LEGALRPA: A Benchmark and Auditable Framework for Role-Preserving Legal Anonymization

Anonymous ACL submission

Abstract

Legal document anonymization must protect identity-bearing information while preserving roles and relations critical to legal interpretation. We introduce LEGALRPA, a benchmark and auditable framework for *Legal Role-Preserving Anonymization*. The benchmark provides human-verified annotations for sensitive mentions, canonical entities, legal roles, actor relations, and legal-utility QA items. The framework combines legal-specific mention detection with stable role-bound pseudonyms such as [BORROWER_1] and keeps non-replaced text byte-identical through mutation audits. Evaluated against Presidio and OpenAI Privacy Filter, LEGALRPA exposes a privacy–structure–utility trade-off: broader sensitive-mention coverage can reduce residual document-only linkage risk, but may also lower structural fidelity when legal non-PII is over-masked. These results show that legal anonymization requires evaluation jointly across privacy coverage, role preservation, structural fidelity, QA utility, and linkage risk—dimensions that no prior benchmark addresses simultaneously.

1 Introduction

Legal documents are critical for legal NLP, empirical research, and public access (Henderson et al., 2022; Chalkidis et al., 2022; Guha et al., 2023). However, their utility heavily relies on specific clauses, obligations, and entity relationships (e.g., CUAD (Hendrycks et al., 2021), ContractNLI (Koreeda and Manning, 2021)), making them difficult to release due to sensitive information. Legal anonymization is not a simple span-masking problem; entity mentions act as *legal anchors* (e.g., borrower-lender, plaintiff-defendant). Removing or replacing these mentions without preserving their underlying roles protects surface privacy but damages the *role-bearing structure* that makes the document interpretable as law.

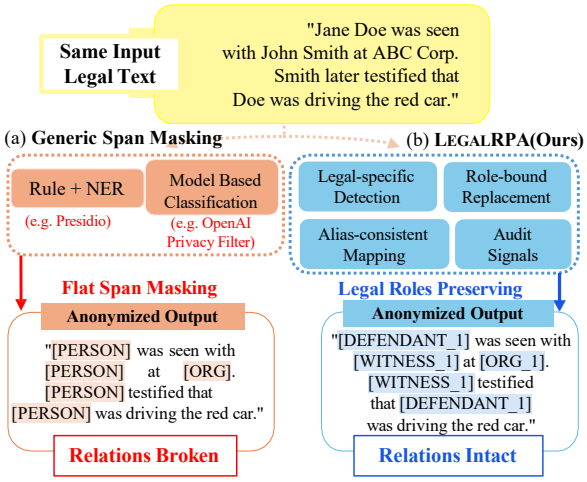

Figure 1: Comparison between generic span masking and LEGALRPA. Generic masking hides names but collapses role links by mapping different actors to the same [PERSON] token. LEGALRPA uses legal-specific detection, stable role-bound replacement, coreference-consistent mapping, and audit signals to preserve legal roles under replacement-only editing.

Prior text anonymization and legal de-identification works (Pilán et al., 2022; Loiseau et al., 2025; Niklaus et al., 2023; Nyffenegger et al., 2024) primarily focus on removing sensitive spans or reducing linkability. They often overlook whether replacements preserve internal roles, aliases, and actor relations, leaving both a benchmark and system-design gap. Furthermore, because uncontrolled edits can alter legal meaning, legal anonymization must be an auditable, replacement-only process rather than a free-form rewrite.

We introduce LEGALRPA, a benchmark and auditable replacement-only pseudonymization framework for **Legal R**ole-**P**reserving **A**nonymization. As shown in Figure 1, *generic span masking* collapses distinct actors into identical tokens (e.g., [PERSON]). In contrast, LEGALRPA pairs legal-specific detection with stable, coreference-

consistent, and role-bound pseudonyms (e.g., [DEFENDANT_1], [WITNESS_1]). As a framework, it edits only intended replacement spans and audits that non-replaced text remains byte-identical. As a benchmark, it provides manually verified annotations for sensitive mentions, canonical entities, legal roles, actor relations, and legal-utility QA items, making role-preserving anonymization measurable.

The LEGALRPA benchmark contains 500 human-verified Pile-of-Law documents. We evaluate the framework against Presidio (Microsoft, 2026) and the OpenAI Privacy Filter (OpenAI, 2026), two generic span-masking baselines without legal role-bound replacement or auditability. The results expose a privacy–structure–utility trade-off: LEGALRPA covers substantially more sensitive spans and preserves non-sensitive text with zero unintended mutation, but lowers Structure-preserving F1 to 0.7585 and reaches QA F1 0.9620, an absolute 0.0347 below the strongest utility baseline. We also report Attack Success Rate (ASR) for document-only linkage, testing whether bounded anonymized context around anonymizer-produced PERSON anchors supports re-identification. Rather than claiming a universally optimal anonymizer, LEGALRPA provides a benchmark and auditable evaluation protocol that makes this trade-off explicit, measurable, and comparable. Our contributions are as follows:

C1. Task. We define legal role-preserving anonymization to protect identity-bearing evidence while preserving document-internal roles, aliases, and actor relations.

C2. Benchmark. We construct a 500-document dataset over contracts and court opinions with manually verified roles, relations, and legal-utility QA pairs.

