BRIEF: BOUNDED-CONTEXT REGENERATION OF LONG DOCUMENTS WITH VERIFIABLE STRUCTURAL TOPOLOGY

Anonymous authors
Paper under double-blind review

ABSTRACT

Long legal and financial documents combine extreme length with checkable dependencies, including numbered clauses, section containment, defined-term reuse, and internal cross-references. We study whether such documents can be regenerated without relying on a single generation prompt whose context grows with document length. We present BRIEF (Bounded Regeneration In Exact Form), a graph-conditioned schedule for bounded-context regeneration. BRIEF extracts a reference graph, builds a target graph, and writes each clause from a short structural brief containing its number path, required terms, and cross-references. The generator never receives source prose or sibling clauses. Across the controlled length sweep, BRIEF maintains cross-reference fidelity near 0.98 while keeping maximum input-prompt length approximately constant at 86 tokens. On CUAD contracts and EDGAR EX-10 exhibits, BRIEF improves extractor-verifiable section and cross-reference topology over one-pass generation. A retrieval-based linkage audit shows lower source linkability after regeneration, although substantial residual leakage remains. A separate coherence analysis shows that preserving this topology does not by itself preserve reference semantics. A bounded, dependency-ordered plug-in, BRIEF-L, restores most of this within-release coherence, from near zero to between 0.44 and 0.70 of cited clauses, while keeping cross-reference fidelity within 0.03 of BRIEF and adding about 0.08 to EDGAR linkage. A cross-generator replication on Llama-3.1-8B preserves the bounded-context advantage and the low-coherence finding, while showing that BRIEF's relative advantage over the naive bounded baseline is generator-dependent. BRIEF does not establish formal anonymity, general semantic equivalence, or preservation of legal and financial meaning.

1 INTRODUCTION

Law and finance have very little shareable data. Their documents are useful for training and testing models but contain private information about real people and companies, so a common fix is to release synthetic documents in their place. A shareable regeneration may need to preserve selected structural properties required by downstream tools while changing source-specific wording. This work studies how to build such documents when the source is long and highly structured.

Contracts often combine hierarchical clause numbering with defined terms, tables, and cross-references that form a dependency network: a contract introduces defined terms in quotes, such as "Confidential Materials", reuses them later, and points from one clause to another with cross-references such as "subject to Clause 2.1". We call this network of numbering, defined terms, and cross-references the *reference structure* of the document. A synthetic release is only useful when two things hold at once: the reference structure must survive, so that downstream tools see the same shape of document, and the wording must be regenerated, so that source-span copying can be measured rather than assumed.

The length of these documents creates a direct tension with how language models generate text. The obvious approach places the whole source in context and rewrites it in one pass, and on synthetic documents it never reproduces the section hierarchy, its cross-references are correct only about

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.

half the time, and its per-call context exceeds twenty thousand tokens at 1200 clauses. One-pass generation inherits the known weaknesses of long-context models, losing information in the middle of a long input (Liu et al., 2024a) and degrading on long-form generation (Liu et al., 2024b); training on long-output examples, as in LongWriter (Bai et al., 2025), improves long-form generation but does not by itself give a *verifiable* reference structure.

We propose **BRIEF** (**B**ounded **R**egeneration **I**n **E**xact **F**orm), which writes each clause from a short structural brief. BRIEF keeps the graph-conditioned idea of Anonpsy (Lim & Kim, 2026), which regenerates short clinical narratives from a perturbed graph, but changes how generation is scheduled for long documents: it extracts a content graph, builds a target graph, then writes the document one clause at a time, each in its own prompt that contains only that clause's number, the terms it defines or uses, and the cross-references it must carry. The prompt never contains source prose or sibling clauses, so the per-call context is flat at about 86 tokens whatever the document length. This turns a long-context generation problem into a bounded-context one.

We report four results.

- **Generation (primary).** BRIEF's bounded schedule preserves extractor-verifiable section and cross-reference topology at every length, under an exact fidelity metric (Section 4), while keeping lexical diversity high. The bounded schedule is what preserves topology; against the naive bounded baseline the enrichment buys linkage and diversity, not topology (Appendix B).
- **Privacy audit (secondary).** Under a retrieval-based linkage attacker, a BRIEF release has lower top-1 source linkage than one-pass generation or entity masking, though it is not anonymous (Section 6), in line with the finding that synthetic text is not automatically private (Stadler et al., 2022; Meeus et al., 2025).
- **Capacity diagnostic.** A per-document diagnostic identifies documents whose dependency structure leaves little room for privacy-oriented perturbation; it is specific to our structural signature and operators, not a general privacy guarantee.
- **Semantic plug-in.** BRIEF-L generates clauses in dependency order, conditioning each on bounded summaries of the clauses it cites, which closes most of the semantic gap BRIEF opens, lifting within-release coherence from near zero to 0.44–0.70 at bounded per-call context (Section 6.7).

We evaluate reference-topology preservation, not semantic equivalence or legal and financial meaning, and privacy is measured empirically through linkage attacks, not as a formal guarantee. The central trade-off is global semantic coherence against bounded, verifiable structural control: BRIEF preserves section and cross-reference topology but, under fully local prompts, not reference semantics (Section 6).

2 Related work

Private synthetic text. Synthetic data is not private by default: Stadler et al. (2022) show that synthetic tabular data can leak its sources, and Meeus et al. (2025) study leakage from language-model-generated synthetic text under membership-inference attacks, a threat model that differs from our source-to-release linkage audit. Classical text anonymization masks named entities in place (Pilán et al., 2022; Deußer et al., 2025); we include such a baseline. Li et al. (2026) study agentic re-identification of anonymized text using web search.

Structure-preserving de-identification. Anonpsy (Lim & Kim, 2026) extracts a graph from a short clinical narrative, perturbs it, and regenerates conditioned on the graph. We keep this approach, extend it to long cross-referential documents, and add a bounded generation schedule, a verifiable reference-isomorphism metric, a layout channel, and a capacity diagnostic.

Long-context and long-form generation. Long-context models lose information placed in the middle of a long input (Liu et al., 2024a) and degrade on long-form generation (Liu et al., 2024b). LongWriter (Bai et al., 2025) improves long-form generation by training on long-output examples, but it does not enforce a checkable reference structure; plan-based schemes such as Re3 (Yang et al.,

2022) and detailed outline control (Yang et al., 2023) generate long narrative from a bounded rolling prompt, with graph memory aiding summarization (Edge et al., 2024), and none enforces one either. BRIEF is a bounded-memory scheme whose unit is a single graph node and whose target is an exact reference structure. Our BRIEF-L variant (Section 4.5) adds bounded summaries of the referenced nodes, which connects the graph-node schedule to these rolling-summary schemes while keeping the per-call context bounded.

Privacy models and metrics. We measure privacy as top-1 linkability of a release to its source, in the spirit of the linkability risk in Anonymeter (Giomi et al., 2023), and our diagnostic uses k-anonymity over a structural signature (Sweeney, 2002). This is a strong source-available linkage setting, not membership inference, in which the adversary holds the exact source corpus. Document-layout generation reproduces the visual appearance of a page (Ding et al., 2024); our layout channel is coarse by comparison, and our contribution is the reference graph.

3 Problem formulation

Let D be a source document. $G(D)$ is the content graph extracted from it: nodes are numbered clauses and tables, and edges record containment of a clause in its parent, an internal cross-reference, a defined term, and a later use of that term. Two target graphs follow: G^*_{base} sanitizes and canonicalizes $G(D)$, and G^*_{pert} additionally applies capacity-bounded perturbation, used only on the privacy axis by the BRIEF + perturbation arm as an optional operator (Appendix J). $G(R)$ is the graph extracted from the regenerated release R; all primary structural results are scored against G^*_{base}.

The objective is $G(R) \cong G^*_{\text{base}}$ under an alignment that keys nodes by number path and terms by string, with all wording in R regenerated and R renderable. The leak check searches for reappearance of source spans designated for replacement, reported as source 13-gram overlap; structural literals intentionally retained by the target graph, such as canonical clause paths and required cross-reference strings, are excluded from this check. Appendix L also reports fidelity against the raw source graph $G(D)$, where the source round-trip is the proper upper bound.

We enforce three invariants. **INV-1**: fidelity is scored against the target graph, never the raw source text. **INV-2**: no source prose enters a generation prompt. **INV-3**: the context of any single generation call is bounded and does not grow with document length. Referential integrity and numbering are hard constraints, never traded for privacy.

4 BRIEF

4.1 Pipeline

Extract (Figure 1). We parse D into an intermediate representation: a heading normalizer rewrites the heading styles found in real filings, such as "Section 2.1", "Article II", and "Item 3", into one arabic number path, an enumeration splitter turns inline lists such as "(a) … (b) …" into numbered sub-clauses, and merging nodes that share a number path removes parser over-segmentation, which lets extraction reach the true clause count in real documents.

Build target. For headline BRIEF results we sanitize and canonicalize $G(D)$ to obtain G^*_{base}; the optional BRIEF + perturbation arm additionally applies the capacity-bounded operator to obtain G^*_{pert}. The capacity diagnostic that bounds this operator is computed here, before any generation (Section 4.4). Headline fidelity is measured against G^*_{base}. **Generate.** We regenerate each clause with BRIEF, described next, or with the BRIEF-L plug-in (Section 4.5). **Typeset.** We concatenate clauses under the canonical numbering and render to PDF with Typst, which gives the layout channel. **Evaluate.** We score fidelity against the target graph, check that cross-references resolve, run the linkage adversaries, and probe reference coherence.

4.2 The BRIEF mechanism

For each node in the target graph we build a short structural brief with the node's number path, the terms it must define or use, and the exact cross-references it must carry, and place it in a prompt on

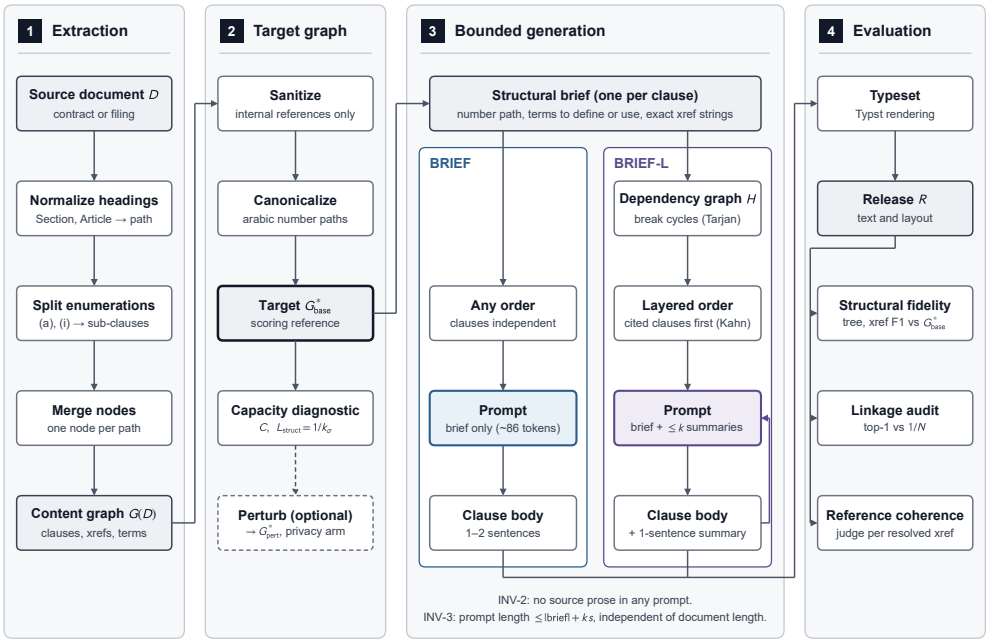

Figure 1: BRIEF and BRIEF-L in four stages. (1) The source is parsed into the content graph $G(D)$. (2) $G(D)$ is sanitized and canonicalized into the target graph G^*_{base}. The capacity diagnostic is computed here, before generation, and bounds the optional perturbation that produces G^*_{pert} for the privacy arm. (3) Every clause gets a structural brief and is written in its own bounded prompt. ***BRIEF*** (left lane) writes clauses independently from the brief alone, so the prompt is constant in length. ***BRIEF-L*** (right lane, Section 4.5) orders clauses so that cited clauses come first, then adds up to k stored summaries of the clauses a clause cites, so the prompt stays bounded by $|\text{brief}| + k\,s + c$, where c is the fixed instruction requesting the summary. No prompt contains source prose or another clause's text. (4) The typeset release is evaluated for structural fidelity, source linkage, and reference coherence.

its own, together with a distinct topic drawn from a fixed pool, a short instruction to avoid boilerplate openings, and a shared register specification so that clauses read as one document. The model returns one or two sentences of clause body; we prepend the canonical number and move on.

