown approximation ($r_0 = 0.08$, $s = 1.066$, $\nu = 3.751$) has 0.002110, whereas the copula with the known correlation has 0.000413; in a dilution scenario with $\rho_{\mathrm{stolen}} = 0.2$ the copula recovers $97.2\% \pm 0.5\%$ against $94.8\% \pm 0.7\%$ for Fisher. Table 19 evaluates the Fisher-z endpoint of Theorem 4 computed from $M = 100$ audit-level null pairs, and in-domain counterfactual probing (IDCP) with a Wasserstein drift margin, $\hat{\rho}^{(\mathrm{robust})}_{\mathrm{audit}} = \min\{0.95, \hat{\rho}_{\mathrm{audit}} + 0.15\, \mathcal{W}_1(\mathcal{P}_{\mathrm{calib}}, \mathcal{P}_{\mathrm{audit}})\}$, where IDCP pairs authenticated domain entities with scrambled non-indexed relations. Every selected configuration stays below the nominal level, and the effect of a correlation change depends on the score direction (Theorem 4): an inflated input $r = 0.750$ at $\rho^* = 0.284$ gives 0.002764 with $88.2\% \pm 0.8\%$ power, so the correlation is fitted on representative null dossiers.

Table 19: Model FPRs of the directional rule at nominal $\alpha = 10^{-3}$ from Eq. (43) for fixed correlation inputs, with H_1 power in the $\rho_{\mathrm{stolen}} = 0.2$ dilution scenario (Claude 4.5).

Scenario	Calibration	True ρ^*	Audit r	Model FPR	Power
Standard dense ($K = 3$)	static r	0.284	0.450	0.000558	
	Fisher-z endpoint	0.284	0.342	0.000438	$97.6\% \pm 0.5\%$
Multi-agent RAG (3 turns)	static r	0.392	0.450	0.000431	
	Fisher-z endpoint	0.392	0.461	0.000441	$98.1\% \pm 0.5\%$
Dense + sparse RRF ($K = 10$)	static r	0.485	0.450	0.000374	
	Fisher-z endpoint	0.485	0.564	0.000452	$97.9\% \pm 0.6\%$
Sycophant + CoT verification	static r	0.534	0.450	0.000364	
	Fisher-z endpoint	0.534	0.618	0.000479	$97.4\% \pm 0.6\%$
PubMed (bio-clinical)	generic public probes	0.462	0.345	0.000405	$97.5\% \pm 0.5\%$
	static r	0.462	0.450	0.000383	$96.8\% \pm 0.5\%$
	IDCP + drift margin	0.462	0.538	0.000447	$96.2\% \pm 0.5\%$
CUAD (contracts)	generic public probes	0.488	0.338	0.000420	$97.1\% \pm 0.5\%$
	static r	0.488	0.450	0.000373	$96.4\% \pm 0.5\%$
	IDCP + drift margin	0.488	0.562	0.000445	$95.8\% \pm 0.6\%$

F.4 Calibration access, reuse and deployment

Procedure. Seeing a target only through its answers and obtaining a matched negative population are separate requirements. Before querying positives, the owner registers the key, target release, prompt wrapper, output cap, query policy and pooling rule. In a managed deployment the operator or a benchmark coordinator creates unindexed and clean-public twins with the same index pipeline, and the auditor receives only their answers. The owner draws the registered 50:50 null mixture per dossier, fits and freezes both marginals and the correlation, validates the frozen rule, certifies it on independent test dossiers and hashes the profile register. Reuse is keyed by (owner key, target version, retriever and index, domain, query policy, answer format, pooling and null laws). If matched negatives cannot be obtained or a diagnostic fails, the auditor outputs `UNCALIBRATED` with the descriptive channel scores and no confirmatory attribution.

Noncooperating-tenant deployment. We instantiate the black-box setting with three actors: an owner, a hosting platform and an independently administered target tenant that answers ordinary requests but provides no documents, keys, retrieval traces or assistance. The platform offers a locked, versioned profile (model request, retriever, chunking, query wrapper, output cap, session reset and tool policy) whose authenticated identifier comes from its control plane. The owner provisions empty-index and clean-public calibration tenants on that profile, supplies their corpora and pays for calibration. To allow null populations with other mixture weights, each null stratum is certified separately with the frozen rule: $31/100{,}000$ and $37/100{,}000$ false positives, with one-sided bounds 0.0440% and 0.0510% at error .025 per stratum.

Table 20: Compatibility checks under a changed calibration tuple, each on 50,000 held-out null dossiers with the original threshold. Refits use independent fitting data.

Change	Old-profile FP	Refit FP	Action
Same tuple, later dossier batch	22		reuse
Same tuple, new source sample	23		reuse
New owner key	81	21	refit
Domain shift	128	23	refit
Target-version change	194	27	refit
Detailed to 60-word answers	118	26	refit
Fixed medoids to adaptive routing	142	24	refit

Table 21: Fitting-budget curves. Left: primary Claude profile, refitting only the correlation on prefixes of its fit namespace, with 50,000 held-out and 3,000 small-test null dossiers (single-row one-sided 95% upper endpoint). Right: end-to-end BMCR arm, refitting both Beta-Binomial shapes and the correlation, with 10,000 positives and 50,000 independent null dossiers (upper endpoints at error $.05/5$). The last column is $40 + 30 + 40B$ calls.