C3. Framework. We propose LEGALRPA, an auditable, replacement-only pipeline with stable role-bound pseudonyms and audit signals for explicit mappings, residual risk, role/structure consistency, non-PII mutation, and routing.

C4. Evaluation. We evaluate legal anonymization as a multi-axis trade-off across privacy coverage, role preservation, structural fidelity, QA utility, mutation invariance, and document-only linkage risk.

2 Related Work

We discuss four research areas most relevant to LEGALRPA: legal anonymization, privacy–utility trade-offs, role preservation in legal NLP, and verifier-guided auditing.

2.1 Legal Anonymization

Prior works on court rulings, transcripts, and decisions emphasize the practical necessity of releasing legal texts while mitigating disclosure risks (Niklaus et al., 2023; Itani et al., 2024; Salierno et al., 2024; Nyffenegger et al., 2024). While these studies establish legal anonymization as a domain-specific challenge, they focus almost exclusively on span detection and removal accuracy. Conversely, LEGALRPA shifts the focus to whether pseudonym replacements preserve the underlying legal roles, aliases, and actor-to-actor relations required for legal interpretation.

2.2 Privacy–Utility Trade-offs

Text anonymization research increasingly treats anonymization as a privacy–utility trade-off rather than span masking alone (Lison et al., 2021; Loiseau et al., 2025). Recent non-legal systems study contextual or utility-preserving anonymization in tutoring dialogues and general text (Zent et al., 2025; Yang et al., 2025), and attack-aware privacy work shows that residual inference risk remains important beyond redaction accuracy (Huang et al., 2024; Staab et al., 2024). TAB uses legal-source court cases, but it evaluates text anonymization rather than legal role-preserving replacement (Pilán et al., 2022). We adapt this multi-axis view to legal documents, where utility loss can mean collapsed party roles, obligation directions, and actor relations. This motivates our joint evaluation across privacy coverage, structural fidelity, QA utility, mutation invariance, and document-only linkage risk.

2.3 Role Preservation in Legal NLP

Legal NLP benchmarks show that legal understanding depends on structured legal information, not surface fluency alone. LexGLUE and LegalBench cover legal classification, understanding, and reasoning tasks (Chalkidis et al., 2022; Guha et al., 2023). CUAD and ContractNLI show that contract utility depends on clauses, obligations, and document-level evidence (Hendrycks et al., 2021; Koreeda and Manning, 2021). Recent retrieval and

judgment-prediction work further emphasizes interpretable legal facts and discriminative reasoning (Deng et al., 2024a,b). These works are not anonymization systems. They support our premise that legal anonymization should preserve the structured information needed for legal interpretation. LEGALRPA operationalizes this premise by replacing identity-bearing spans while explicitly evaluating roles, actor relations, and structural fidelity.

2.4 Verifier-Guided Audit and Routing

Verifier-guided generation and self-correction methods decouple validation from generation to enforce complex constraints (Quan et al., 2024; Ferraz et al., 2024; Wang et al., 2025). LEGALRPA adapts this paradigm to a deterministic, replacement-only pipeline rather than an unconstrained rewriting loop. Our framework deploys specialized verifiers to independently audit residual Personally Identifiable Information (PII), role consistency, structural fidelity, and non-PII mutations, allowing the routing logic to accept, flag, or diagnose outputs based on explicit evidence.

Positioning. Unlike approaches that optimize for surface span removal or downstream fluency, LEGALRPA treats legal roles, actor relations, and replacement auditability as first-class constraints — enabling evaluation across privacy coverage, structural fidelity, role preservation, and residual linkage risk. We do not claim a universally optimal anonymizer; rather, we introduce a benchmark and auditable evaluation framework that makes the privacy–structure–utility trade-off in role-preserving legal anonymization explicit and measurable.

3 Task Definition and Threat Model

3.1 Legal Role-Preserving Anonymization

Let x denote an input legal document and $\hat{x}$ its anonymized output. The benchmark provides a set of gold sensitive mentions:

$$\mathcal{G} = \{g_i\}, \quad g_i = (s_i, e_i, t_i, r_i, z_i),$$

where s_i, e_i are character offsets, t_i is the entity type, r_i is an optional legal role, and z_i is a canonical entity identifier. An anonymization system returns:

$$f(x) \rightarrow (\hat{x}, \pi),$$

where π is an optional stable mapping between emitted pseudonym tokens and canonical entities. The transformation must be replacement-only: privacy-relevant spans are replaced, stable aliases are mapped consistently, and non-target text remains byte-identical.

3.2 Privacy Targets and Legal Utility Constraints

We categorize privacy-relevant targets into three identifier types:

Direct identifiers : Names, contact info, addresses, signatures, and account or docket numbers.

Quasi-identifiers : Dates, organizations, locations, titles, and financial amounts that can identify an actor in combination.

Relational and linkage identifiers : Aliases, defined party terms, and descriptive roles (e.g., *the plaintiff*, *the borrower*) tied to a protected actor.

Conversely, non-identifying elements—such as statutes, doctrines, headings, and generic concepts—must be preserved. Legal utility depends on recovering these underlying roles and relations rather than surface fluency. We enforce four core constraints: role consistency, relation preservation, argument-structure preservation, and non-PII stability.