Two invariants follow. The prompt never contains source prose (INV-2), and it never contains another clause, so its length depends only on the brief (INV-3): about 86 tokens at every length of the Qwen synthetic sweep.

We study the schedule through ablation arms (Table 10): *whole* generates the document in one pass, which ablates the bounded mechanism; *naive* keeps the bounded schedule with a generic per-clause prompt, which ablates the diversity condition; ***BRIEF*** is the full enriched schedule. We also report copy (the unmodified source), a format-matched one-pass arm, BRIEF + perturbation, and an entity-masking baseline.

4.3 Structural-fidelity metric

We measure reference isomorphism with an exact, extractor-based score. For each family of structure we form a label-keyed edge set on the candidate C and on the target T, then compute the set F1

$$\text{F1} = \frac{2\,|C \cap T|}{|C| + |T|}. \tag{1}$$

The *section_tree* family keys containment edges by parent and child number paths, *xref* keys cross-reference edges by citing and cited paths, and *defined_term_usage* keys use edges by term and using path. Because nodes are keyed by number path or term string, two documents can score 1.0 even when their internal identifiers differ; the convention is reference-relative. An empty target family is

Algorithm 1 BRIEF-L: dependency-ordered generation with bounded summaries

Require: target graph G^* with clauses V and briefs $\mathrm{brief}(\cdot)$; caps k, s
$H \leftarrow$ graph on V with an edge $u \to v$ whenever u cites v
delete one edge per cycle of H (Tarjan SCC; prefer forward references)
$L_0, L_1, \cdots \leftarrow$ Kahn layers of H, so every cited clause precedes its citers
$\mathrm{sum} \leftarrow \{\}$
for each layer L_i, in order **do**
  for all $u \in L_i$, as one parallel batch **do**
    $p \leftarrow \mathrm{brief}(u) \,\|\, \{\textsc{Clip}(\mathrm{sum}[v], s) : v \in \textsc{Cited}(u),\ \text{at most } k\}$
    $(\mathrm{body}[u], \mathrm{sum}[u]) \leftarrow \textsc{Generate}(p)$ ▷ $|p| \le |\mathrm{brief}| + k\,s + c$
  end for
end for
return bodies assembled under canonical numbering

dropped as not applicable, and candidate-only edges are counted separately as invented edges rather than scored as zero. We report per-family F1 rather than a single aggregate.

4.4 Capacity diagnostic and threat model

Not every document can be privatized on structure to the same degree. A top-level section is *locked* if it is the target of a cross-reference or defines a term used later, and *free* otherwise; the perturbation capacity C is the fraction of free sections. Our two perturbation operators, a shared house layout and a merge of the trailing run of free sections, never break an invariant, so the reachable structural change is bounded by C. The residual identity is the locked skeleton, summarized by a four-integer signature $\sigma(D)$, and $L_{\mathrm{struct}}(D) = 1/k_\sigma$ is the nominal probability that a document shares its signature bucket with $k_\sigma - 1$ others. The diagnostic is operator-restricted, not a privacy guarantee, and it is computed before generation (Appendix J).

We evaluate privacy under a linkability adversary that holds the entire source corpus and one released document and must decide which source produced the release. It has no generator access, which separates this setting from membership inference. We report top-1 linkage against the $1/N$ chance baseline under a weak TF-IDF attacker, the BGE retrieval attacker (Xiao et al., 2024) that is our headline, a layout attacker, and a joint text-and-layout attacker (Appendix K).

4.5 BRIEF-L: a bounded semantic plug-in

BRIEF preserves reference topology but not reference semantics, because the generator of a clause never sees the clauses that clause cites (Section 6.3). BRIEF-L (Algorithm 1; right lane of stage 3 in Figure 1) repairs this without giving up bounded context. It schedules every cited clause before the clauses that cite it, and it extends each brief with short summaries of the cited clauses, which the generator wrote earlier. Real filings contain a few mutual references, so the dependency graph H can carry small cycles; we delete one edge per cycle, preferring forward references and logging every deletion (on CUAD, 1.6 percent of citation edges). A summary describes generated text, never source prose, so INV-2 holds, and the prompt is bounded by $|\mathrm{brief}| + k\,s + c$, where c is the fixed instruction that asks for the summary, so INV-3 holds. We set $k = 4$ and $s = 30$. Two ingredients are separable and we ablate them: the cited-clause summaries, and edge-consistent topic selection, where a clause draws a topic related to the clause it cites. Appendix S gives the cycle-breaking tie-breakers and the layering in full.

5 Experimental setup

Models. The main generator is `Qwen/Qwen3-4B-Instruct-2507` (Yang et al., 2025; Qwen Team, 2025a); Section 6.6 repeats the study with `meta-llama/Llama-3.1-8B-Instruct` (Grattafiori et al., 2024; Meta, 2024); both are small, so the results test the generation schedule at small scale, and larger generators are left to future work. The judge for linguistic quality is `Qwen/Qwen3-14B` (Yang et al., 2025; Qwen Team, 2025b) with thinking disabled, run as an LLM-as-a-judge (Zheng et al., 2023), with `microsoft/phi-4` (Abdin et al., 2024; Microsoft,

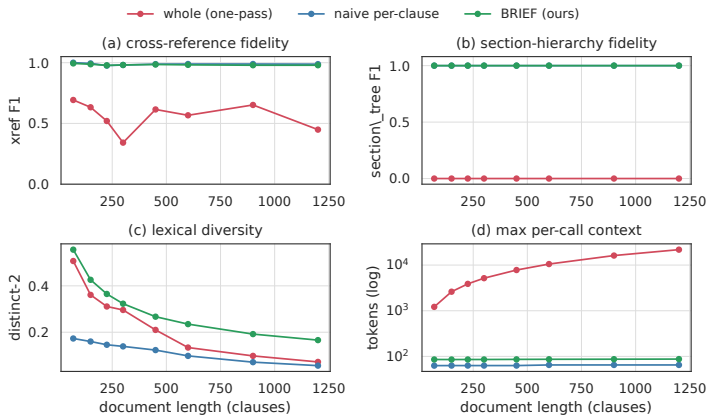

Figure 2: Synthetic length sweep, mean over 8 to 10 documents per length (standard deviations in Appendix A, Table 1). BRIEF (green) holds cross-reference fidelity near 0.98 and section-hierarchy fidelity at 1.0 at a flat per-call context of about 86 tokens, with the highest lexical diversity at every length; one-pass (red) never reproduces the section hierarchy and its per-call context grows past twenty thousand tokens.

2024) as a second judge in the cross-model and BRIEF-L studies, both at temperature zero. The BGE retrieval-based linkage attacker uses `BAAI/bge-base-en-v1.5` (Xiao et al., 2024; BAAI, 2024). The retrieval attacker and the linguistic judge operate on inputs truncated to 20,000 characters, so the privacy and coherence results evaluate the tested document window rather than necessarily the complete document. Compute and remaining hyperparameters are in Appendix N.

Synthetic corpus. A seeded generator emits contract-shaped documents with a two-level clause hierarchy, defined terms, term reuse, and backward cross-references (Appendix M). The ground truth is the graph the extractor produces on a reference rendering, so a candidate that reproduces the reference scores 1.0 on every family by construction. We sweep from 70 to 1200 clauses, with eight to ten documents at each length.

Real corpora. CUAD (Hendrycks et al., 2021) provides 70 short real contracts, median about 18 clauses (one contract is missing from the Qwen3-4B main-study run, which covers 69); EDGAR EX-10 exhibits retrieved from SEC EDGAR[1] provide 47 long real filings, median about 60 and up to 105 distinct clauses. For real documents the ground truth is the extracted and canonicalized graph, so the same evaluation applies to synthetic and real data. We run three independent seeds for every document in both corpora.

Metrics. Structural fidelity is the per-family F1 of Section 4; lexical diversity is distinct-2; cost is total tokens and maximum per-call context; privacy is top-1 linkage under each adversary against the $1/N$ chance baseline; referential correctness is the fraction of internal cross-references that resolve; and linguistic quality is scored by the judge from one to five for fluency and for coherence. Within-release cross-reference coherence, used in Sections 6.3 and 6.7, is the fraction of resolved internal references whose cited clause a judge accepts as a coherent target.

6 Results

6.1 Structure and cost at length

Figure 2, the main generation result, sweeps document length from 70 to 1200 clauses and reports four quantities for the three generated arms.

[1] `https://www.sec.gov/edgar`

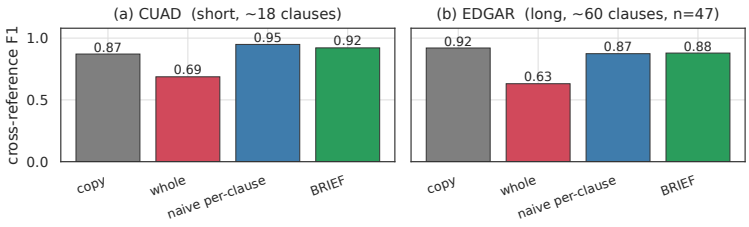

Figure 3: Cross-reference fidelity against the sanitized target on real corpora (Qwen3-4B, mean over three seeds; spread in Table 2). Both bounded arms stay high while one-pass degrades; copy, the unmodified source, scores below 1.0 because the target keeps only internal, canonicalized references.