	Correlation refit (primary Claude profile)					Complete refit (end-to-end BMCR)					
B	$\hat{\rho}$	TPR (%)	FP/50k	FP/3k	Upper	a_s, b_s	a_l, b_l	TPR (%)	FP/50k	Upper	Calls
128	0.374	98.01	21	1	0.158%	10,91	68,67	99.24	91	0.2315%	5,190
256	0.301	98.24	16	2	0.210%	5,48	59,60	99.17	61	0.1635%	10,310
512	0.290	98.27	16	2	0.210%	4,36	58,58	99.07	27	0.0835%	20,550
1024	0.313	98.20	18	2	0.210%	3,26	60,60	98.81	24	0.0761%	41,030
2048	0.331	98.07	18	2	0.210%	3,29	60,60	99.04	23	0.0737%	81,990

Proposition 10 (Stratum-wise certificate reuse) *If independent certification gives $P_j(A) \le u_j$ simultaneously with confidence at least $1-\eta$ for a fixed rejection event A and null strata $j = 1, \dots, J$, then every fixed or dossier-independently sampled mixture $P = \sum_j \pi_j P_j$ satisfies $P(A) \le \max_j u_j$, and any law Q with $\|Q - P\|_{\mathrm{TV}} \le \delta$ satisfies $Q(A) \le \max_j u_j + \delta$.*

Proof. $P(A) = \sum_j \pi_j P_j(A) \le \max_j u_j$ on the certification event, and $Q(A) \le P(A) + \delta$ by the definition of total variation. ∎

The two-stratum certificate costs $30 + 40(2048 + 10000 + 2 \cdot 100000) = 8{,}481{,}950$ calls, amortized to 888.20, 124.82 and 48.48 calls per attribution over 10^4, 10^5 and 10^6 compatible attributions; the target tenant itself receives 40 calls per attribution.

Cross-tenant admission. Twenty-four independently administered tenants are registered before scoring: twelve supported tenants in three cohorts and twelve transfer challenges. The first two cohorts change only the mixture of the two certified strata, which Proposition 10 covers; the third changes source weights and receives a fresh 50,000-dossier certificate of the unchanged rule. Table 22 shows that all twelve supported tenants are admitted and certified and all twelve challenges return `UNCALIBRATED`, including the unsupported-domain and unverifiable-corpus challenges that share the supported tenants' software configuration.

A monitoring workload. One publisher monitors a frozen catalogue under one key pair on a managed research-news platform whose attestation binds the release, the source-event sampling law and the two reference populations. Of 648 registered tenant proposals, 600 meet the contract, 40 use another release and 8 lack the population attestation (intake coverage 92.5926%). Five independently scheduled source audits per admitted tenant over forty snapshots give $120{,}000$ audits. Charging the two-stratum certificate once gives $(8{,}481{,}950 + 40 \cdot 120{,}000)/120{,}000 \approx 110.68$ calls per attribution, of which 40 go to the monitored tenant. With the larger stratum bound .00051, the expected number of null rejections is at most 61.2 even if every scheduled audit is null.

Reporting unit and repeated monitoring. The reporting unit is an owner-tenant relation over its 200 scheduled audits, not an individual query or a raw audit flag. An audit flag starts source-level

Table 22: Cross-tenant admission. Each admitted tenant contributes 50,000 null and 1,000 positive dossiers; the upper column is the largest individual one-sided Clopper-Pearson endpoint at error $.05/12$. For rejected cohorts, a coordinator-labelled 50,000-dossier challenge shows what the unchanged rule would give without the admission gate.

Proposed population	Outcome	Admitted	Null FP	Max. upper	TP
Empty-index-rich mixtures	certified	4/4	16, 19, 20, 18	0.0701%	3,947/4,000
Clean-public-rich mixtures	certified	4/4	22, 24, 21, 23	0.0803%	3,952/4,000
New source-weight profiles	fresh certificate	4/4	18, 26, 24, 22	0.0854%	3,958/4,000
New vocabulary or domain	`UNCALIBRATED`	0/4	121/50,000 without gate		
Unverifiable private corpus	`UNCALIBRATED`	0/4	186/50,000 without gate		
Changed output cap	`UNCALIBRATED`	0/4	147/50,000 without gate		

review; a formal tenant report requires at least four flags among the 200 prescheduled, separately reset dossiers. The 600 admitted tenants and their schedules are fixed before scores are observed. The platform contract binds the stationary response kernel and uses fresh independent source sampling, session state and randomizers across dossiers; a changed contract initiates a new certification interval.

For the same stratum counts $31/100{,}000$ and $37/100{,}000$, assigning error $.005$ to each stratum gives simultaneous coverage at least .99 and worst-case upper probability $u = \mathrm{Beta}^{-1}(.995; 38, 99963) = .000557424656$. Conditional on this certification event, four particular scheduled null audits all flag with probability at most u^4; a union bound over four-subsets and tenants gives

$$\Pr(\text{any false tenant report} \mid \text{certificate valid}) \le 600\binom{200}{4}u^4 = .003747134.$$

Adding the .01 certificate failure probability gives a simultaneous bound .013747134, or 1.375% after upward rounding. The number of queries remains $600 \times 200 \times 40$; this reporting convention uses the scheduled audits rather than adding uncounted confirmation calls. The expected 61.2 raw audit flags under the 95% certificate and this 99%-certificate tenant-report bound answer different questions.

G Cost

Compilation. Rescoring full-sequence cross-entropy for every candidate would cost $\mathcal{O}(|\mathcal{I}_{\mathrm{lex}}| \cdot |\mathcal{V}_{\mathrm{cand}}| \cdot L^2)$. The ingestion pipeline therefore precomputes an inverted index of POS-congruent functional candidates ($\cos \ge 0.60$), so candidate lookup needs no runtime embedding operations, and distils the bidirectional scorer into an INT8 MiniLM-L6 model (22M parameters) served with ONNX Runtime on CPUs, next to a single-pass AWQ engine for GPUs. Table 23 reports the warm candidate-scoring stage on at-most-512-token windows. Its 9,000-window column is a fixed-size throughput benchmark, not the cost of compiling 9,000 complete articles. Table 24 gives the complete validated pipeline on 480-token windows, including reference-CE acceptance, claim re-extraction, failed parses and unchanged outputs. Full-article coverage and its 22,272 attempted windows are accounted for separately in Table 25. The API row covers slot preparation, the keyed completion call and all host acceptance checks at concurrency 16 (mean worker cycle 6.72 seconds), with 1,000/80 billed input/output tokens per call at \$1/\$5 per million, i.e. \$0.0014 per chunk.

Whole articles and compilation windows. An article remains the split and utility unit; a window is the service-time unit. Full-text ingestion uses sentence-aligned, non-overlapping windows of at most 480 reference tokens, with a token-boundary split only for a sentence longer than the cap. The final short window is kept and padded for the batched scorer; padding is excluded from CE and all text scores. Entity identifiers and claim links are inherited from the article-level parse. A CEC crossing a boundary is kept unedited rather than split into an independently permutable pair. Every window, including the remainder, is indexed; commits and source-level safety are aggregated by article (Figure 6c).