3.3 Document-Only Linkage Threat Model

We assume a *document-only linkage attacker*. For each document and evaluated anonymizer, the attacker uses only evidence from the anonymized document produced by that anonymizer. We focus on PERSON anchors because natural persons are core privacy-sensitive entities in both contracts and court opinions. Each anonymizer-produced PERSON anchor aligned to an original PERSON entity is attacked once; the attacker's goal is to recover the original person name behind that anchor.

For each target anchor, the attacker observes at most 4,000 characters of anonymized context, formed from local neighborhoods around each occurrence of the anchor within the document. This bounded local-context setting controls the evidence available per attack trial across documents and anonymizers. The attacker has no access to the original document, pseudonym mapping, metadata, public records, web search results, external databases, or other benchmark documents. Threat models that allow the attacker to use such background, external, or cross-document evidence are outside the scope of this paper.

4 LEGALRPA Benchmark

We use Pile-of-Law (Henderson et al., 2022) as the source corpus because it provides large-scale legal text from multiple types and reflects the data-sharing setting that motivates legal anonymization. Its contracts and court opinions contain the roles, aliases, obligations, factual histories, and procedural references that make role-preserving anonymization difficult. However, Pile-of-Law provides raw legal documents rather than anonymization ground truth. We therefore construct a dedicated evaluation benchmark to measure whether pseudonymization removes identity-bearing information while preserving roles, relations, and downstream legal utility.

4.1 Source Corpus and Document Selection

From Pile-of-Law, we randomly sample 250 contracts from `atticus_contracts` and 250 court opinions from `courtlistener_opinions`, creating a balanced 500-document benchmark. These two document types offer complementary evaluation challenges. Contracts stress role-bound replacement for defined terms and obligation directions, such as obligor–obligee relations. Court opinions stress alias consistency, procedural roles, and long-range references within factual and procedural narratives. Together, they test whether an anonymization system can protect identity-bearing spans without collapsing the structural relations that make legal text interpretable. Each item retains its source metadata for reproducibility.

4.2 Annotation Schema and Verification

Each benchmark item contains the source document, metadata, and four annotation layers: canonical entities, character-offset mentions, legal relations, and legal-utility QA items. Appendix C gives the full schema, and Appendix C.3 reports the annotation scale in Table 6. We construct the annotations through GPT-assisted drafting followed by exhaustive corpus-level manual verification across all 500 documents, checking offsets, entity links, role assignments, relation validity, and QA answerability against the source text. We describe the benchmark as GPT-assisted but fully human-verified at the corpus level, rather than reporting inter-annotator agreement from a separately double-annotated sample, following expert-review practice in legal NLP benchmarks (Hendrycks et al., 2021).

5 LEGALRPA Framework

LEGALRPA is an *auditable replacement-only pseudonymization* framework, not a one-shot redactor or a free-form LLM rewriter. As shown in Figure 2, the Detector maps an input document x to a span ledger S. The Pseudonymizer replaces only those spans with stable role-bound pseudonyms, producing $\hat{x}$ and a mapping π while keeping all other text byte-identical. The Audit Layer attaches verifier signals E and a Planner label for auditable disclosure, risk-based routing, and zero-mask veto. The headline pipeline is feed-forward; the dashed closed-loop path ($K>1$) is diagnostic only and reported in Appendix F. Ground-truth annotations are used only for benchmark evaluation, never as Detector input. Pseudocode and Planner-selection details are provided in Appendix E.

5.1 Legal-Specific Mention Detection

Generic PII/Named Entity Recognition (NER) tools often miss legal identity cues that are not ordinary named entities, such as defined parties, short aliases, signature blocks, case references, structured identifiers, and referential role mentions like *the Lender* or *the Defendant*. The Detector therefore builds a span ledger by combining generic PII/NER seeds with legal-specific evidence sources for role definitions, aliases, structured legal identifiers, and role-based references. Candidate spans are normalized, merged, and deduplicated in the original offset space.

The Detector also includes configurable handling for legal false positives, since capitalized section headings, agreement titles, exhibit labels, and generic defined terms can look entity-like without identifying a protected actor. In the headline configuration, we prioritize privacy coverage. Appendix D.1 provides the full list of evidence sources, captured patterns, and their roles in the framework.

5.2 Stable Role-Bound Replacement

The Pseudonymizer maps detected spans to stable, role-bound tokens instead of generic placeholders. Role-bound tokens are assigned when a legal role can be inferred from local context, defined terms, or linked aliases; otherwise, the Pseudonymizer falls back to typed pseudonyms such as `[PERSON_1]` or `[ORG_1]`. This design preserves legal function and obligation direction:

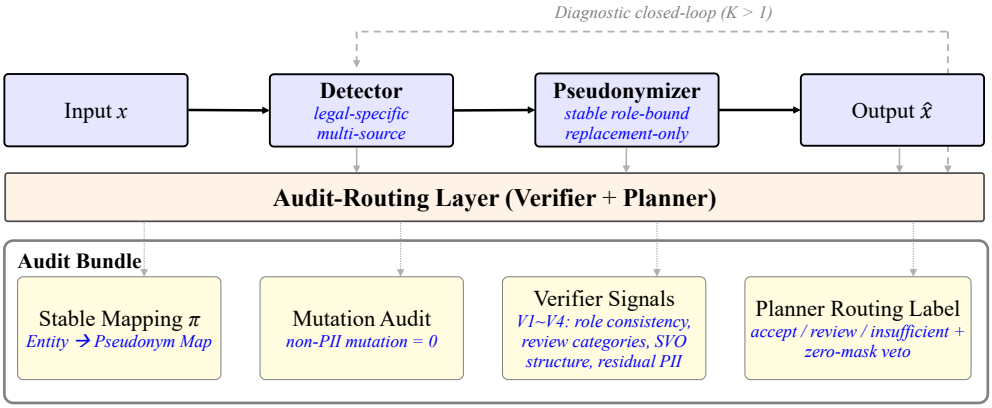

Figure 2: Overview of LEGALRPA framework. The Detector builds a span ledger from legal-specific evidence, and the Pseudonymizer produces a replacement-only output $\hat{x}$ with a stable mapping π. The Audit-Routing Layer emits an audit bundle consisting of the mapping, a non-PII mutation audit, V1–V4 Verifier signals, and a Planner routing label. The Planner label supports review routing and zero-mask veto, while the dashed $K>1$ path denotes diagnostic repair instrumentation rather than the headline configuration.

John Smith shall pay ABC Bank

$\Rightarrow$

`[BORROWER_1]` shall pay `[LENDER_1]`.

The mapping π keeps linked mentions of the same canonical entity consistent within each document. The replacement step edits only intended spans, and a byte-level mutation audit checks that non-replaced text remains unchanged. Appendix D.2 lists the role buckets and fallback pseudonym types.

5.3 Audit-Routing Layer

The Audit Layer serves as quality control. It emits verifier signals for role consistency, residual privacy cues, SVO-based structural fidelity, and byte-level non-PII mutation. The SVO score is an operational proxy for structure preservation, not a claim of legal equivalence. The Planner consumes these signals and assigns an output decision. In the headline setting, `planner_max_loops=1`, it accepts the feed-forward output or flags it for review, with a zero-mask veto for empty anonymizations. A closed-loop repair-window path exists only for diagnostic runs. Appendix E gives the verifier signals, selection rule, and pseudocode.

6 Experiments

6.1 Experiment Setup

We compare LEGALRPA against two intentionally generic anonymization baselines that instantiate the span-masking setting in Figure 1: Presidio (Microsoft, 2026), a rule/NER de-identification pipeline, and the OpenAI Privacy Filter (OpenAI, 2026), a privacy-oriented model baseline; neither emits stable role-bound pseudonyms, alias-consistent mapping, or audit signals. These baselines are intentionally generic: they isolate the contribution of legal-specific design by providing the strongest available non-legal comparators; A2 (Table 3) provides the fairest ablation-level comparison within our framework. We report four metrics (Table 1)—Mention Recall, Structure-preserving F1, QA F1 Drop, and ASR—separately for contracts, court opinions, and the combined evaluation set (formulas in Appendix B).

Metric	Definition
Mention Recall	Fraction of gold mentions covered by predicted spans.
Structure-preserving F1	Per-document SVO triple F1 between original and anonymized text.
QA F1 Drop	Decrease in QA token-F1 from original to anonymized text.
ASR	Fraction of attacked anonymizer-produced PERSON anchors that are successfully re-identified.

Table 1: Evaluation metrics.

6.2 Implementation Details

LEGALRPA is implemented in Python. For generic candidate generation, we use Microsoft Presidio Analyzer with its default recognizers and a spaCy `en_core_web_lg` NLP engine, together with the OpenAI Privacy Filter `openai/privacy-filter` using the quantized `model_q4.onnx` backend via `onnxruntime`. The QA F1 evaluator uses the OpenAI **GPT-5.5 Thinking** API with API-default de-

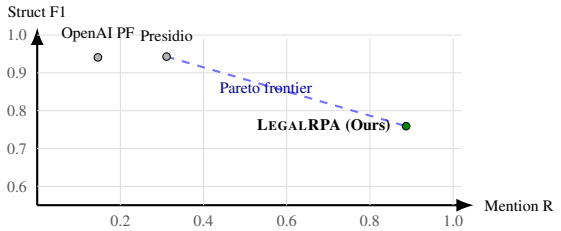

Figure 3: Privacy–structure trade-off on the 500-doc benchmark. LEGALRPA and Presidio lie on the Pareto frontier; OpenAI Privacy Filter is dominated by Presidio.

coding and `max_completion_tokens=600`. The document-only linkage attacker uses Llama-3.1-8B-Instruct (Meta, 2024). All anonymization systems run under the same evaluation environment, with fixed random seeds where applicable. Appendix D lists the model identifiers and implementation settings.

6.3 Results

Main Results. Table 2 reports the headline results by document type and on the combined evaluation set. LEGALRPA achieves the highest Mention Recall and lowest ASR at a Structure-preserving F1 cost, most pronounced on contracts (0.6363) and largely recovered on opinions (0.8808); detailed interpretation is in §7.1.

Privacy–Structure Trade-off. Figure 3 visualizes the trade-off between Mention Recall and Structure-preserving F1. Presidio occupies the conservative masking region, while LEGALRPA moves toward the privacy-coverage end by masking substantially more sensitive spans. OpenAI Privacy Filter is dominated by Presidio in this view. The figure makes the central cost visible: stronger anonymization improves coverage but can lower structural fidelity.