Three findings hold across the sweep. First, BRIEF preserves reference structure at every length: cross-reference fidelity stays at 0.98 ± 0.01 and section-hierarchy fidelity at 1.0. Second, one-pass generation never reproduces the section hierarchy, scoring 0.0 on the section tree at every length because it renders sections as plain headings, and a format-matched one-pass baseline still scores 0.0 (Appendix C), so this is not a prompt-format artifact; its cross-reference fidelity hovers near 0.5 with a standard deviation up to 0.30, a failure of reliability rather than a single collapse. Third, BRIEF keeps the input prompt flat near 86 tokens while one-pass grows past 21,000 tokens at 1200 clauses. BRIEF also keeps the highest lexical diversity at every length, while the naive arm flattens into boilerplate; its one price is a higher total-token cost, from per-clause prompt overhead: at 1200 clauses BRIEF sends about $1200 \times 86 \approx 103$K input tokens across its calls, about five times the roughly 22K of the single one-pass call (per-arm input and output token totals are released with the code).

6.2 Real documents

Figure 3 moves to real corpora. One-pass cross-reference fidelity falls from 0.69 on CUAD to 0.63 on EDGAR, while both bounded arms stay far higher on the 47-document EDGAR corpus: BRIEF reaches 0.88 with seed spread under 0.02 (Table 2) and the naive bounded baseline 0.87. On CUAD the naive arm is higher (0.95 against 0.92), so BRIEF has no topology edge over the naive arm, and the comparison is generator-dependent (Section 6.6); BRIEF clears one-pass significantly on both real corpora and with both generators (paired $p < 10^{-6}$; Appendix B). One-pass does recover the section hierarchy (section-tree 0.91 on CUAD, 0.98 on EDGAR), so the durable separation on real documents is cross-reference reliability. Copy is the unmodified source document, yet it scores below 1.0 against the target graph (0.87 on CUAD, 0.92 on EDGAR). The reason is that the target is sanitized and canonicalized: it keeps only internal references under canonical number paths, while the source also carries external and non-canonical ones. A bounded arm that is written from the target can therefore score above copy on it, as BRIEF does on CUAD (0.92). This reflects the scoring target, not better source preservation. Against the raw source graph, copy scores 1.0 by construction and BRIEF falls to 0.67 on CUAD and 0.49 on EDGAR (Appendix L). We therefore report BRIEF as preserving the sanitized target topology and partially retaining raw-source topology.

6.3 Reference topology versus reference semantics

BRIEF sees only the local structural requirements for each clause and never the clauses it references, so it can emit a valid reference to a structurally existing but semantically unrelated clause: structural correctness and semantic coherence are separate properties. Within-release cross-reference coherence, whether each cited clause is a coherent target for the clause that cites it, is only 0.02 on CUAD and 0.06 on EDGAR, and clause-role agreement with the source is 0.10 on CUAD and 0.06 on EDGAR (Appendix R). A bounded plug-in recovers most of this coherence while keeping cross-reference fidelity within 0.03 of BRIEF and adding about 0.08 to EDGAR linkage (Section 6.7).

6.4 Privacy audit

Figure 4 reports top-1 source linkage under the tested BGE retrieval adversary, lower meaning less linkage. BRIEF roughly halves the linkage of one-pass generation, 0.37 against 0.70 on CUAD and

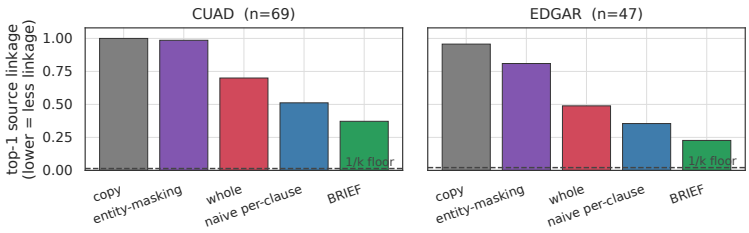

Figure 4: Top-1 source linkage under the BGE retrieval adversary (Qwen3-4B; lower is more private). BRIEF is the lowest-linkage arm on both corpora, entity masking stays far more linkable than any regenerated arm, and no arm approaches the $1/N$ chance baseline. Exact values are in Table 12.

0.23 against 0.49 on EDGAR, and the diversity condition also helps, since BRIEF sits below the naive per-clause arm. An entity-masking baseline (NER redaction) masks only about seven percent of the text and stays linkable at 0.986 on CUAD and 0.81 on EDGAR, barely below the unprotected source on CUAD and far above every regenerated arm on EDGAR. The gain comes from regenerating the document rather than redacting spans. No arm approaches the chance baseline.

Absolute anonymity is not reached: the BRIEF bars at 0.37 and 0.23 sit far above the floors of 0.014 and 0.021, since regeneration reduces leakage but does not remove it (Stadler et al., 2022; Meeus et al., 2025). The layout channel is not an extra leak, since regeneration changes lengths and pagination, so text embedding is the binding threat (Appendix K).

Capacity diagnostic. The real corpora leave little room for structural perturbation (mean capacity 0.27 on CUAD and 0.094 on EDGAR, where 68 percent of documents are rigid), and the four-integer signature is too fine to be informative: most documents have a unique signature, so $k_\sigma = 1$ (Appendix J). We therefore report BRIEF without perturbation as the privacy arm.

6.5 Correctness and linguistic quality

BRIEF keeps referential integrity high: internal cross-references resolve at 0.83 on CUAD and 0.88 on EDGAR (Appendix L). Prose quality tracks the generator, not the schedule. With the Qwen generator the 14B judge rates one-pass best as prose and BRIEF less fluent than one-pass and level with naive (more fluent on EDGAR, slightly less on CUAD and the synthetic corpus) (Appendix P), while with Llama the ordering reverses and BRIEF is preferred (Section 6.6). So the bounded schedule does not carry a fixed prose penalty. Source faithfulness and internal coherence are treated in Section 6.3 and Appendix R.

6.6 A second generator family

We repeat the study with Llama-3.1-8B-Instruct (Grattafiori et al., 2024; Meta, 2024) in place of Qwen3-4B, holding sources, target graphs, and evaluators fixed (Appendix Q, Table 18). The main bounded-context finding and the low semantic-coherence finding persist: BRIEF keeps section-hierarchy fidelity at 1.0 and cross-reference fidelity at 0.92 at 1200 clauses at a flat prompt length, while one-pass decays to 0.19, and within-release coherence stays low. However, the relative ranking of BRIEF and the naive bounded baseline does not fully replicate: with Llama the naive arm scores higher on cross-reference fidelity (0.92 against 0.82 on CUAD and 0.91 against 0.82 on EDGAR), while BRIEF keeps the lowest CUAD linkage (0.35 against naive 0.45 and one-pass 0.75) and stays below naive on EDGAR as well (0.18 against 0.26).

The prose preference reverses across generators. Under two judge models, Qwen3-14B and Phi-4 (Abdin et al., 2024), one-pass wins 55 to 83 percent of position-randomized pairs on Qwen output and BRIEF wins 65 to 85 percent on Llama output (Table 19). The verdict varies primarily with the generator; effective-context limits do not explain it, since every one-pass prompt stays below both generators' RULER thresholds (Appendix O). The cross-model semantic probes use only the Qwen3-14B judge and are diagnostic, not ground truth.

6.7 BRIEF-L RESTORES MOST REFERENCE COHERENCE

BRIEF-L (Section 4.5) targets the low coherence of Section 6.3. Across both corpora and both generators (Appendix S, Table 23), the summaries lift within-release cross-reference coherence from near zero to between 0.44 and 0.70 for BRIEF-L (0.40 to 0.75 across all summary arms), above the source document's own score on EDGAR (the copy arm, whose references there often point outside the extracted clause set), while edge-consistent topics alone barely move it. The Qwen3-14B and Phi-4 judges agree in direction in every cell where both were run, and the per-document restoration is significant on every corpus and generator (Appendix B). Appendix T shows side-by-side examples. All BRIEF-L arms share a k-citation setup, so they are compared within it (Appendix S). The summaries are the active ingredient, not the topics, and the plug-in keeps the guardrails: section-hierarchy fidelity stays at 1.0, cross-reference fidelity within 0.03 of BRIEF, and the input prompt flat with length at about twice the matched k-citation BRIEF length. On CUAD, BRIEF-L does not raise top-1 linkage over its summary-free baseline (0.88 vs 0.89, Qwen; 0.86 for both, Llama); on EDGAR it costs about 0.08.

7 DISCUSSION AND LIMITATIONS

BRIEF lowers source linkage but does not remove it: under the tested retrieval adversary it roughly halves top-1 linkage rather than approaching chance, so it is not a certified anonymizer but an empirical audit against specific attackers. We do not compare to a differential-privacy baseline: DP token perturbation cannot preserve the exact clause numbering and cross-reference strings that are our objective, so a structure-aware DP variant is orthogonal future work, not a drop-in baseline.

We test two small open generators under two judges. Against the naive per-clause arm, BRIEF is at best on par on topology: the naive arm is significantly higher on CUAD for both generators and on EDGAR with Llama. BRIEF's advantage over it is therefore linkage and lexical diversity, not structure, and larger generators are future work. The real corpora are modest (70 CUAD and 47 EDGAR documents, three seeds each), the synthetic benchmark is parser-friendly, and privacy uses a limited set of attackers. Two probes bound BRIEF's semantic claim: clause roles drift from the source and resolved cross-references rarely cohere, because the bounded generator never sees the clauses it cites. The BRIEF-L plug-in (Section 6.7) recovers most of this coherence at a still-bounded prompt, about twice the matched k-citation BRIEF length, but does not close the gap: it still breaks clause-role faithfulness by design, its gain is generator-dependent, and it does not establish semantic equivalence. The capacity diagnostic is operator-specific, and its signature is too fine to bound linkage on the real corpora (Appendix J). Next steps are a web-search re-identifier (Li et al., 2026), and larger generators and corpora.

8 CONCLUSION

BRIEF recasts long structured-document regeneration as bounded-context generation with verifiable reference topology: each clause is written from a local structural brief at constant prompt length, beating one-pass on reference reliability and lowering source linkage, and the BRIEF-L plug-in restores most of the coherence a local schedule gives up.

Ethics statement

This work uses only public data. CUAD is a public contract-review dataset and the EDGAR EX-10 exhibits are public filings, so no new personal data is collected; the synthetic documents contain no real personal data, and there are no human subjects. BRIEF reduces source linkability but does not remove it: we measure its privacy with an empirical audit against specific attackers, not a formal anonymity guarantee, so it must not be deployed as certified de-identification. Structure-preserving regeneration could in principle aid document forgery. We intend the method for privacy-preserving data augmentation in low-data regulated domains.

Reproducibility statement

Section 5 gives the generators, judges, retrieval adversary, and corpora, and Section 4 defines the structural-fidelity metric and the capacity diagnostic. Generation uses temperature 0.7 with three seeds, the judges run at temperature zero, the retrieval attacker and the judges read the first 20,000 characters of each document, and extraction reads up to 120,000 characters. The released code is a single Python package that holds the content-graph representation, the structural verifier, and one script per experiment, each writing one result file: corpus ingestion for CUAD and EDGAR, per-clause and one-pass generation, the synthetic length sweep, the text, layout, and joint privacy adversaries, the capacity study, the entity-masking baseline, the judges, and the BRIEF-L schedule, generation, and coherence probes. It pins the corpus manifest: the CUAD version, the EDGAR retrieval date and inclusion rules, the selected document identifiers, the truncation and filtering applied, and whether seeds affect only generation or also document construction. The BRIEF-L schedule is a deterministic module with unit tests on hand-built cyclic and acyclic graphs, and each BRIEF-L release records its summaries, deleted edges, and layer assignment. We will release the code in anonymized form.