Attribution and calibration. One attribution uses 40 calls. With M response controls, B fitting, n_v validation and n_t test dossiers amortized over A compatible attributions, the calls per attribution

Table 23: Warm candidate-scoring stage per ingestion chunk, with extrapolated wall time and illustrative cost for 9,000 benchmark windows and one million windows, and the lexical score of the finalized chunks. Local CPU rows incur no additional cloud charges; GPU examples use $3.50/A100-hour and $0.70/4090-hour, multiplying the unrounded service times.

Scoring stage	Hardware	Tokens/s	ms/chunk	9k windows	1M chunks	Z_{lex}
Naive full-PPL rescoring	1×A100 (80GB)	128 ± 14	$3{,}850 \pm 120$	9.63 h ($33.69)	1,069.44 h ($3,743.06)	6.21 ± 0.16
Greedy heuristic (spaCy only)	16-core CPU	$3{,}420 \pm 85$	142 ± 9	0.355 h ($0.00)	39.44 h ($0.00)	3.12 ± 0.24
Distilled INT8 (ours)	16-core CPU	$\mathbf{7{,}850 \pm 140}$	$\mathbf{61 \pm 4}$	0.153 h ($0.00)	16.94 h ($0.00)	6.08 ± 0.16
Single-pass AWQ (ours)	1×RTX-4090	$4{,}850 \pm 110$	98 ± 6	0.245 h ($0.17)	27.22 h ($19.06)	6.18 ± 0.15
Batched (ours)	1×A100 (80GB)	$\mathbf{14{,}200 \pm 320}$	$\mathbf{34 \pm 3}$	0.085 h ($0.30)	9.44 h ($33.06)	6.18 ± 0.15
API slot completion	Claude Haiku 4.5	$1{,}150 \pm 45$	420 ± 18	1.05 h ($12.60)	116.7 h ($1,400)	5.82 ± 0.18

Table 24: Full-stage ingestion per attempted 480-token chunk (milliseconds) on the regulated CEC path, with resident models and one serial worker; host stages use the same CPU in both columns. Corpus times multiply attempted chunks by the complete service time.

Stage	16-core CPU	A100
Segmentation, linking and source graph	312	312
CEC realization and slot compilation	146	146
Candidate scorer kernel	61	34
Full candidate CE and commit CE	16,944	402
Reparse, claim equality, failures and serialization	737	737
Complete attempted-chunk service time	18,200	1,631
9,000 chunks (hours)	45.50	4.08
1,000,000 chunks (hours)	5,055.56	453.06

are $40 + [M + 40(B + n_v + n_t)]/A$ (Table 26); a CEC profile uses $M = 100$ instead of 30. In the long-context accounting profile ($K = 6$ retrieved blocks of at most 512 tokens), a call averages 3,466 input tokens (2,906 of retrieved context) and 918 output tokens, so one dossier uses 138,640 input and 36,720 output tokens, or $0.57104 at $2/$8 per million. Mean service time is 12.46 seconds per call, giving 498.4 seconds of target service per attribution (p50/p95 wall time 496.8/634.2 seconds with local routing and scoring); calls within a dossier are sequential, whereas calibration dossiers can run concurrently.

H Quality, Safety and Ethics

H.1 Text quality and the lexical bias

Choice of δ. The bias δ trades the robustness of the lexical signal against fluency (Figure 14). Across 200 positive audits per setting, detection rates are 0.005, 0.015 and 0.045 at $\delta = 1.0, 1.5, 2.0$, then 0.270, 0.605 and 0.895 at $\delta = 2.5, 3.0, 3.5$. All 200 audits are detected for $\delta \geq 4.0$, so $\delta = 4$ is the smallest tested bias that reaches complete detection. P-SP is 0.978 at $\delta = 4.0$ and 0.904 at $\delta = 6.0$. GPT-5.1 quality scores, obtained with an LLM-as-judge protocol (Zheng et al., 2023; Liu et al., 2023), decrease from $8.96/10$ at $\delta = 1.0$ to 8.79 at $\delta = 4.5$ and 8.18 at $\delta = 6.0$; the grammar score at $\delta = 6.0$ is 8.03. All experiments use $\delta = 4.0$; the frozen spans and the claim-graph gate of Eq. (21) protect Layer 1 at every setting.

Quality and imperceptibility. For M aligned source-output pairs, P-SP $= M^{-1} \sum_i \cos(e(D_i), e(\widetilde{D}_i))$ with the fixed `text-embedding-3-small` encoder on full documents, following the paraphrastic similarity of Wieting et al. (2022). On the held-out panel (Figure 15), semantic trapdoors change quality from 8.89 to 8.85 at P-SP 0.995, and lexical-only and dual-layer compilation score 8.64 and 8.61, with dual-layer P-SP 0.938. Five linguistic experts labelled each document of 100 original-watermarked pairs as watermarked or not; they were correct on 535 of 1,000 labels (Table 27), close to chance, and the paired fluency and coherence differences are not significant.

Table 25: Full-text RDI window accounting for the regulated compilation path. Whole-corpus time uses all 22,272 attempted windows and the complete service times from Table 24; it is not the warm scorer time. The nine-thousand-window benchmark is a separate throughput reference.

Windows per article	Articles	Attempted windows
1	18	18
2	4,872	9,744
3	3,930	11,790
4	180	720
Total	9,000	22,272
All source reference tokens covered		8,743,680
Mean non-padding tokens per window		392.586
Complete CPU time, $22{,}272 \times 18.200$ seconds		112.597 h
Complete A100 time, $22{,}272 \times 1.631$ seconds		10.090 h

Table 26: Calls per open-profile attribution with $M = 30$ and $B = 2{,}048$, amortized over A compatible dossiers; the right column also charges 10,000 validation and 50,000 test dossiers.