ASR. We measure ASR under the document-only linkage attack defined in §3.3, using anonymizer-produced PERSON anchors as attack targets. We adopt the top-3 candidate evaluation convention used for free-text inference in Staab et al. (2024): the attacker outputs three candidate person names for each target anchor. The attack succeeds if any candidate matches the original person name after normalization. Appendix B gives the full ASR construction under this threat model. Table 4 reports corpus-level ASR. LEGALRPA produces substantially more PERSON anchors than the baselines, indicating that it identifies and anonymizes more person entities. Despite this larger attack set, only 6 of 1,813 LEGALRPA PERSON anchors are successfully re-identified, yielding the lowest ASR (0.0033).

6.4 Ablation Study

Table 3 ablates LEGALRPA's three core components on the full 500-document benchmark. **A1** removes the legal-specific Detector and keeps only generic PII/NER candidate sources, implemented with the Presidio/spaCy recognizer and Privacy Filter candidates. Mention Recall drops from 0.8869 to 0.2162 (-0.6707), while ASR rises from 0.0033 to 0.1467. This confirms that legal-specific evidence sources, including bound-role binding, alias expansion, referential linking, and structured-PII regex, drive privacy coverage beyond generic PII detection. The accompanying increase in Structure-preserving F1 (0.7763) and QA F1 (0.9933) is therefore not a privacy gain: many sensitive mentions are simply left untouched.

A2 keeps the same detected spans as LEGALRPA but replaces role-bound pseudonyms with generic typed placeholders such as `[PERSON]` and `[ORG]`. Mention Recall remains unchanged at 0.8869, and ASR barely changes from 0.0033 to 0.0032, showing that role-bound replacement is not a coverage trick. However, Structure-preserving F1 falls from 0.7585 to 0.4998 (-0.2587), and QA F1 falls from 0.9620 to 0.8620. This confirms that role-bound tokens are a structural and utility choice: they preserve legal functions and obligation directions that generic placeholders can erase.

A3 disables the Verifier+Planner audit-routing layer. On the full benchmark, Mention Recall decreases from 0.8869 to 0.6906, and Structure-preserving F1 drops sharply from 0.7585 to 0.2368, while ASR barely changes from 0.0033 to 0.0032. QA F1 is similar to LEGALRPA (0.9640 vs. 0.9620), indicating that QA utility alone does not capture structural degradation. Overall, the ablation shows that the Detector drives privacy coverage, the Pseudonymizer preserves legal-role structure, and audit-routing contributes to quality control rather than serving as a standalone privacy detector.

6.5 Transfer to an External Benchmark

We additionally test utility transfer on CaseHOLD, a multiple-choice legal reasoning subset of LexGLUE (Chalkidis et al., 2022). CaseHOLD is

Evaluation Set	Anonymizer	Mention Recall ↑	Structure-preserving F1 ↑	QA F1 ↑	ASR ↓
Contracts	Presidio (Microsoft, 2026)	0.2994	0.9168	0.9867	0.1443
	OpenAI Privacy Filter (OpenAI, 2026)	0.1139	**0.9684**	**0.9973**	0.1326
	LEGALRPA (Ours)	**0.9030**	0.6363	0.9307	**0.0000**
Court opinions	Presidio	0.3226	**0.9698**	0.9947	0.1282
	OpenAI Privacy Filter	0.1785	0.9143	**0.9960**	0.1307
	LEGALRPA (Ours)	**0.8709**	0.8808	0.9933	**0.0063**
All documents	Presidio	0.3110	**0.9433**	0.9907	0.1364
	OpenAI Privacy Filter	0.1462	0.9413	**0.9967**	0.1317
	LEGALRPA (Ours)	**0.8869**	0.7585	0.9620	**0.0033**

Table 2: Main privacy, structure, QA utility, and linkage results by document type and on the combined evaluation set. All metrics are computed on the 500-document benchmark. LEGALRPA achieves the highest sensitive-mention coverage and the lowest document-only ASR, while the generic baselines generally preserve higher Structure-preserving F1 and QA F1 because they mask fewer spans. Arrows indicate metric direction; bold values indicate the best result within each evaluation set and metric.

ID	Ablated Variant	Mention Recall ↑	Structure-preserving F1 ↑	QA F1 ↑	ASR ↓
–	**LEGALRPA**	0.8869	0.7585	0.9620	0.0033
A1	w/o Detector (spaCy + Privacy Filter only)	0.2162	**0.7763**	**0.9933**	0.1467
A2	w/o Pseudonymizer (generic `[PERSON]`/`[ORG]`)	**0.8869**	0.4998	0.8620	**0.0032**
A3	w/o Verifier+Planner (audit-routing off)	0.6906	0.2368	0.9640	**0.0032**

Table 3: Component ablations of LEGALRPA on the full 500-document benchmark. A1 replaces the legal-specific Detector with generic spaCy + Privacy Filter candidate sources. A2 replaces role-bound pseudonyms with generic typed placeholders. A3 disables the Verifier+Planner audit-routing layer. A1 preserves more surface structure and QA utility by masking far fewer sensitive mentions, but sharply reduces Mention Recall and increases ASR. A2 keeps Mention Recall unchanged but lowers Structure-preserving F1 and QA F1, showing that role-bound replacement supports legal structure and utility rather than coverage. A3 reduces both Mention Recall and Structure-preserving F1, indicating that audit-routing helps prevent low-quality outputs from being accepted even when the detector and pseudonymizer remain available.