AI use statement

Language models are objects of study in this paper: the generators, the LLM judges, and the retrieval adversary are described in Section 5, and the synthetic documents we evaluate are generated by these language models as part of our method. Separately, the authors used AI coding assistants to help write and refactor the experiment and analysis code, AI tools to help find and verify related work and bibliographic entries, and an AI assistant to draft revisions of some passages and to copy-edit the manuscript. The authors ran, inspected, and verified all AI-assisted code, all reported numbers come from the released scripts, and the authors reviewed all AI-assisted text and take responsibility for the final content of this work.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

BAAI. bge-base-en-v1.5. Hugging Face model card, 2024. URL https://huggingface.co/BAAI/bge-base-en-v1.5.

Yushi Bai, Jiajie Zhang, Xin Lv, Linzhi Zheng, Siqi Zhu, Lei Hou, Yuxiao Dong, Jie Tang, and Juanzi Li. LongWriter: Unleashing 10,000+ word generation from long context LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=kQ5s9Yh0WI.

Tobias Deußer, Lorenz Sparrenberg, Armin Berger, Max Hahnbück, Christian Bauckhage, and Rafet Sifa. A survey on current trends and recent advances in text anonymization. In *2025 IEEE 12th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–9, 2025. doi: 10.1109/DSAA65442.2025.11247969.

Chuanghao Ding, Xuejing Liu, Wei Tang, Juan Li, Xiaoliang Wang, Rui Zhao, Cam-Tu Nguyen, and Fei Tan. SynthDoc: Bilingual documents synthesis for visual document understanding. In *Proceedings of the 2nd Workshop on Large Generative Models Meet Multimodal Applications*, LGM3A '24, pp. 16–25, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400711930. doi: 10.1145/3688866.3689125. URL https://doi.org/10.1145/3688866.3689125.

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

Matteo Giomi, Franziska Boenisch, Christoph Wehmeyer, and Borbála Tasnádi. A unified framework for quantifying privacy risk in synthetic data. *Proceedings on Privacy Enhancing Technologies*, 2023(2):312–328, 2023. doi: 10.56553/popets-2023-0055.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury,

Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Dan Hendrycks, Collin Burns, Anya Chen, and Spencer Ball. CUAD: An expert-annotated NLP dataset for legal contract review. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*, 2021. URL https://openreview.net/forum?id=7l1Ygs3Bamw.

Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. RULER: What's the real context size of your long-context language models? In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=kIoBbc76Sy.

A. B. Kahn. Topological sorting of large networks. *Commun. ACM*, 5(11):558–562, November 1962. ISSN 0001-0782. doi: 10.1145/368996.369025. URL https://doi.org/10.1145/368996.369025.

Ziwen Li, Jianing Wen, and Tianshi Li. LLM anonymization against agentic re-identification, 2026. URL https://arxiv.org/abs/2605.30848.

Kyungho Lim and Byung-Hoon Kim. Anonpsy: A graph-based framework for structure-preserving de-identification of psychiatric narratives. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 19288–19311, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-395-1. doi: 10.18653/v1/2026.findings-acl.963. URL https://aclanthology.org/2026.findings-acl.963/.

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, 2024a. doi: 10.1162/tacl_a_00638. URL https://aclanthology.org/2024.tacl-1.9/.

Xiang Liu, Peijie Dong, Xuming Hu, and Xiaowen Chu. LongGenBench: Long-context generation benchmark. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 865–883, Miami, Florida, USA, November 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.48. URL https://aclanthology.org/2024.findings-emnlp.48/.

Matthieu Meeus, Lukas Wutschitz, Santiago Zanella-Béguelin, Shruti Tople, and Reza Shokri. The canary's echo: Auditing privacy risks of LLM-generated synthetic text. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 43557–43580, 2025. URL https://openreview.net/forum?id=f3mQ0xYA1I.

Meta. Llama-3.1-8B-Instruct. Hugging Face model card, 2024. URL https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct.

Microsoft. Phi-4. Hugging Face model card, 2024. URL https://huggingface.co/microsoft/phi-4.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. The text anonymization benchmark (TAB): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101, December 2022. doi: 10.1162/coli_a_00458. URL https://aclanthology.org/2022.cl-4.19/.

Qwen Team. Qwen3-4B-Instruct-2507. Hugging Face model card, 2025a. URL https://huggingface.co/Qwen/Qwen3-4B-Instruct-2507.

Qwen Team. Qwen3-14B. Hugging Face model card, 2025b. URL https://huggingface.co/Qwen/Qwen3-14B.

Theresa Stadler, Bristena Oprisanu, and Carmela Troncoso. Synthetic data – anonymisation Groundhog Day. In *31st USENIX Security Symposium (USENIX Security 22)*, pp. 1451–1468, 2022. URL https://www.usenix.org/conference/usenixsecurity22/presentation/stadler.

Latanya Sweeney. k-anonymity: a model for protecting privacy. *Int. J. Uncertain. Fuzziness Knowl.-Based Syst.*, 10(5):557–570, October 2002. ISSN 0218-4885. doi: 10.1142/S0218488502001648. URL https://doi.org/10.1142/S0218488502001648.

Robert Tarjan. Depth-first search and linear graph algorithms. *SIAM Journal on Computing*, 1(2):146–160, 1972. doi: 10.1137/0201010.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-Pack: Packed resources for general Chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, pp. 641–649, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400704314. doi: 10.1145/3626772.3657878. URL https://doi.org/10.1145/3626772.3657878.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Kevin Yang, Yuandong Tian, Nanyun Peng, and Dan Klein. Re3: Generating longer stories with recursive reprompting and revision. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 4393–4479, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.296. URL https://aclanthology.org/2022.emnlp-main.296/.

Kevin Yang, Dan Klein, Nanyun Peng, and Yuandong Tian. DOC: Improving long story coherence with detailed outline control. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3378–3465, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.190. URL https://aclanthology.org/2023.acl-long.190/.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

A STANDARD DEVIATIONS

The figures in the main text plot means only. This section reports the spread behind them. Table 1 gives the standard deviation of the two metrics that vary across the synthetic length sweep. The section-hierarchy score has zero standard deviation at every length, since it is a clean 1.0 for the bounded arms and 0.0 for the one-pass arm. The one-pass cross-reference standard deviation is one to two orders of magnitude larger than that of BRIEF at every length.

Table 1: Standard deviation over 8 to 10 documents per length for the synthetic sweep. BRIEF is stable while the one-pass baseline is high variance.

	xref F1 std			distinct-2 std		
clauses	whole	naive	BRIEF	whole	naive	BRIEF
70	0.124	0.000	0.010	0.043	0.013	0.008
150	0.183	0.008	0.010	0.071	0.010	0.005
225	0.224	0.020	0.009	0.049	0.006	0.003
300	0.225	0.014	0.005	0.045	0.004	0.004
450	0.297	0.006	0.006	0.052	0.004	0.004
600	0.284	0.006	0.005	0.040	0.003	0.003
900	0.258	0.010	0.005	0.030	0.002	0.003
1200	0.194	0.007	0.003	0.017	0.002	0.002

Table 2 gives the real-corpus cross-reference fidelity as a mean with its standard deviation over three seeds. The mean gaps between one-pass and BRIEF are substantially larger than the reported seed-to-seed spread; the document-level paired tests are in Appendix B.

Table 2: Real-corpus cross-reference fidelity, mean and standard deviation over three seeds. The reported mean $\pm$ standard-deviation ranges are separated, although the experiment uses only three generation seeds per document; the deviations quantify seed variability and do not replace document-level uncertainty estimates.

corpus	whole (one-pass)	BRIEF (ours)
CUAD	0.687 ± 0.017	0.921 ± 0.008
EDGAR	0.631 ± 0.009	0.879 ± 0.008

B SIGNIFICANCE OF THE FIDELITY COMPARISONS

We test the two structural claims the paper rests on, BRIEF against one-pass and BRIEF against the naive bounded baseline, on cross-reference F1. For each corpus and generator we average a document's score over the three seeds to get one paired value per document, then report the mean paired difference Δ (BRIEF minus the comparison arm), a bootstrap 95% confidence interval over documents (10^4 resamples), and a two-sided Wilcoxon signed-rank p. A document is scored only when its target graph contains an internal cross-reference, since cross-reference F1 is undefined otherwise (Appendix I); this rule reads only the target graph, so it drops the same documents for every arm, and we pair on the documents common to both generators, giving one n per corpus. Cross-reference F1 is the exact extractor metric of Appendix I. We apply the same paired framework to the privacy and coherence claims (Tables 4 and 5): for privacy we test the binary BGE top-1 re-identification of each release instance with a two-sided McNemar test, reporting the paired difference in re-identification rate; for coherence we test the per-document within-release coherence with the same signed-rank and bootstrap procedure as the fidelity comparison.

The BRIEF-over-one-pass gap is large and significant on every corpus and both generators. The BRIEF-over-naive comparison is the opposite: small and generator-dependent, statistically indistinguishable on Qwen EDGAR ($p = 0.45$) and significantly favoring the naive arm on Llama and on Qwen CUAD, which is the quantitative basis for the claim that the schedule's advantage over the naive bounded baseline is not uniform across generators.

Table 3: Paired significance of the cross-reference F1 comparisons, seed-averaged per document. Δ is BRIEF minus the comparison arm, with a bootstrap 95% CI over documents and a two-sided Wilcoxon signed-rank p. Positive Δ favors BRIEF.

			BRIEF vs one-pass		BRIEF vs naive	
generator	corpus	n	Δ (95% CI)	p	Δ (95% CI)	p
Qwen3-4B	CUAD	60	$+0.23\,[+0.19, +0.28]$	$<10^{-6}$	$-0.03\,[-0.06, +0.00]$	2×10^{-3}
Qwen3-4B	EDGAR	45	$+0.25\,[+0.18, +0.32]$	$<10^{-6}$	$+0.01\,[-0.03, +0.04]$	0.45
Llama-3.1-8B	CUAD	60	$+0.48\,[+0.41, +0.54]$	$<10^{-6}$	$-0.10\,[-0.13, -0.07]$	$<10^{-6}$
Llama-3.1-8B	EDGAR	45	$+0.28\,[+0.21, +0.34]$	$<10^{-6}$	$-0.09\,[-0.13, -0.05]$	8×10^{-6}

Privacy. Table 4 tests source linkage under the headline BGE attacker, pairing arms on the same release instances. BRIEF is significantly less linkable than one-pass on every corpus and generator, and significantly below the naive per-clause arm as well, so the diversity condition of the enriched schedule lowers linkage beyond the bounded schedule alone.