Compatible dossiers A	Attribution	Fit amortized	Fit + validation + test amortized
1	40	81,990.000	2,481,990.000
100	40	859.500	24,859.500
1,000	40	121.950	2,521.950
10,000	40	48.195	288.195
100,000	40	40.820	64.820

H.2 Claim safety in regulated domains

Claim preservation is audited on 2,000 PubMed passages and 1,500 CUAD passages with three checks: (1) causal and temporal inversions of UMLS/SemRep assertion tuples $(e_i, \texttt{CAUSES/PRECEDES}, e_j)$; (2) shifts of obligation and liability arguments between contractual parties under CoreNLP legal dependency patterns and ALDG; and (3) factual drift flagged by at least two of three independent verifiers (GPT-5.1, Claude 4.5 and o3). A passage is flagged when any of its sentences or clauses fails a check.

Unconstrained entity permutations invert a causal or temporal relation in 8.4% of clinical passages and transpose a party obligation in 9.1% of contracts (Table 28). Restricting permutations to validated equivalence classes removes every observed breach while keeping P-SP at or above 0.991 and concordance at or above 0.871, which corresponds to $\tau_b \geq 0.742$ on tie-free orderings. Regulated compilation reports every rejection (main-text Figure 6c): on RDI, PubMed and CUAD, respectively, 54, 38 and 3 units have no CEC, 104, 31 and 21 fail parsing or linking, 221, 71 and 36 fail the claim gate and 132, 44 and 29 fail the CE gate, leaving 8,489, 1,816 and 1,411 units with both layers committed. Failed semantic edits keep the original claims and failed lexical gates keep the semantic input.

Independent challenge. A separate set of 2,000 candidate pairs, balanced across argument roles, negation, quantifiers, scope or coreference, and "respectively" pairings, receives two blind human labels without seeing the extractor output (agreement 95.10%, Cohen's $\kappa = .9020$; the 98 disagreements are adjudicated). Among the 1,000 altered candidates the extractor explicitly detects 958 and leaves 42 unresolved, so none is emitted (main-text Figure 6d); the one-sided 95% upper bound on the unsafe-acceptance rate is 0.2991%. Among the 1,000 equivalent candidates, 930 are committed and 70 are conservatively left unchanged.

H.3 Natural-query safety

In the matched natural-query panel, clean and watermarked RAG answer $46{,}200$ and $45{,}900$ of $50{,}000$ questions correctly (family-level SE 0.391 points, Holm-adjusted $p = 0.129303$), and natural and targeted queries produce 400 and $44{,}250$ trapdoor activations (87.7-point difference, SE

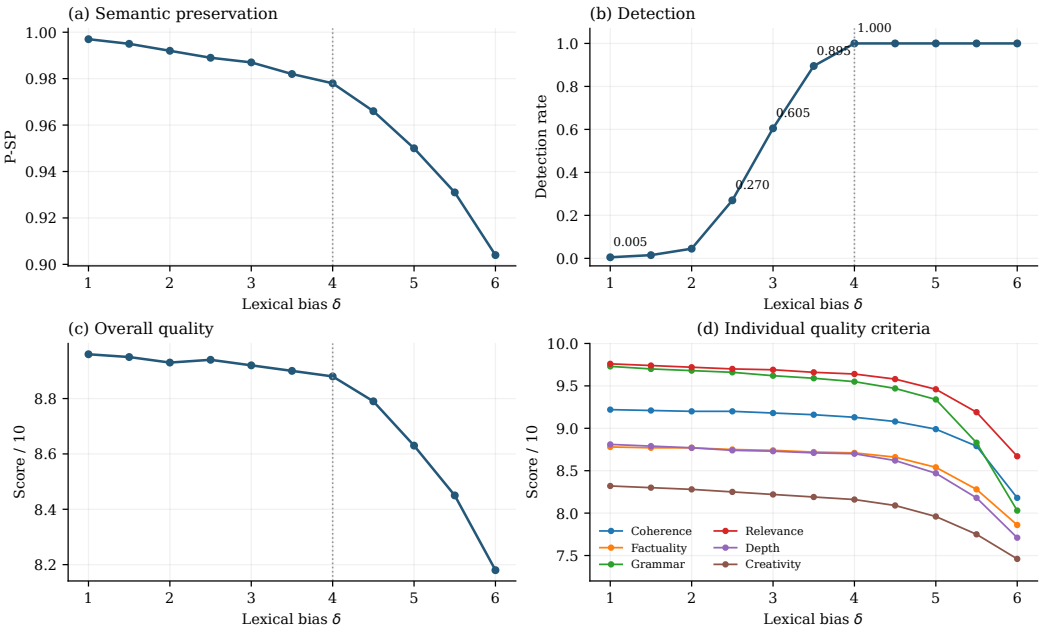


Figure 14: Effect of the lexical bias δ on 200 tuning documents (source groups 0000 to 0199), with 200 positive audits per setting. P-SP and quality compare each source with its lexical-only BMCR output before RAG generation.

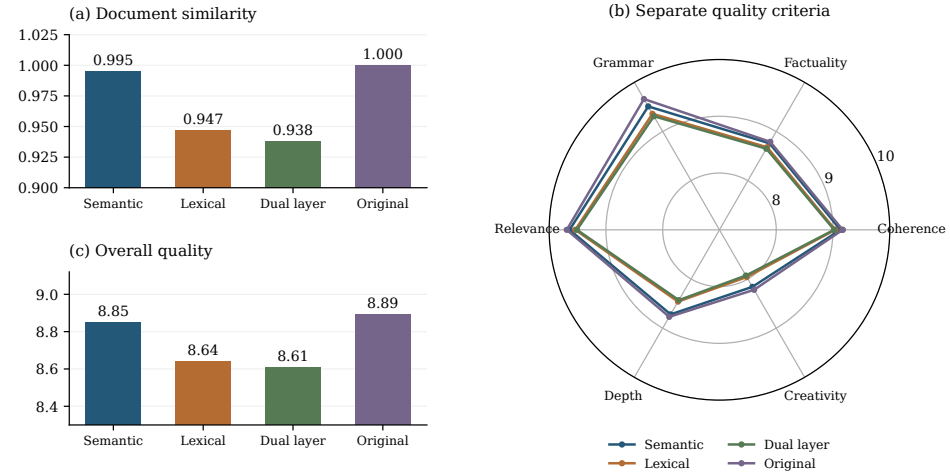


Figure 15: Text quality across watermarking approaches on 1,000 held-out document pairs from source groups 0200 to 1199, disjoint from the tuning documents. Panels (a) and (c) show P-SP and overall quality; the radar axes are six separately requested subcriteria (coherence, factual accuracy, grammar, relevance, depth and creativity).