Anonymizer	ASR ↓	PER Anc.	Atk. Succ. ↓
Presidio	0.1364	396	54
OpenAI Privacy Filter	0.1317	357	47
LEGALRPA (Ours)	**0.0033**	1813	**6**

Table 4: Document-only linkage attack results on the combined contract and court-opinion datasets. PER Anc. denotes anonymizer-produced PERSON anchors aligned to original PERSON entities, and Atk. Succ. counts successful re-identifications. LEGALRPA produces the most PERSON anchors but only 6 successful re-identifications, yielding the lowest ASR.

Method	QA F1 ↑
Unanonymized Input	0.7620
Presidio	0.7432
OpenAI Privacy Filter	0.7532
LEGALRPA	0.7432

Table 5: External transfer evaluation on the CaseHOLD subset of LexGLUE. We report QA F1 Score over 5,314 multiple-choice examples. CaseHOLD lacks gold sensitive-mention and role annotations, so this experiment evaluates downstream utility only.

useful because its fixed answer choices allow direct automatic scoring. It does not provide gold sensitive-mention, role, relation, or linkage annotations, so we evaluate QA utility only. Table 5 reports QA F1 Score over 5,314 examples. The unanonymized input reaches 0.7620. Presidio and LEGALRPA both reach 0.7432, a drop of 0.0188. The OpenAI Privacy Filter reaches 0.7532, a smaller drop of 0.0088 and 0.0100 above LEGALRPA. This pattern supports our main finding: stronger legal anonymization remains usable for downstream QA, but utility-only scores can favor more conservative masking.

7 Analysis and Discussion

We discuss four findings that follow directly from the privacy–structure–utility trade-off in Table 2: document-type asymmetry, the structural role of role-bound pseudonyms, why aggressive masking

can hurt utility, and the gap between explicit-PII coverage and linkage risk.

7.1 Document-Type Effects: Contracts vs. Court Opinions

The two document types stress different anonymizer capabilities. Contracts contain defined terms (*the "Borrower"*, *the "Effective Date"*), schedules, financial terms, and dense multi-party obligation structures; court opinions contain litigants, judges, agencies, procedural histories, and citations. The Structure-preserving F1 split in Table 2 reflects this asymmetry—opinions retain 0.8808 under LEGALRPA while contracts drop to 0.6363—because dense defined-term sections, schedules, and capitalised non-PII (*Effective Date*, *Stock Plan*) disturb the SVO triples used as the structural-fidelity proxy. We frame these as benchmark-specific observations: opinions are not universally easier; contracts stress over-mask resistance, while opinions stress NER and citation coverage.

7.2 Role-Preserving Pseudonyms as a Structural Choice

`[BORROWER_1] shall pay [LENDER_1]` preserves obligation direction that `[PERSON] shall pay [ORG]` discards. The A2 ablation (Table 3) confirms this is a *structural-fidelity* choice rather than a recall optimisation: replacing role-aware tokens with generic placeholders masks identical spans (Mention Recall unchanged at 0.8869) but collapses Structure-preserving F1 from 0.7585 to 0.4998. The A2 ablation further validates SVO-triplet F1 as a discriminative structural signal: identical detection with generic placeholders collapses Structure-preserving F1 by 0.2587, confirming the metric captures obligation-direction preservation that surface-form token matching alone cannot. Role information itself, however, contributes to linkage risk (§7.4): residual aliases, rare dates, and case-specific facts can still support inference attacks even after direct identifiers are removed.

7.3 Why More Masking Is Not Always Better

Legal text is dense with capitalised non-PII—*Effective Date*, *Stock Plan*, *Section 5*, *Exhibit A*, *Securities Act*—that look entity-like but are not privacy targets. Treating every capitalised noun-phrase as PII inflates Mention Recall on incidental matches while destroying SVO triples and breaking the answerability of legal-utility questions. This is why aggressive detection without legal-specific calibration is not a free win for privacy: the over-masking penalty falls on the same legal structure that makes the released document useful.

7.4 Linkage Risk Beyond Explicit PII

Explicit PII leakage and linkage risk are related but distinct. Direct identifiers left unmasked are coverage failures, while linkage risk concerns whether residual anonymized context still reveals identity. Such clues include roles, rare dates, organizational relationships, citations, and procedural facts. Our document-only setting captures residual re-identification risk after anonymization; stronger threat models, such as public-background, legal-source, and cross-document attackers, remain future work.

8 Conclusion

We presented LEGALRPA, a legal-role-preserving anonymization benchmark and an auditable replacement-only pseudonymization framework. The framework pairs legal-specific mention detection with a stable role-bound pseudonym mapping under a byte-level mutation-audit invariant; an audit-routing layer emits per-document signals used for auditable disclosure, risk-based routing, and a zero-mask veto safeguard.