Table 4: Paired significance of BGE top-1 re-identification, per release instance. Δrate is the comparison arm's re-identification rate minus BRIEF's, with a bootstrap 95% CI and a two-sided McNemar p. Positive Δrate means BRIEF is more private.

			one-pass vs BRIEF		naive vs BRIEF	
generator	corpus	n	Δrate (95% CI)	p	Δrate (95% CI)	p
Qwen3-4B	CUAD	207	$+0.33\,[+0.26, +0.40]$	$<10^{-6}$	$+0.14\,[+0.09, +0.19]$	1×10^{-6}
Qwen3-4B	EDGAR	141	$+0.26\,[+0.18, +0.35]$	$<10^{-6}$	$+0.13\,[+0.06, +0.19]$	3×10^{-4}
Llama-3.1-8B	CUAD	210	$+0.40\,[+0.33, +0.47]$	$<10^{-6}$	$+0.11\,[+0.04, +0.17]$	3×10^{-3}
Llama-3.1-8B	EDGAR	141	$+0.26\,[+0.18, +0.34]$	$<10^{-6}$	$+0.08\,[+0.03, +0.14]$	7×10^{-3}

Coherence. Table 5 tests the two coherence claims, pairing per document over all three seeds. Coherence is defined only where a release has a resolved internal cross-reference to judge, so unlike the target-only fidelity rule this drop depends on the generated text, and a document enters a comparison only when both arms produce at least one judged pair; n is that paired count. Within-release coherence is significantly lower for BRIEF than for one-pass, confirming the semantic gap of Section 6.3, and the BRIEF-L plug-in significantly restores it over the summary-free BRIEF arm on every corpus and generator.

Table 5: Paired significance of within-release coherence, pooled over all three seeds per document, two-sided Wilcoxon with a bootstrap 95% CI. The main comparison is one-pass minus BRIEF (positive means one-pass is more coherent); the plug-in comparison is BRIEF-L minus BRIEF (positive means the summaries restore coherence).

comparison	generator	corpus	n	Δ (95% CI), p
one-pass − BRIEF	Qwen3-4B	CUAD	56	$+0.27\,[+0.20, +0.33]$, $<10^{-6}$
one-pass − BRIEF	Qwen3-4B	EDGAR	35	$+0.26\,[+0.20, +0.32]$, $<10^{-6}$
BRIEF-L − BRIEF	Qwen3-4B	CUAD	60	$+0.54\,[+0.49, +0.59]$, $<10^{-6}$
BRIEF-L − BRIEF	Qwen3-4B	EDGAR	35	$+0.57\,[+0.51, +0.63]$, $<10^{-6}$
BRIEF-L − BRIEF	Llama-3.1-8B	CUAD	64	$+0.31\,[+0.26, +0.37]$, $<10^{-6}$
BRIEF-L − BRIEF	Llama-3.1-8B	EDGAR	35	$+0.39\,[+0.32, +0.47]$, $<10^{-6}$

C FULL SYNTHETIC SWEEP

Table 6 gives the per-length means behind Figure 2. The section-hierarchy score is omitted because it is constant, at 0.0 for the one-pass arm and 1.0 for both bounded arms at every length. One-pass cross-reference fidelity is erratic in the range 0.34 to 0.69. BRIEF cross-reference fidelity stays near

0.98. BRIEF keeps the highest diversity at every length. The one-pass per-call context grows by more than an order of magnitude while both bounded arms stay flat.

Table 6: Per-length means of the synthetic sweep, reported as whole / naive / BRIEF for each metric. Section-hierarchy F1 is constant and omitted.

	xref F1			distinct-2			max per-call context		
clauses	whole	naive	BRIEF	whole	naive	BRIEF	whole	naive	BRIEF
70	0.69	1.00	0.99	0.51	0.17	0.56	1215	62	85
150	0.63	0.99	0.99	0.36	0.16	0.43	2611	63	85
225	0.52	0.98	0.98	0.31	0.15	0.36	3876	63	85
300	0.34	0.98	0.98	0.30	0.14	0.32	5165	63	85
450	0.61	0.99	0.98	0.21	0.12	0.27	7790	63	86
600	0.57	0.99	0.98	0.13	0.10	0.23	10528	65	86
900	0.65	0.99	0.98	0.10	0.07	0.19	16160	65	86
1200	0.45	0.99	0.98	0.07	0.06	0.17	21761	65	87

A format-matched one-pass baseline. To check that the one-pass section-hierarchy failure is not merely an unstated convention, we add *whole-matched*, a one-pass arm whose prompt states the two-level numbering format explicitly and matches the reference layout. Table 7 reports it. Even with the format spelled out, whole-matched scores section-tree F1 of 0.0 at every length, exactly like the plain one-pass arm (Table 6), and its cross-reference F1 stays high variance and no more reliable. So the hierarchy collapse is a property of one-pass generation at length, not of the prompt wording.

Table 7: Format-matched one-pass baseline. Even with the two-level numbering stated explicitly, whole-matched never reproduces the section hierarchy. Plain one-pass values are in Table 6.

clauses	section-tree F1	cross-reference F1
70	0.0	0.20
150	0.0	0.30
300	0.0	0.56
600	0.0	0.54

D Positioning against Anonpsy

Table 8 summarizes how this work generalizes Anonpsy along each axis.

Table 8: How BRIEF generalizes Anonpsy from short clinical narratives to long, cross-referential, layout-bearing documents.

dimension	Anonpsy	this work
documents	short clinical narratives	long contracts and filings
length handling	one-shot graph-conditioned	bounded-context schedule
structure	clinical entity graph	cross-reference and term graph, verified
modality	text only	text and layout
privacy audit	similarity and judge proxy	explicit linkage adversary
perturbation	graph-constrained	capacity-bounded diagnostic

E Scope and experimental arms

Table 9 states what the paper evaluates and what it does not, so the claims are not read more broadly than the evidence. Table 10 lists every arm. The headline BRIEF results are the enriched bounded schedule *without* the perturbation operators. The perturbation arm differs from BRIEF only on the privacy axis; the capacity diagnostic leaves little room for structural perturbation on either real corpus (mean capacity 0.27 on CUAD and 0.094 on EDGAR), so we report BRIEF without it.

Table 9: What BRIEF is shown to preserve, and what it is not.

property	status
number paths and section hierarchy	evaluated, preserved
internal cross-reference topology	evaluated, preserved
referential resolution	evaluated, preserved
source 13-gram overlap	evaluated; approximately zero for generated arms
cross-reference semantics	evaluated, not preserved (bounded schedule)
clause role relative to the source	regenerated, not matched (by design)
defined-term identity and usage	regenerated by design, not preserved
empirical source linkage	evaluated with a limited set of attackers
legal or financial meaning	not established
formal privacy	not provided

Table 10: Experimental arms. Headline BRIEF excludes the perturbation arm.

arm	description
copy	the unmodified source document; source ceiling for raw-source fidelity, with no privacy transformation
whole	one-pass full-document generation
whole-matched	one-pass with the two-level numbering stated explicitly
naive	bounded per-clause generation from a generic prompt
BRIEF (ours)	bounded per-clause generation from a structural brief
BRIEF + perturbation	BRIEF plus capacity-bounded perturbation (privacy axis only)
entity-masking	NER redaction of the source in place

Entity-masking baseline. The entity-masking arm runs a named-entity recognizer over the source and replaces each detected entity span with a type placeholder, leaving every other token and the document structure unchanged. It masks about seven percent of the text on our corpora, so the numbering, the cross-references, and most of the wording survive verbatim, which is why it stays highly linkable under the BGE attacker at 0.986 on CUAD and 0.81 on EDGAR. It is a defense arm, not an adversary, and it is the standard point of comparison for regeneration.

F THE CONTENT GRAPH AND ITS STRUCTURAL VERIFIER

Every document, real or synthetic, is parsed into one intermediate representation before anything else runs. The representation holds three things. It holds an ordered list of nodes, where each node is a clause or a table and carries its text and its number path such as `2.1`. It holds every cross-reference found in the text, each tagged with whether its target exists in the same document. It holds every defined term, with the place it is first defined and the places it is later used. All four structural families in the paper are read off this representation.

A release is accepted only when three checks pass. The first check is a leak check. The generator writes a map of every real value it replaced. The check searches for exact source 13-grams among spans designated for replacement, with canonical structural literals intentionally retained by the target graph excluded. The second check is a structural-consistency check. It pulls plain structural facts from the source and runs the same facts on the release. Cross-references must resolve to a clause that exists. Clause numbers must increase within a section. A defined term must be used somewhere after it is defined. Working structure must not be deleted wholesale. The third check is a coherence check. It never sends the whole document to a model in one prompt. Instead it extracts a short, specific list first and asks a model to judge only that list. It flags undefined terms, numbers that contradict each other, two scenarios stitched into one document, and text that breaks off mid sentence. Every flag must cite the exact sentence it rests on. This design keeps each prompt small while covering the whole document, which is what makes checking a document of sixty pages or more possible without truncation; the separate LLM judge and retrieval attacker of Section 5 read only the first 20,000 characters.

G EXTRACTION PIPELINE

Real filings do not number their clauses the way our graph expects. Two preprocessing steps close this gap. The heading normalizer rewrites the common heading styles into one arabic number path. It maps `Section N.N`, `SECTION N.N`, and `Sec. N.N` to `N.N`. It maps `Article` written in roman numerals to arabic. It maps `Item N` to a numbered path. Numeric headings are left untouched. Body cross-references are left intact, since the parser already resolves them. The enumeration splitter turns inline lists such as `(a)` and `(b)`, and roman lists such as `(i)` and `(ii)`, into numbered sub-clauses one level deeper. Canonicalization then merges any nodes that share a number path, which removes parser over-segmentation, and it keeps the original numbering so that source references stay valid. Without these steps the extractor, not the document, was the limit on how many clauses a long filing appeared to contain.

H THE BRIEF MECHANISM

Generation walks the target graph one node at a time. For each node the system assembles a structural brief and places it in a single prompt. The brief states the clause number path. It states the term the clause must define, if the node is a definer. It states the term the clause must use by name, if the node is a user. It states the exact cross-reference string the clause must contain, if the node cites another clause. Three prompt families use this brief.

The *bounded* prompt asks for one or two sentences of a services-contract clause body and injects the brief. The *enriched* prompt is the one we call BRIEF. It adds a distinct topic drawn from a pool of 21 topics indexed by clause position, one of five opening steers, a shared register specification so that clauses read as one document, and an instruction not to open with boilerplate such as "The parties shall". The *whole* prompt is the one-pass baseline. It asks for a complete document that contains exactly the listed numbered clauses, in order, one short paragraph each, with the per-clause instructions placed in brackets next to each number.

Each bounded call returns at most 100 tokens. The number path is prepended afterward, so the model never has to reproduce it. The prompt never contains source prose and never contains any other clause. Its length depends only on the brief, which is why the per-call context is flat at about 86 tokens across all document lengths.

I STRUCTURAL-FIDELITY METRIC

For each family we build a label-keyed edge set on the candidate and on the target, then score the set F1 as $2\,|C \cap T|/(|C| + |T|)$. The section-tree family keys a containment edge by the pair of parent number path and child number path. The cross-reference family keys an edge by the pair of citing path and cited path. The defined-term family keys an edge by the pair of term string and using path. The arithmetic family keys operand and result triples. Nodes are aligned by number path, or by term string for definitions, so two documents can match even when their internal identifiers differ.