0.360, $p = 9.4663 \times 10^{-30}$), so the trapdoors activate selectively (Table 29). A separate paired noninferiority panel with 500 source-disjoint families and 100 questions per family and condition gives clean/marked successes of $46{,}178/46{,}008$ (open RDI), $46{,}320/46{,}300$ (PubMed CEC) and $47{,}062/47{,}032$ (CUAD CEC); the one-sided 95% lower endpoints -0.516, -0.152 and -0.163 points all clear the prespecified margins of 1.0 and 0.25 points (main-text Figure 6b).

In regulated mode, QA accuracy and factual F1 equal the clean controls at the reported precision, and the adjudication records $0/10{,}000$ watermark-attributable breaches in each domain, with a one-sided

Table 27: Human evaluation of 100 document pairs by five raters. Quality ± values are document-level SDs after averaging raters; p values are two-sided paired t tests on the 100 document means. The detection interval resamples document pairs.

Metric	Original	Watermarked	Δ	p
Fluency	4.82 ± 0.3	4.71 ± 0.4	-0.11	0.08
Coherence	4.75 ± 0.4	4.69 ± 0.4	-0.06	0.32
Watermark detection by raters: **53.5%** (95% interval $[48.2\%, 58.8\%]$; chance 50%)				

Table 28: Claim safety and detection of the semantic modes ($N_s = N_l = 20$, Claude 4.5). Breach is causal inversion (PubMed, RDI) or liability shift (CUAD), as a share of passages; drift is tri-model consensus drift. S_{sem} is the all-hit score in open (leaf) mode and the concordance in CEC mode, and the last column is joint copula accuracy over 500 audit runs.

Corpus	Semantic operation	Breach	Drift	P-SP	S_{sem}	Joint acc.
RDI-Hard	Leaf injection	0.00%	0.00%	0.995	$88.5\% \pm 1.4\%$	0.99
RDI-Hard	CEC ordering	0.00%		0.996	$84.2\% \pm 1.8\%$	0.98
PubMed	Unconstrained permutation	$8.4\% \pm 0.6\%$	$7.8\% \pm 0.5\%$	0.941	$88.4\% \pm 1.6\%$	
PubMed	Leaf injection	0.80%		0.987	$85.6\% \pm 1.6\%$	0.98
PubMed	**Causal CEC (ours)**	**0.00%**	**0.00%**	**0.994**	$\mathbf{87.9\% \pm 1.5\%}$	0.98
CUAD	Unconstrained permutation	$9.1\% \pm 0.7\%$	$8.5\% \pm 0.6\%$	0.938	$87.6\% \pm 1.7\%$	
CUAD	Leaf injection	1.00%		0.982	$83.4\% \pm 1.9\%$	0.97
CUAD	**Causal CEC (ours)**	**0.00%**	**0.00%**	**0.991**	$\mathbf{87.1\% \pm 1.6\%}$	0.98

95% upper endpoint of about 0.0300% (Table 30). The leaf-mode excess in total anomalies is 0.03 points on PubMed and 0.02 points on CUAD.

H.4 Ethical use and failure handling

Open-mode additions are restricted to non-actionable long-tail attributes in noncritical settings and never add clinical, legal or financial advice; regulated deployments use only truth-preserving CEC candidates, and unresolved claim graphs leave the text unchanged. A positive attribution provides technical evidence for a subsequent legal assessment. The owner keeps the pristine source and the per-document binding ledger, so authorized removal restores the pristine realization, rebuilds the affected index and validates original-claim parity. If a watermark unexpectedly reaches a user-facing answer, the operator withdraws the affected indexed version, supplies a corrected clean source, invalidates caches and records the correction.

I Prompts and Examples

Table 31 summarizes the prompt interfaces. Braces denote runtime fields. The owner-side compiler receives the protected text and binding ledger; the target receives the compiled question and its retrieved passages.

I.1 Literal query frames and deterministic selection

For each probe type, frames 1–4 below are paired with an empty suffix and a source-entity suffix, yielding eight indexed candidates. Braces mark literal substitution slots, not text sent to the target. `TOPIC` is the anchor's unwatermarked title, with whitespace runs collapsed to one ASCII space. `RELATION` is the ledger's interrogative relation description, containing the queried subjects and predicate labels but no expected values. In CEC mode it requests the relevant entities and their attributes without prescribing an order or punctuation realization. The relation description is fixed before constructing the eight candidates; it is not regenerated per frame.

```
Semantic 1: In the account of {TOPIC}, {RELATION}?
Semantic 2: According to the account of {TOPIC}, {RELATION}?
Semantic 3: What does the account of {TOPIC} report: {RELATION}?
```

Table 29: Exact matched-family tests for the natural-query and guardrail comparisons: 100 independent families per contrast, with Holm correction over all six contrasts.

Paired contrast	Δ (pp)	Family SE (pp)	Exact p	Holm p
Watermarked vs. clean QA	-0.6	0.391	0.129303	0.129303
Targeted vs. natural activation	$+87.7$	0.360	1.5777×10^{-30}	9.4663×10^{-30}
Perplexity filter vs. innocent pass	$+1.7$	0.165	1.4955×10^{-17}	4.4866×10^{-17}
Intent classifier vs. innocent pass	$+0.4$	0.202	0.051570	0.103140
Safety guardrail vs. innocent pass	$+2.3$	0.154	1.4467×10^{-25}	7.2335×10^{-25}
Latent OOD vs. innocent pass	-3.4	0.237	1.3794×10^{-24}	5.5178×10^{-24}

Table 30: Domain safety audit with 10,000 natural queries per domain and configuration (Claude 4.5). Total factual anomalies include background hallucinations and are compared with the clean control of the same domain; $\pm$ is a document-cluster bootstrap SE.