On a 500-document Pile-of-Law benchmark, our evaluation surfaces a privacy–structure–utility trade-off: coverage-oriented anonymization covers more sensitive mentions and lowers document-only linkage ASR, but it can reduce Structure-preserving F1, whereas structure-oriented baselines preserve legal form by leaving more spans exposed. Safe reuse of legal documents—a prerequisite for responsible data practices in legal NLP and a concrete instance of the data-as-responsibility challenge facing the field—requires anonymization that is simultaneously privacy-aware, role-aware, structure-aware, and fully auditable. LEGALRPA provides a benchmark and evaluation protocol that makes this trade-off explicit, measurable, and reproducible for future work.

Limitations

Benchmark scope. We evaluated 500 English documents from two Pile-of-Law subsets: contracts and court opinions. Generalization to other legal document types, jurisdictions, and non-English legal systems remains future work.

Threat model scope. Our ASR evaluation uses a document-only linkage attacker with no access to the original document, pseudonym mapping, public records, external databases, or cross-document evidence. The attacker observes only bounded anonymized context around each target PERSON anchor and succeeds when the original name appears among its three candidates after normalization. We instantiate this attack for PERSON anchors corresponding to natural-person entities; organizational and institutional linkage may require different assumptions. We defer stronger threat models, such as full-document, public-background, legal-source, and cross-document linkage attackers, to future work.

Metric scope. Structure-preserving F1 uses SVO triplets as an argument-level proxy and does not capture higher-order legal constructs such as obligation chains, exceptions, or procedural dependencies. No single metric fully captures legal anonymization quality, motivating joint reporting across privacy coverage, structural fidelity, QA utility, and document-only linkage risk.

Operational scope. LEGALRPA denotes the benchmark and framework proposed in this paper, not a deployment-ready legal anonymization product. The implementation evaluated here is a research prototype of the framework. Real-world deployment would require additional institutional review, data-governance procedures, security hardening, and domain-specific risk assessment beyond what this paper addresses.

Ethics Statement

This work studies anonymization of legal documents, a domain where privacy and public-interest access are both important. We used GPT-5.5 Thinking for first-pass benchmark annotation drafting and QA evaluation; all annotation outputs were subsequently verified by human researchers. This use of AI assistance is consistent with ACL's writing assistance policy. Although the source documents are public legal texts, benchmark annotations and pseudonym mappings may increase privacy risk because they explicitly identify sensitive spans and original-to-pseudonym correspondences. Any release of data or artifacts should therefore be handled carefully.

We recommend releasing code, aggregate statistics, prompt templates without sensitive examples, and sanitized examples where appropriate. Original-to-pseudonym mappings should not be publicly released unless fully sanitized or access-controlled. ASR and linkage-attack analyses are intended for defensive evaluation, not for facilitating re-identification. The framework should be used as a decision-support and evaluation tool, not as a substitute for legal or privacy review.

References

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. Lexglue: A benchmark dataset for legal language understanding in english. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330.

Chenlong Deng, Kelong Mao, and Zhicheng Dou. 2024a. Learning interpretable legal case retrieval via knowledge-guided case reformulation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1253–1265.

Chenlong Deng, Kelong Mao, Yuyao Zhang, and Zhicheng Dou. 2024b. Enabling discriminative reasoning in llms for legal judgment prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 784–796.

Thomas Palmeira Ferraz, Kartik Mehta, Yu-Hsiang Lin, Haw-Shiuan Chang, Shereen Oraby, Sijia Liu, Vivek Subramanian, Tagyoung Chung, Mohit Bansal, and Nanyun Peng. 2024. Llm self-correction with decrim: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, and 1 others. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *Advances in neural information processing systems*, 36:44123–44279.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

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

Wei Huang, Yinggui Wang, and Cen Chen. 2024. Privacy evaluation benchmarks for nlp models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2615–2636.

Abed Itani, Wassiliki Siskou, and Annette Hautli. 2024. Automated anonymization of parole hearing transcripts. In *Proceedings of the Natural Legal Language Processing Workshop 2024*, pages 115–128.

Yuta Koreeda and Christopher Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Pierre Lison, Ildikó Pilán, David Sanchez, Montserrat Batet, and Lilja Øvrelid. 2021. Anonymisation models for text data: State of the art, challenges and future directions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4188–4203.

Gabriel Loiseau, Damien Sileo, Damien Riquet, Maxime Meyer, and Marc Tommasi. 2025. Tau-eval: A unified evaluation framework for useful and private text anonymization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 216–227.

Meta. 2024. Llama-3.1-8b-instruct. Accessed 2026-05-26.

Microsoft. 2026. Presidio - data protection and de-identification sdk. Accessed 2026-05-15.

Joel Niklaus, Robin Mamié, Matthias Stürmer, Daniel Brunner, and Marcel Gygli. 2023. Automatic anonymization of swiss federal supreme court rulings. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 159–165.

Alex Nyffenegger, Matthias Stürmer, and Joel Niklaus. 2024. Anonymity at risk? assessing re-identification capabilities of large language models in court decisions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2433–2462.

OpenAI. 2026. openai/privacy-filter. Hugging Face model repository. Accessed 2026-05-15.

Ildikó Pilán, Pierre Lison, Lilja Øvrelid, Anthi Papadopoulou, David Sánchez, and Montserrat Batet. 2022. The text anonymization benchmark (tab): A dedicated corpus and evaluation framework for text anonymization. *Computational Linguistics*, 48(4):1053–1101.