The convention is reference-relative. When the target family is empty, the family is not applicable and is dropped from the average. When the target family is not empty but the candidate family is empty, the family scores zero. Candidate-only edges that have no counterpart in the target are counted separately as invented edges and are not folded into the F1. The scorer defines four families for an optional weighted aggregate, but our corpora exercise only three. No target contains arithmetic edges, so the arithmetic family is always not applicable here and is dropped. We therefore report per-family F1 rather than an aggregate. The defined-term family is low for every generated arm by design, since the arms regenerate terms rather than reuse the source strings, so the isomorphism claim rests on the term-agnostic families of section tree and cross-reference.

J CAPACITY AND THE STRUCTURAL PRIVACY DIAGNOSTIC

A top-level section is locked when it is the target of a cross-reference or when it defines a term that is used later. Otherwise it is free. The capacity is the fraction of free sections. The structural signature is the tuple of the section count, the locked count, the defined-term count, and the internal

cross-reference count. Two perturbation operators never break an invariant. The first typesets the document to a shared house template, which is always available because it touches no content. The second merges the trailing run of free sections, which renumbers no locked section. The reachable structural change is bounded by the length of the trailing free run, which is at most the capacity times the section count. The residual identity after perturbation is the locked skeleton, so the operator-restricted structural collision baseline is one over the signature bucket size (at most the corpus size, so the corpus baseline is never the binding term). This is a nominal structural-collision probability, not a lower bound on an attacker's accuracy. A rigid document has low capacity and a rare signature, so its value is high; a flexible document can be moved into a common bucket, so its value approaches $1/N$. The diagnostic is computed before generation, so a caller knows in advance which documents are rigid.

Table 11 gives the capacity distribution on the real corpora, which is the input the diagnostic reads. CUAD contracts span the full range and are only partly rigid. EDGAR EX-10 exhibits are mostly rigid, with only about four percent flexible, which leaves the perturbation operators little to act on there. We compute the per-document signature offline from the saved sources, which reproduces these capacity means and so validates the extraction. The signature bucket size is degenerate. On CUAD all 69 documents have a unique signature, so $k_\sigma = 1$ and $L_{\mathrm{struct}} = 1$ for every document. On EDGAR 36 of 47 are unique and the rest fall into four pairs and one triple, so L_{struct} averages 0.87. The raw four-integer signature is therefore too fine for useful k-anonymity at these corpus sizes. The actionable signal on real data is the capacity, not L_{struct}, and a coarser or learned signature is the right next step for real deployments.

Table 11: Perturbation-capacity distribution on the real corpora. Rigid is capacity below 0.2, flexible is capacity above 0.5.

corpus	mean	max	rigid fraction	flexible fraction
CUAD	0.27	0.79	0.42	0.20
EDGAR	0.094	0.57	0.68	0.04

K PRIVACY ADVERSARIES

The weak text attacker is a TF-IDF nearest neighbor. The BGE retrieval-based linkage attacker embeds each document with a retrieval embedder. It splits the text into chunks of 350 words, keeps at most 12 chunks, embeds each chunk, mean-pools the chunk vectors, and normalizes the result. It ranks sources by cosine similarity to the release. For the layout channel, each document, source or release, is first rendered to PDF with the Typst engine under a fixed page geometry, then parsed with `pdfplumber` into per-page block, figure, and table structure; the same renderer and parser are applied to real and regenerated documents so the channel measures re-typesetting rather than the original filing's formatting. The layout attacker reduces this parsed layout to an eight-dimensional signature. The signature records the page count, the section count, the figure count, the table-break count, the pages per section, the blocks per page, the figure page spread, and the count of figure placement violations. It ranks sources by z-normalized Euclidean distance on this vector. The joint attacker min-max fuses the text score and the layout score with equal weight. In every case we report top-1 accuracy against the chance baseline of one over the corpus size, and lower means more private.

Table 12 gives every arm on every channel. Three points support the claims in the main text. The weak TF-IDF attacker underestimates leakage at every generated arm except BRIEF on EDGAR, where the two attackers roughly agree (0.24 against 0.23), so it is not a reliable upper estimate and the BGE attacker is the headline. Layout-only sits near the chance baseline for every generated arm. Fusing layout with BGE lowers linkage for every generated arm where it was run on both corpora; on EDGAR, for example, BRIEF's linkage falls from 0.23 to 0.13 and one-pass from 0.49 to 0.09. Layout is therefore not an extra leak for regenerated documents.

The entity-masking baseline uses a spaCy named-entity tagger and redacts the PERSON, ORG, GPE, DATE, and MONEY types. It masks only about seven percent of the text and preserves structure at a section-tree fidelity of 0.998. This is why it stays highly linkable. On the synthetic corpus the picture is different. With 18 documents and a floor of 0.056, the BGE top-1 is 0.22 for whole, 0.00

Table 12: Top-1 re-identification across adversaries and channels, lower is more private. Floors are 0.014 (CUAD) and 0.021 (EDGAR). "n/a" means the channel was not run for that arm.

	CUAD				EDGAR			
arm	BGE	TF-IDF	layout	joint	BGE	TF-IDF	layout	joint
copy	1.00	1.00	0.84	1.00	0.96	0.96	0.62	0.96
entity-masking	0.99	n/a	n/a	n/a	0.81	n/a	n/a	n/a
whole	0.70	0.53	0.04	0.54	0.49	0.42	0.04	0.09
naive	0.51	0.18	0.09	0.41	0.36	0.21	0.04	n/a
BRIEF (ours)	**0.37**	0.20	0.09	**0.33**	**0.23**	0.24	0.09	0.13

for naive, and 0.056 for BRIEF, so the bounded arms sit at or near the floor. This is an artifact of the homogeneous templated corpus, where every document shares the same body wording and there is little document-specific signal to link. The real corpora are therefore the meaningful privacy test, and we report the synthetic privacy only for completeness.

L CORRECTNESS AND DEFINED TERMS

Table 13 gives referential integrity per arm, measured as the fraction of internal cross-references that resolve to a clause that exists. The one-pass arm keeps the highest internal consistency because it drafts in one pass, but it does so on a different structure that scores low against the target graph. BRIEF integrity is higher at longer lengths in this experiment; the mechanism behind this pattern is not isolated. The defined-term-usage F1 is low for every generated arm, at about 0.23 to 0.27 on CUAD and about 0.22 on EDGAR, against 0.71 to 0.83 for copy. This reduces source-term reuse in the tested representation, because the generated arms invent their own terms rather than reuse the source strings that the family keys on.

Table 13: Referential integrity, reported as the fraction of internal cross-references that resolve.

corpus	copy	whole	naive	BRIEF
CUAD	0.68	0.965	0.80	0.83
EDGAR	0.50	0.99	0.87	0.88

Table 14 reports invented cross-reference edges, meaning edges in the candidate that have no counterpart in the target graph. On the long EDGAR EX-10 exhibits the one-pass arm invents 12.2 cross-reference edges per document, against 1.6 for BRIEF, so its low fidelity is over-generation of references that do not match the target rather than omission. No arm invents section-tree edges. The defined-term family is omitted here because generated arms invent terms by design.

Table 14: Invented cross-reference edges per document, meaning candidate edges with no counterpart in the target. Lower is more disciplined.

corpus	copy	whole	naive	BRIEF
CUAD	2.1	4.3	0.7	0.7
EDGAR	2.3	12.2	2.1	1.6

Fidelity to the source graph, and source-span overlap. The main text scores each arm against the target graph G^*, which is the sanitized and canonicalized source graph. Because G^* keys only internal, canonicalized references, a release can score above the source round-trip on it, as Figure 3 shows. Table 15 instead scores each release graph $G(R)$ against the *raw* source graph $G(D)$. Here copy is 1.0 by construction, which restores a proper upper bound, and BRIEF sits below it while remaining the strongest generated arm on section structure and on EDGAR cross-references; on CUAD cross-references the naive arm is level (0.68 against 0.67). So the earlier above-copy scores come from target construction, not from BRIEF preserving more source structure. The same table reports source-span overlap, the fraction of source word 13-grams that reappear verbatim in the

release. It is 1.0 for copy and effectively zero for every generated arm; shorter copied spans can still escape a 13-gram test.

Table 15: Fidelity of each release to the raw source graph $G(D)$ (section-tree / cross-reference F1), and source-span overlap. Copy is the true upper bound here.

corpus	metric	copy	whole	naive	BRIEF
CUAD	section-tree F1	1.00	0.70	0.73	0.75
CUAD	cross-reference F1	1.00	0.46	0.68	0.67
CUAD	source-span overlap	1.00	0.000	0.000	0.000
EDGAR	section-tree F1	1.00	0.53	0.57	0.61
EDGAR	cross-reference F1	1.00	0.31	0.45	0.49
EDGAR	source-span overlap	1.00	0.000	0.000	0.000

M SYNTHETIC CORPUS GENERATOR

The synthetic testbed is deterministic and seeded. Given a target section count and a fixed count of four subsections, it emits a two-level hierarchy of numbered clauses. It threads three dependency families through them. It introduces defined terms from a pool of distinctive multi-word phrases, each stated in quotes at an early subsection. It has a later subsection reuse an existing term by name with probability 0.7. It has a later subsection cite an earlier one with probability 0.6. A reference document renders each clause with wording chosen to hit the extractor conventions, so a quoted term becomes a definition, a reuse becomes a term-use edge, and a "Clause a.b" string becomes a cross-reference. The ground truth is the graph the extractor produces on that reference document, so a candidate that reproduces the reference exactly scores 1.0 on every family by construction. We vary the section count to reach 70 to 1200 clauses and vary the seed to change the term and cross-reference assignment.

N MODELS, COMPUTE, AND HYPERPARAMETERS

The main generator is `Qwen/Qwen3-4B-Instruct-2507`; the cross-model study uses `meta-llama/Llama-3.1-8B-Instruct`. The judge is `Qwen/Qwen3-14B` with thinking disabled and temperature zero, with `microsoft/phi-4` at temperature zero as the second judge in the cross-model study. The linkage attacker is `BAAI/bge-base-en-v1.5`. Model licenses: Qwen3 under Apache-2.0, Llama-3.1 under the Llama 3.1 Community License, Phi-4 under MIT, and BGE-base-en-v1.5 under MIT. Generation temperature is 0.7. We use three seeds, with eight to ten documents per length in the sweep. The joint fusion weight is 0.5. Text is truncated to 120k characters for extraction and to 20k characters for the adversary and the judge. All runs use the model chat template. Experiments ran on L4 and A100-80GB GPUs. The full project used more compute than the reported runs, since early pilots did not make the paper.

O EFFECTIVE CONTEXT AND THE ONE-PASS BUDGET

The one-pass baseline must hold the whole document in a single context window. We set its context budget from each generator's effective context, measured with RULER (Hsieh et al., 2024), using the common-word extraction (cwe) and frequent-word extraction (fwe) aggregation tasks. We use the NVIDIA/RULER implementation; the task configuration, number of examples, and random seed are given with the released code. Table 16 reports accuracy by context length. Qwen3-4B remains above 82 accuracy on the common-word extraction task through 128K, so we use 128K as its one-pass budget. On the common-word extraction task, Llama-3.1-8B degrades sharply by 32K, falling from 75 at 16K to 15 at 32K; the frequent-word task shows a different, non-monotonic pattern. We use 32K as its one-pass budget. BRIEF does not use this budget. Its per-call context stays near 86 tokens at every document length, so the budget bounds the one-pass arm alone.