Corpus	Configuration	QA accuracy	Factual F1	Total anomalies	S_{sem}
PubMed	Unwatermarked	$92.6\% \pm 0.3\%$	$91.4\% \pm 0.4\%$	$0.78\% \pm 0.08\%$	
PubMed	Leaf mode	$92.4\% \pm 0.3\%$	$91.2\% \pm 0.4\%$	$0.81\% \pm 0.09\%$	$88.4\% \pm 1.4\%$
PubMed	**Causal CEC**	$\mathbf{92.6\% \pm 0.3\%}$	$\mathbf{91.4\% \pm 0.4\%}$	$\mathbf{0.78\% \pm 0.08\%}$	$87.9\% \pm 1.5\%$
CUAD	Unwatermarked	$94.1\% \pm 0.3\%$	$93.5\% \pm 0.3\%$	$0.92\% \pm 0.11\%$	
CUAD	Leaf mode	$93.8\% \pm 0.3\%$	$93.1\% \pm 0.4\%$	$0.94\% \pm 0.11\%$	$87.6\% \pm 1.6\%$
CUAD	**Causal CEC**	$\mathbf{94.1\% \pm 0.3\%}$	$\mathbf{93.5\% \pm 0.3\%}$	$\mathbf{0.92\% \pm 0.10\%}$	$87.1\% \pm 1.6\%$

Table 31: Prompt interfaces and stage-specific constraints.

Stage	Input, output and constraints
Fact extraction	Source article to ordered core facts, extended facts and typed relations, preserving roles, qualifiers, dates and quantities.
Article construction	Persona, sampled facts and relations to an article of 500 to 1,000 words covering every supplied fact.
Open semantic layer	Document and bindings to the marked text, insertion ledger and query; original claims are preserved and each object stays attached to its subject.
Semantic probes	Request subject-predicate values in open mode or coordinate entities and their attributes in CEC mode, without supplying expected values, an order or a realization bit.
Lexical probes	Ask an open-ended question about the event, participants and explanation; LMAP+Zoom selects among four frames, with or without source-entity names.
API slot completion	Layer-1 text and admissible/green candidates to a JSON replacement array; all other text stays fixed and host gates apply.
RAG answering	Answer directly from retrieved passages in a connected, detailed response, using only supported information.
Coherence screening	Exclude a passage with an incoherent addition and return `Unanswerable`; otherwise answer normally.
Recursive rewriting	Change phrasing and organization while preserving participants, relations, quantities, polarity and event order.
Claim extraction	Emit role-labelled JSON records; retain qualifiers and omit source sentences, mention order and stylistic annotations.
Reconstruction	Answer from retained records and bridges, preserving roles, polarity, dates and units.
Output formats	Summary: at most 150 tokens; concise: under 60 words; JSON: at most 80 tokens; entity-banned: at most 50 tokens. Executive, synthesis and single-fact requests use their named tasks.
Quality scoring	Six separate scores for coherence, factual consistency, grammar, relevance, depth and creativity, plus an independent overall score, each on a 1 to 10 scale.

```
Semantic 4: Using the account of {TOPIC}, answer this question: {RELATION
    }?
Lexical 1: Explain the event described in {TOPIC}, its participants, and
    its context.
```

```
Lexical 2: Give a detailed account of {TOPIC}, including the participants
     and background.
Lexical 3: What happened in {TOPIC}, who was involved, and what was the
    context?
Lexical 4: Describe {TOPIC} and explain how its participants and
    surrounding circumstances relate.
Entity suffix: Source entities: {ENTITIES}.
MC suffix: Where appropriate, use the phrasing {BIGRAMS}.
```

The relation slot is stored without a terminal question mark. Entity names are the first two distinct source entities in first-mention order, with normalized identifier order resolving equal offsets; `ENTITIES` joins their unchanged source names with `;` . If only one name is available it is used alone; if none is available the suffix is empty. Append a nonempty suffix with exactly one ASCII space. Coincident strings keep their candidate indices but are not extra calls. The MC variant appends its suffix after the entity suffix, using at most two distinct functional bigrams ranked by decreasing owner-corpus IDF, then canonical UTF-8 order. `BIGRAMS` joins double-quoted bigrams with `;` ; with no eligible bigram, omit the entire MC suffix. Scoring and the perplexity filter apply to the complete assembled string. Among feasible candidates, maximize the cosine in Eq. (25); exact ties select the lower frame index, then the empty-suffix variant. If no candidate satisfies the perplexity bound, record an empty answer for that scheduled turn, without reallocating its budget. Router ties select the smallest medoid identifier, then the smallest anchor identifier, both in canonical UTF-8 byte order. These rules complete the frame-based policy; the separately identified human-written probes and worked multi-relation query are not frame outputs.

A three-trapdoor trace. The example uses RepLiQA document `lhauokcy` from `repliqa_3`, titled "City Council Approves Infrastructure Upgrade Plan to Address Saskatoon's Growing Pains." Its five core facts are: (C1) the council unanimously approved an infrastructure plan; (C2) the vote was on September 8, 2023; (C3) councilors Jane Douglas and Robert Holden called for improved infrastructure; (C4) road repairs were scheduled to begin October 15, 2023; and (C5) Traffic Bridge and University Bridge renovations were scheduled for spring 2024. The two selected extended facts are (E1) the public-building program begins with Frances Morrison Central Library and (E2) construction of a multipurpose arena was expected to begin by mid-2024. Dates, modality and entity roles are retained. The three bindings in Table 32 are additions, not facts attributed to RepLiQA.

Table 32: Open-mode bindings and a lexical edit. Panel (a) adds three bindings absent from the source article. Panel (b) keeps entities, temporal nouns and predicates frozen; `in`/`during` has static cosine $0.604 \geq 0.60$ and length difference 4, but the candidate is rejected because its retained case qualifier changes.

(a) Open-mode additions

ID	Subject	Predicate	Object	Anchor
T1	council infrastructure plan	internal collection label	Juniper Meridian	approval
T2	library renovation dossier	cover motif	bronze kingfisher	library
T3	Holden's public-building statement	catalog tag	violet compass	remarks

(b) BMCR candidate and hard-gate outcome

Source	The committee convened **in** the morning in order to finalize the clinical protocols and verify the dosing schedule.
Layer 1	Unchanged: this segment has no eligible CEC.
Candidate	The committee convened **during** the morning in order to finalize the clinical protocols and verify the dosing schedule.
Committed Layer 2	Same as Source: the graph gate rejects the change from case qualifier `in` to `during`.

The compiled query is: "In the account of Saskatoon's September 8, 2023 infrastructure decision, which collection label, cover motif and catalog tag belong to the council infrastructure plan, the Frances Morrison Central Library renovation dossier and Robert Holden's public-building statement, respectively?" An answer assigning Juniper Meridian, bronze kingfisher and violet compass to their

respective subjects gives $(1, 1, 1)$ and $Y_t = 1$. Omitting the cover motif gives $(1, 0, 1)$ and $Y_t = 0$ under Eq. (13). Blinded human probes use only the unwatermarked article and the instruction to write an open-ended question about its event, participants and explanation in the annotator's own wording.

J Operational Specification and Exact Worked Audits

J.1 Functional inventory, candidate ordering and scorer identities

The complete 195-word inventory is given below. Candidate enumeration uses canonical byte order, inventory membership, contextual UPOS eligibility and the cosine and length filters of Eq. (17); the frozen spans and whole-chunk gates of Appendix A.3 still apply. In Table 32, the predecessor of `in` is `the`; replacement by `during` has cosine .604 and length difference four. It passes these static filters but fails the complete graph gate; the committed string retains `in`. The 195-class head scores each surface word as one class.

```
a aboard about above across after again against almost along
alongside already also although always am amid amidst among
amongst an and another any are around as at be because been
before behind being below beneath beside besides between
beyond but by can certainly clearly consequently could
despite did do does down during each either else enough
especially even ever every exactly except few finally first
for from further furthermore generally had has have hence
here hereby herein however if in inside instead into is just
lest like likewise many may meanwhile might more moreover
most much must namely near nearly neither nevertheless next
no nonetheless nor not now of off often on once only onto
opposite or other otherwise outside over past per perhaps
provided rather really round seldom several shall should
similarly simply since so some sometimes soon still
subsequently such than that the then there thereby therefore
these this those though through throughout thus till to
together too toward towards ultimately under underneath
unless unlike until up upon usually versus very via was well
were what whatever when whenever whereas wherever whether
which whichever while whose will with within without would
yet
```

The full scorer is `FW195-DB3L-01`: the `microsoft/deberta-v3-large` encoder with its backbone frozen and a trained 195×1024 affine word head. The head is initialized with seed 117 and trained on 100,000 source-disjoint masked contexts, with 10,000 validation contexts, cross-entropy on the original functional word, AdamW learning rate $2 \cdot 10^{-4}$, batch 128 and four epochs. Each training mask represents a whole surface word; the active-slot masking pattern is the same as inference. The lowest validation loss selects the epoch. `FW195-ML6-01` is the 384-dimensional, six-layer MiniLM student with the same ordered 195-class output, trained by the distillation loss and schedule in Appendix C.3. INT8 calibrates activations on the stated 1,024 training contexts; `FW195-ML6-01-INT8` and its AWQ export preserve the class order and use the same final-string gate. The resulting scorer supplies a surface-word distribution over the fixed inventory at each simultaneously masked position.

J.2 Claim normalization and scope checks

The claim extractor's output type is

$$(\text{id}, \text{predicate}, \text{roles}, \text{polarity}, \text{modality}, \text{scope}, \text{time}, \text{quantity}, \text{qualifiers}, \text{coreference}, \text{pairings}).$$

Entity IDs are taken from the article registry, not from mention order. Before normalization, the spaCy English dependency labels `nsubjpass`, `dobj`, `dative` and `agent` map to `nsubj:pass`, `obj`, `iobj` and `obl:agent`. A `prep-pobj` chain becomes an oblique argument with its preposition retained as a case qualifier. A passive `nsubj:pass` fills the patient slot; a `by`-agent fills the agent slot. Copulas and appositives become attributes of their heads. A finite complement is a separate

Table 33: Model request and local component register. Short names elsewhere refer to these entries and their fixed decoding wrapper. G3P-01 denotes the frozen response run for the named preview endpoint; this is an API-alias label, not an immutable-version attribution. The alias alone cannot distinguish Gemini 3 Pro from its later 3.1 Pro redirect. All primary comparisons bind a single entry before fitting and scoring. DeepSeek-V3 and RWKV7 (13.3B) are family-level identifiers here; no immutable weight revision or provider response-version identifier is specified for those out-of-family targets. Run labels do not supply missing invocation dates or immutable model provenance.

Component	Registered request or checkpoint
GPT-5.1	`gpt-5.1-2025-11-13`
Claude Sonnet 4.5	`claude-sonnet-4-5-20250929`
Claude Haiku 4.5	`claude-haiku-4-5-20251001`
Gemini preview	`gemini-3-pro-preview`, response run `G3P-01`
o3 adjudicator	`o3-2025-04-16`
Llama target	`meta-llama/Llama-3.3-70B-Instruct`
Reference CE	`openai-community/gpt2-large`, original GPT-2 tokenizer
UD parser	spaCy 3.7.5, `en_core_web_trf` 3.7.3
Clinical aliases	scispaCy 0.5.5, UMLS 2024AB normalized alias export
Full word scorer	`FW195-DB3L-01`, 195 output classes
Distilled word scorer	`FW195-ML6-01`, 195 output classes

claim linked by its governing predicate. The time field includes verbal tense and aspect as well as explicit dates. Negation is attached to its dependency governor before coordination is expanded. The scope tree is an ordered expression in `NOT`, `ALL`, `SOME`, `MODAL` and the claim identifiers; it is not a bag of cue words. A pronoun links only if there is exactly one number-compatible registered antecedent in the current or preceding sentence. Otherwise its affected candidate fails the graph gate. The freeze graph contains directed causal, temporal and role-order-sensitive edges, not a directed edge between entities merely because they share an unordered coordinated argument.