Table 16: RULER accuracy by context length for the two generators. cwe is common-word extraction and fwe is frequent-word extraction. On the two reported RULER aggregation tasks, Qwen remains relatively stable through 128K, whereas Llama degrades sharply by 32K. These measurements motivate the one-pass budgets but do not by themselves establish the cause of the structural failures.

model	task	16K	32K	64K	128K
Qwen3-4B	cwe	96.8	92.4	89.5	82.1
Qwen3-4B	fwe	99.0	96.3	93.3	90.7
Llama-3.1-8B	cwe	75.0	14.7	2.0	–
Llama-3.1-8B	fwe	88.9	93.9	85.1	–

P LINGUISTIC QUALITY

The judge and the retrieval adversary each receive the leading 20,000 characters of a document, the same window for every arm. Table 17 gives the full 14B judge scores for the Qwen generator. Here one-pass generation reads best as prose, which is the counter-signal for this generator. The advantage of BRIEF is structural fidelity and privacy, not prose fluency. A position-randomized pairwise comparison agrees with the rubric scores. The one-pass arm is preferred over BRIEF in about 55 to 83 percent of pairs across the corpora and judges, the low end coming from the 47-document EDGAR set where the preference is near a coin flip. This preference is generator-specific and reverses on Llama, where BRIEF is preferred by both judge families (Appendix Q). The judge reads excerpts rather than whole documents, which under-weights the late-document structural failure that the length sweep in Section 6 catches.

Table 17: Linguistic quality from the 14B judge, scale 1 to 5.

corpus	metric	copy	whole	naive	BRIEF
CUAD	fluency	4.03	**4.40**	4.00	3.97
CUAD	coherence	3.40	**3.83**	2.70	3.00
EDGAR	fluency	3.49	**3.81**	3.35	3.62
EDGAR	coherence	2.68	**2.81**	2.22	2.62
synth	fluency	2.00	**4.56**	4.33	4.00
synth	coherence	1.39	**3.33**	3.06	3.00

Q CROSS-MODEL STUDY ON LLAMA-3.1-8B

We repeat the full study with `meta-llama/Llama-3.1-8B-Instruct` as the generator, holding the sources, the target graphs, and the evaluators fixed. This tests whether the findings persist across generators, but it does not isolate scheduling effects from model-family or model-capacity differences. Table 18 collects the structure, privacy, and coherence numbers. The main bounded-context finding and the low semantic-coherence finding persist: BRIEF preserves section and cross-reference topology at constant prompt length while one-pass degrades with length, and within-release coherence stays low because the bounded generator never sees the clauses it cites. The relative ranking of BRIEF and the naive bounded baseline does not fully replicate. The naive per-clause arm scores higher than BRIEF on cross-reference fidelity with Llama (0.96 against 0.92 on the synthetic sweep, 0.92 against 0.82 on CUAD, 0.91 against 0.82 on EDGAR), while BRIEF has the lower EDGAR linkage (0.18 against 0.26). BRIEF's advantage over the naive arm with this generator is therefore linkage and lexical diversity, not topology.

Prose preference also differs across generators. Table 19 gives the position-randomized pairwise win rate of BRIEF over one-pass under both judge families, and Table 20 the rubric scores. On Qwen output one-pass is rated more fluent by both judges. On Llama output the gap closes and the pairwise test prefers BRIEF. The preference varies primarily with the generator and is consistent with the effective-context measurements in Appendix O, which do not by themselves establish a causal explanation.

Table 18: Cross-model results with the Llama-3.1-8B generator, three seeds. Section tree and cross-reference fidelity are against the sanitized target graph. Linkage is BGE top-1, lower is less linkage. Coherence is within-release cross-reference coherence. Bold marks the best generated arm in each row. Copy is the source ceiling for raw-source fidelity, but it is not necessarily an upper bound against the sanitized target graph because the target graph may differ from the source graph.

corpus	metric	copy	whole	naive	BRIEF
synth (1200)	section_tree	1.00	0.01	**1.00**	**1.00**
synth (1200)	xref	1.00	0.19	**0.96**	0.92
CUAD	xref	0.88	0.35	**0.92**	0.82
EDGAR	xref	0.92	0.55	**0.91**	0.82
CUAD	BGE linkage	1.00	0.75	0.45	**0.35**
EDGAR	BGE linkage	0.96	0.44	0.26	**0.18**
CUAD	coherence	0.58	0.21	0.10	0.05
EDGAR	coherence	0.31	0.18	0.13	0.05

Table 19: Pairwise win rate of BRIEF over one-pass, position-randomized, under two judge models; per-cell pair counts are released with the code. Above 0.5 means BRIEF is preferred. The preference varies with the generator rather than the judge.

generator	corpus	Qwen3-14B judge	Phi-4 judge
Qwen3-4B	CUAD	0.17	0.28
Qwen3-4B	EDGAR	0.45	0.42
Llama-3.1-8B	CUAD	0.67	0.65
Llama-3.1-8B	EDGAR	0.85	0.70

R SOURCE FAITHFULNESS AND INTERNAL COHERENCE

We separate faithfulness to the source from internal usability. The first probe tests whether a regenerated clause plays the same role as the source clause it replaces. For clauses aligned by number path, the 14B model labels the source clause and the release clause into a nine-way contract-role taxonomy, and we report the fraction that agree (Table 21). BRIEF agreement falls below the majority-class floor. The role probe confirms that BRIEF is not source-faithful, while the within-release coherence probe shows that topology does not imply reference semantics. The BRIEF-L plug-in (Appendix S) adds bounded summaries of referenced clauses to improve this trade-off. Internal usability is the separate question. We measure it by referential integrity (Table 13), by the judge coherence score (Table 17), and by a within-release cross-reference coherence probe (Table 22) that asks the 14B judge whether each cited clause is a coherent target for the clause that cites it.

Table 22 reports the within-release cross-reference coherence. For each internal cross-reference that resolves, the 14B judge decides whether the cited clause is a coherent target for the clause that cites it. Copy is the source document itself and serves as the reference point. BRIEF scores lowest of all arms. Its references resolve structurally but rarely point to a semantically related clause. This is a generation-context effect. The bounded generator writes each clause without seeing the clauses it references, so it cannot make the reference cohere at any length. One-pass drafts the whole document at once and does better on this axis, though it too drops on the long EDGAR EX-10 exhibits. The source's own lower EDGAR value reflects references to material outside the extracted clause set, so it is a noisy reference point there, not a ceiling. Conditioning each brief on a short summary of the clauses it references is the fix that BRIEF-L implements (Appendix S), and Appendix T shows side-by-side examples. This table scores a random subsample of documents (20 CUAD, 37 EDGAR); the paired tests of Table 5 rerun the probe over every document and all three seeds and agree in direction. The copy arm's coherence is not comparable across tables, because each table judges a different document set or citation setup: this table a random 20/37-document subsample, Table 23 the k-citation setup of Appendix S. Every coherence comparison in this paper is made within one table.

Table 20: Fluency and within-document coherence rubric (1 to 5) by generator and judge, mean over the two real corpora. The one-pass fluency lead on Qwen output disappears on Llama output under both judges.

generator	judge	whole fl	BRIEF fl	whole coh	BRIEF coh
Qwen3-4B	Qwen3-14B	4.11	3.79	3.33	2.81
Qwen3-4B	Phi-4	3.73	3.39	2.92	2.39
Llama-3.1-8B	Qwen3-14B	3.24	3.33	2.20	2.33
Llama-3.1-8B	Phi-4	3.03	3.21	1.92	2.21

Table 21: Clause-role agreement between each release clause and its aligned source clause, labeled by the 14B model. The floor is the majority-class rate. BRIEF does not preserve semantic role.

corpus	whole	naive	BRIEF	majority-class floor
CUAD	0.29	0.32	0.10	0.39
EDGAR	0.32	0.27	0.06	0.30

S BRIEF-L: DEPENDENCY-ORDERED SEMANTIC PLUG-IN

BRIEF-L (Section 4.5) writes clauses in a dependency order and gives each prompt bounded summaries of the clauses it cites. This appendix reports the schedule statistics, the full coherence matrix under both judges, and the privacy guardrail. The schedule module is deterministic and unit-tested, and each release carries a generation-graph artifact recording the cited-clause summaries, the deleted edges, and the layer assignment. The right lane of stage 3 in Figure 1 shows the pipeline.

Relation to the headline BRIEF. The BRIEF-L setup carries up to k cross-references per clause, while the headline BRIEF study of Section 6 carries the single primary cross-reference produced by canonicalization. Emitting more exact cross-reference strings makes a release more source-identifiable, so the summary-free BRIEF arm in this section has higher absolute BGE linkage than the headline BRIEF (about 0.89 against 0.37 on CUAD) and slightly higher within-release coherence (about 0.09 against 0.02). Every BRIEF-L arm shares this k-citation setup, so the ablation isolates the effect of the summaries and the topics, and the numbers in this section are compared within these tables rather than against the headline study.

Schedule construction. Algorithm 2 builds the schedule. The dependency graph H has one edge from a citing clause to each clause it cites, so a valid generation order writes every cited clause before the clause that cites it. We find cycles with Tarjan's strongly connected component algorithm (Tarjan, 1972). A component of size one is already acyclic. For a component of size greater than one we delete edges until it is acyclic, and we choose each deleted edge by three tie-breakers in order: prefer a forward reference whose citing path sorts after the cited path, then the edge whose removal leaves the fewest edges on any cycle, then the lexicographically smallest citing and cited path pair. This is a deterministic feedback-arc heuristic, not a minimum feedback-arc set, which is NP-hard. Every deleted edge is logged, and the clause on a deleted edge writes that one reference without a summary. We then layer the acyclic graph with Kahn's algorithm (Kahn, 1962) in the reverse-topological form of Algorithm 3, so layer zero holds the clauses that cite nothing and each later layer holds clauses all of whose cited clauses are already scheduled. Layers are parallel generation batches, and the number of layers is the scheduling depth reported below.

Generation. Algorithm 1 in the main text generates in schedule order. Each prompt is the structural brief of the clause plus the stored summaries of up to k clauses it cites, chosen by number path and each clipped to s tokens. When the returned summary is missing or longer than s tokens we fall back to one bounded summarization call over the generated body. A clause that cites nothing, such as every clause in the first layer, carries no summaries and is written from its structural brief alone, so BRIEF-L reduces exactly to BRIEF wherever a clause has no dependencies to honor.