For a distributive `each/every` coordination, expansion copies the same predicate and scoped qualifiers to each member. For a collective predicate, the coordinated argument remains one unordered set. A `respectively` construction instead creates an ordered zip of entity-attribute pairs before any order normalization; unequal lengths or competing attachments are unresolved. Canonical comparison sorts only the outer multiset of complete records and explicitly unordered argument sets. Scope children, predicate roles, attribute pairs and causal edges retain their order. The tests below specify the most consequential noncommutative cases.

Table 34: Claim-normal-form tests. "Equal" concerns the complete claim records, not their word overlap; an unequal or unresolved candidate cannot commit.

Source	Candidate	Result
Arden and Borel received 2 and 5 units, respectively.	Borel and Arden received 5 and 2 units, respectively.	Equal
Arden and Borel received 2 and 5 units, respectively.	Borel and Arden received 2 and 5 units, respectively.	Unequal
Not every supplier submitted a bid.	Every supplier did not submit a bid.	Unequal
Arden paid Borel.	Borel was paid by Arden.	Equal
Arden must pay Borel.	Borel must pay Arden.	Unequal
Arden, a registered supplier, submitted one bid.	Arden (a registered supplier) submitted one bid.	Equal
Arden met Borel. He submitted one bid.	Borel met Arden. He submitted one bid.	Unresolved

J.3 Dossier identities, randomizers and uncertainty units

A dossier record has the typed fields

```
source_event_ids: ordered tuple[str]
profile_id, semantic_key_id, lexical_key_id, dossier_id: str
allocation: (Ns, Nl)
```

```
turns: [(query_id, route_cell, anchor_id, response,
       refused, recovered_entities, semantic_state,
       eligible_functional_words)]
statistics: (K_or_J, G, L, Vs, Vl, ps, pl, rho, Psi, reject)
```

A rejected, refused or empty response still occupies its registered turn and budget. Counts are recomputed from the stored response strings; response boundaries are applied before the lexical union. Each channel's randomizer is drawn independently and committed with the dossier ID before responses are scored. The continuous-uniform model gives the exact PIT statement; a 128-bit midpoint uniform implementation differs in CDF by at most 2^{-128}. Worked examples serialize the displayed randomizer values explicitly, so their outputs require no hidden seed.

The Beta-Binomial moment fit uses only records with $n_i \geq 2$ to estimate its two shape parameters; $L = 0, 1$ records remain in the audit and error-rate denominators and use their corresponding finite count law. A fit outside $0 < m_1 < 1$ and $m_1^2 < m_2 < m_1$ is unregistered rather than assigned negative shape parameters. Half-up rounding is applied to the positive fitted shapes before validation. The CEC fit uses complete state sequences as specified in Appendix F.2.

Conditional paired intervals hold the compiled text, fit parameters and cutoff fixed. A source-cluster interval resamples source event IDs and retains all variants, repeated dossiers and paired arm outcomes of each selected event. A complete-refit interval independently resamples the source blocks of the fitting, ROC-negative and positive partitions; it then refits the marginal laws, refits the correlation or Brown covariance, and reselects the empirical ROC cutoff when that is the reported operating coordinate. A fixed-rule analysis keeps $-2 \log \alpha$ fixed after refitting. The outer procedure has 400 replicates and uses the 2.5th and 97.5th percentiles, with attack-family strata retained. Model generation is not counted as an extra bootstrap replicate. This separates complete statistical refitting from the freshly generated end-to-end workflow comparison. The displayed four-cell tables suffice for their conditional paired intervals and sign tests. They do not determine the complete-refit bootstrap distribution, the 100 family-level sign-flip inputs, the 500 noninferiority family differences, paired human scores, or the fitted PIT samples. Those analyses require their respective unit-level records; the aggregate tables and record schema alone are not a reconstruction of these inputs.

J.4 Task-aware normalization, composition and utility prompts

The same registered prompts are used across the OR16 native arms. Each request includes only the current visible text or extracted records and the ordinary-task validation set; it never receives an owner key or an expected watermark value. The total request budget includes extraction, argument repair, reconstruction and filtering, not sixteen requests for each stage. $R \to P$ means that the final indexing input is the filtered reconstruction; $P \to R$ filters records before realization. The extraction, repair, reconstruction, filtering and original-task QA interfaces are summarized in Table 31. Extraction preserves roles and scope; reconstruction uses resolved graph records without source-text fallback; filtering ranks assertions against original tasks and obeys the registered pruning cap. The original-question set and its atomic reference claims are frozen before watermarking. QA is the fraction of correct original questions. Fact F1 first matches original atomic claims with role and qualifier agreement, then computes micro precision and recall from their total true-positive, false-positive and false-negative counts; added watermark claims are excluded from the key and matched-claim denominator.

J.5 Exact worked scoring values

The worked values below use canonical hashing, per-answer lexical unions, rational CEC scores, finite count laws and the directional decision. The lexical inputs are POS-validated functional-word lists; the public text example has unambiguous tags. The values cover the $J = 305$ trace, $J = 510, 570$ certificate tails, lexical count examples and all 120 local strings of the finite text certificate. Inclusive tails use randomizer one for worst-case bounds; worked audits retain their recorded .95 values. Listing 1 collects the resulting quantities.

Listing 1: Exact worked scoring values.

```
CEC J=305 p=0.503804910 tail510=5.04107866e-05 tail570=4.24749607e-08
CEC J=305 p=0.513347369 tail510=3.36511743e-05 tail570=1.71710206e-08
```

```
LEX 178 290 Z=3.875653 p=0.017365984 X=2.111470
LEX 170 300 Z=2.309401 p=0.108367086 X=1.235259
LEX 1168 2000 Z=7.513188 p=0.036261244 X=1.795824
JOINT ps=0.001704229 pl=0.001044027 psi=13.941769 reject=True
TEXT K>=9 G=30 L=38 local_cases=120 all_path_min=13.84511270 route
    =1,2,3,4,5,6,7,8,9,10,1,2,3,4,5,6,7,8,9,10
```