Table 22: Within-release cross-reference coherence: the fraction of resolved internal references whose cited clause the 14B judge accepts as a coherent target. Copy is the source document itself. BRIEF preserves section and cross-reference topology but not reference semantics.

corpus	copy	whole	naive	BRIEF
CUAD	0.85	0.40	0.06	0.02
EDGAR	0.31	0.30	0.19	0.06

Algorithm 2 BRIEF-L schedule construction

Require: target graph G^* with clause nodes V; caps k, s
$H \leftarrow$ graph on V with edge $u \to v$ for each clause v cited by u
optionally add $u \to v$ when u uses a term that v defines
$\mathcal{C} \leftarrow$ TARJAN-SCC(H)
$deleted \leftarrow [\,]$
for each component $S \in \mathcal{C}$ with $|S| > 1$ **do**
  while the subgraph $H[S]$ still contains a cycle **do**
    $E_c \leftarrow$ edges of H that lie inside a nontrivial component
    pick $e = (u, v) \in E_c$ minimizing, in order:
      (a) 0 if $\mathrm{path}(u) > \mathrm{path}(v)$ else 1; (b) remaining cycle edges if e removed; (c) $(\mathrm{path}(u), \mathrm{path}(v))$
    delete e from H; append e to $deleted$
  end while
end for
$layers \leftarrow$ KAHN-LAYERS(H)
return $layers,\ deleted$

Schedule statistics. The dependency graph is built offline from the target graph. On CUAD, 7 of 70 documents contain any cycle, the largest strongly connected component over the whole corpus has 3 clauses, and cycle-breaking deletes 9 of 570 citation edges (1.6 percent); the $k = 4$ cap on cross-references per brief drops about 1.2 more citations per CUAD document and about 2.1 per EDGAR document. On EDGAR 1 of 47 documents contains a cycle, and cycle-breaking deletes 1 of 807 citation edges. The generation schedule is shallow, at mean depth 2.5 layers on CUAD and 2.9 on EDGAR, and the maximum out-degree is 14 on CUAD and 16 on EDGAR, which is why the brief caps cross-references and summaries at $k = 4$.

Coherence, both judges. Table 23 gives the full within-release coherence matrix. The summary arms (BRIEF-L, sum-only, naive-L) lift coherence from near zero to between 0.40 and 0.75 in every cell, while the summary-free arms (BRIEF, edge) stay near the floor. Wherever both judges were run they agree on direction; the Phi-4 judge was not run on Llama-3.1-8B CUAD output, so that cell is absent. The summaries are the active ingredient and edge-consistent topics alone add at most 0.09.

Table 23: Within-release cross-reference coherence for BRIEF-L and ablations, by generator, corpus, and judge. Higher is more coherent. Copy is the source document itself, a reference point rather than a ceiling. Each row scores a random subsample of 20 documents; the paired tests of Table 5 use every document and all three seeds. The Phi-4 judge was not run on Llama-3.1-8B CUAD output.

generator	corpus	judge	copy	BRIEF	edge	BRIEF-L	sum-only	naive-L
Qwen3-4B	CUAD	Qwen3-14B	0.57	0.09	0.09	0.69	0.58	0.70
Qwen3-4B	CUAD	Phi-4	0.40	0.04	0.12	0.49	0.42	0.58
Qwen3-4B	EDGAR	Qwen3-14B	0.41	0.05	0.14	0.70	0.63	0.75
Qwen3-4B	EDGAR	Phi-4	0.36	0.08	0.17	0.59	0.57	0.65
Llama-3.1-8B	CUAD	Qwen3-14B	0.57	0.13	0.10	0.48	0.42	0.53
Llama-3.1-8B	EDGAR	Qwen3-14B	0.41	0.08	0.12	0.55	0.45	0.70
Llama-3.1-8B	EDGAR	Phi-4	0.36	0.09	0.08	0.44	0.40	0.59

Algorithm 3 KAHN-LAYERS: layered reverse-topological order

Require: acyclic H, where edge $u \to v$ means u cites v
$\mathrm{out}[u] \leftarrow$ out-degree of u for every u
$ready \leftarrow \{u : \mathrm{out}[u] = 0\}$; $layers \leftarrow [\,]$
while $ready \neq \emptyset$ **do**
  append $ready$, sorted by number path, to $layers$
  $next \leftarrow \emptyset$
  for all $v \in ready$ and each u citing v **do**
    $\mathrm{out}[u] \leftarrow \mathrm{out}[u] - 1$
    if $\mathrm{out}[u] = 0$ **then** add u to $next$
  end for
  $ready \leftarrow next$
end while
return $layers$

Privacy and structure guardrails. Table 24 gives BGE top-1 linkage. On the better-powered CUAD corpus, 70 documents against a 0.014 chance baseline, BRIEF-L matches BRIEF for both generators, while the generic-prompt naive-L variant leaks more. The coherence gain therefore does not by itself cost privacy, and the edge-consistent topics of the full BRIEF-L keep the regenerated wording diverse enough to hold linkage. On the 47-document EDGAR corpus, against a 0.021 baseline, BRIEF-L raises top-1 linkage over the summary-free BRIEF arm by a small margin (about 0.08 for both generators), and it stays well below the copy arm, the unmodified source at 0.96, and below the generic-prompt naive-L variant.

Table 25 reports the structural and cost guardrails. Section-hierarchy fidelity stays at 1.0 in every cell, cross-reference fidelity stays within 0.03 of BRIEF on Qwen and improves on Llama, and lexical diversity rises over BRIEF. The mean per-document maximum prompt is about 1.9 times the BRIEF length and flat with document length, which is the cost of carrying up to k cited-clause summaries and the fixed summary instruction. On Llama the summaries have one cost: Llama often omits the summary delimiter, so the bounded fallback summarizer runs about 10 to 17 times per document, against under once for Qwen. The fallback stays within the token budget and uses no source prose, so the invariants hold. Citation truncation at $k = 4$ is rare on CUAD (about 1.2 per document) and more frequent on the denser EDGAR filings (about 2.1 per document).

Table 24: BGE top-1 source linkage for BRIEF-L and ablations, lower is more private. CUAD is the reliable corpus. BRIEF-L matches BRIEF there.

generator, corpus	copy	BRIEF	edge	BRIEF-L	sum-only	naive-L
Qwen3-4B, CUAD	1.00	0.89	0.87	0.88	0.91	0.93
Llama-3.1-8B, CUAD	1.00	0.86	0.86	0.86	0.89	0.93
Qwen3-4B, EDGAR	0.96	0.46	0.50	0.55	0.53	0.59
Llama-3.1-8B, EDGAR	0.96	0.44	0.45	0.52	0.48	0.59

Table 25: BRIEF-L structural and cost guardrails, BRIEF baseline shown as "BRIEF → BRIEF-L". Section-tree fidelity is 1.0 throughout. Cross-reference fidelity is against the sanitized target graph, max context is the mean per-document maximum prompt length in tokens, and fallbacks is the mean number of bounded summarization fallbacks per document.

generator, corpus	xref F1	sec-tree	max ctx	distinct-2	fallbacks / doc
Qwen3-4B, CUAD	$0.91 \to 0.88$	1.00	$123 \to 221$	$0.65 \to 0.69$	0.2
Qwen3-4B, EDGAR	$0.92 \to 0.89$	1.00	$137 \to 265$	$0.55 \to 0.58$	0.6
Llama-3.1-8B, CUAD	$0.80 \to 0.83$	1.00	$124 \to 220$	$0.63 \to 0.67$	10.4
Llama-3.1-8B, EDGAR	$0.83 \to 0.86$	1.00	$137 \to 263$	$0.55 \to 0.58$	17.0

T QUALITATIVE EXAMPLES: ONE-PASS, BRIEF, AND BRIEF-L

Table 26 shows the same cross-reference as regenerated by three arms of the Qwen3-4B study, for one CUAD contract and one EDGAR exhibit (seed 0). We picked citations that occur in all three releases, so each row shows how one arm fills the same structural slot: the citing clause and the clause it cites. The examples are chosen by hand to illustrate the aggregate pattern, and they are not judge outputs. Text is reproduced verbatim, apart from the elision marked [. . .]. Upper-case tokens such as PRODUCTS or EVENT are defined terms carried by the target graph. The one-pass text comes from the main study; the BRIEF and BRIEF-L text comes from the BRIEF-L setup (Appendix S).

The pattern follows Tables 22 and 23. BRIEF writes the citing and cited clauses independently, so the reference resolves but links two unrelated obligations: a data-breach duty cites a service-level clause, and a consequential-damages exclusion cites a payment deadline. BRIEF-L writes the cited clause first and gives its one-sentence summary to the citing clause, so both sides of the reference describe the same obligation (force majeure in the first example, allocation of fees and expenses in the second). One-pass is often locally coherent, as in these examples, because it drafts the whole document in one context. Its failure is structural and shows up at length (Section 6), not in single reference pairs.

Table 26: The same internal cross-reference regenerated by one-pass, BRIEF, and BRIEF-L (Qwen3-4B, seed 0). BRIEF preserves the reference but not its meaning; BRIEF-L preserves both.

arm	citing clause	cited clause
CUAD outsourcing agreement, clause 3.7 cites clause 3.8		
one-pass	3.7. The review shall be conducted in accordance with the procedures outlined in Clause 3.8 of the AGREEMENT.	3.8. The term "REVIEW" shall be conducted quarterly and shall include an assessment of forecast accuracy, component performance, and supply chain health.
BRIEF	3.7. Notwithstanding any other provision of this Agreement, the Contractor shall ensure that all data processing related to the PRODUCTS complies with the NICE Data Protection Principles, and shall notify the relevant party immediately in the event of a data breach or failure to meet the requirements set out in Clause 3.8 by the DUE DATE.	3.8. The CONTRACTOR shall ensure that the PRODUCTS are delivered and maintained in accordance with the service levels specified in the NICE, as agreed upon in the AGREEMENT, to guarantee performance, availability, and responsiveness as defined therein.
BRIEF-L	3.7. Notwithstanding Clause 3.8, if force majeure events prevent delivery of the PRODUCTS under the NICE and AGREEMENT prior to the DUE DATE, the CONTRACTOR shall be excused from performance as provided in Clause 3.8.	3.8. Upon occurrence of force majeure events beyond the CONTROL of the CONTRACTOR, the CONTRACTOR shall be excused from performance of its obligations under the AGREEMENT and the NICE, provided such events materially affect the delivery of the PRODUCTS.
EDGAR EX-10 security agreement, clause 2.6 cites clause 9		
one-pass	2.6. This Agreement is governed by the Uniform Commercial Code (UCC), specifically Article 9, and is subject to the rules and procedures set forth in Clause 9.	9. The UCC governs the perfection, enforcement, and priority of the Security interest in the pledged assets.
BRIEF	2.6. The obligated party warrants that all services provided under the AGREEMENT shall be performed in a merchantable and workmanlike manner, in compliance with the UCC and the PROVISIONS of the SECURITY, and expressly excludes any liability for consequential damages beyond those expressly stated in Clause 9.	9. Notwithstanding any other provision of this Agreement, payment for the OBLIGATIONS shall be made within thirty (30) days of the EVENT, in accordance with the UCC's rules on performance and breach, and shall be due and payable under the UNDER.
BRIEF-L	2.6. During the EVENT, the UNDER shall bear all fees and expenses under UCC [. . .] and Clause 9, except as otherwise provided in the PROVISIONS or by prior written agreement with respect to the SECURITY under the AGREEMENT.	9. Subject to the UCC, all fees and expenses incurred in connection with the OBLIGATIONS shall be borne by the UNDER, except as otherwise agreed in writing prior to the EVENT.