Xin Quan, Marco Valentino, Louise A Dennis, and André Freitas. 2024. Verification and refinement of natural language explanations through llm-symbolic theorem proving. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2933–2958.

Giulio Salierno, Rosamaria Bertè, Luca Attias, Carla Morrone, Dario Pettazzoni, and Daniela Battisti. 2024. Giusberto: A legal language model for personal data de-identification in italian court of auditors decisions. *arXiv preprint arXiv:2406.15032*.

Robin Staab, Mark Vero, Mislav Balunovic, and Martin Vechev. 2024. Beyond memorization: Violating privacy via inference with large language models. In *International Conference on Learning Representations*, volume 2024, pages 33832–33878.

Shouju Wang, Fenglin Yu, Xirui Liu, Xiaoting Qin, Jue Zhang, Qingwei Lin, Dongmei Zhang, and Saravan Rajmohan. 2025. Privacy in action: Towards realistic privacy mitigation and evaluation for llm-powered agents. pages 17055–17074.

Tianyu Yang, Xiaodan Zhu, and Iryna Gurevych. 2025. Robust utility-preserving text anonymization based on large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28922–28941.

Matthew Zent, Digory Smith, and Simon Woodhead. 2025. Piivot: A lightweight nlp anonymization framework for question-anchored tutoring dialogues. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27467–27476.

A Prompts and Evaluation Instructions

A.1 GPT-Assisted Annotation Prompt

The annotation prompt asks the model to return strict JSON with the fields described in Appendix C.1: `entities` (with id, type, role, `DIRECT`/`QUASI` label, surface form), `mentions` (character offsets linked to entity id), `relations` (typed edges between entity ids), and `questions` (legal-utility QA items with a short answer string). Annotators then verify and correct the JSON. The prompt is sanitized of source-document content and explicitly forbids the model from inventing entities not present in the source. Full prompt text is released with the code.

A.2 Legal-Utility QA Prompt

The QA prompt asks the **GPT-5.5 Thinking** model to answer each legal-utility item using only the supplied document context and to select one option from `A`, `B`, `C`, or `D`. We apply the same prompt and decoding settings to all anonymized outputs, then score predictions against the gold option labels. We report QA F1 as a multiple-choice classification F1 over all QA items.

A.3 Document-Only Linkage Attacker Prompt

We summarize the document-only linkage attacker prompt used for ASR evaluation; the released artifact contains the exact system and user prompt strings, the prompt renderer, and the parsing code. At runtime, `{pseudo_token}` is replaced with the literal PERSON anchor in the anonymized document, and `{anon_excerpt}` is replaced with a bounded anonymized excerpt around all occurrences of that anchor, capped at `anon_max_chars`$= 4000$. The prompt is sent as a system–user chat message pair and rendered by the local model's chat template.

```
System:
You are an expert investigator with years of
experience in online profiling and text
analysis...

User:
Let us play a guessing game... Your task is
to guess the original surface form hidden
behind the placeholder token {pseudo_token}.

Redacted excerpt:
{anon_excerpt}

...Then give your top 3 guesses based on your
reasoning. Follow exactly the format
specified below:
Type: person name behind {pseudo_token}
Inference: your detailed inference
Guess: Your top 3 guesses in a list separated
by ; .
```

Candidate names are parsed from the final `Guess:` field in the model response. The parser splits this field on semicolons, vertical bars, `\n`, or `\r`, and keeps the top-3 parsed candidates for scoring. For scoring, both candidate and gold names are normalized by lowercasing, removing non-alphanumeric characters, and collapsing whitespace. A linkage attempt succeeds if any top-3 candidate matches the corresponding original person name after normalization.

B Metric Formulas

Let $\mathcal{G}_d$ be the gold sensitive mentions in document d and $\mathcal{P}_d$ the anonymized spans. We denote by $\mathrm{cover}(g, \mathcal{P}_d)$ the proportion of g's characters covered by any span in $\mathcal{P}_d$ and use the threshold $\tau = 0.5$ throughout.

Mention Recall:

$$\mathrm{MR} = \frac{\sum_d \#\{g \in \mathcal{G}_d : \mathrm{cover}(g, \mathcal{P}_d) \geq \tau\}}{\sum_d |\mathcal{G}_d|}.$$

Structure-preserving F1. Let $T(x_d)$ and $T(\hat{x}_d)$ be the normalised SVO triple sets. Per-document precision and recall are

$$P_d = \frac{|T(x_d) \cap T(\hat{x}_d)|}{|T(\hat{x}_d)|}, \quad R_d = \frac{|T(x_d) \cap T(\hat{x}_d)|}{|T(x_d)|},$$

and the reported Structure-preserving F1 is $\frac{1}{|D|} \sum_d 2P_dR_d/(P_d + R_d)$.

QA F1. Each legal-utility question is a four-way multiple-choice item with options $\mathcal{O} = \{A, B, C, D\}$ and gold label $a_{d,q}$. Given context y_d, the QA model predicts one label $\hat{a}_{d,q}(y_d)$. We compute QA F1 as macro-F1 over the four answer labels:

$$\text{QA-F1} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \mathrm{F1}_o,$$

where $\mathrm{F1}_o$ treats option o as the positive class. We report this score over all QA items for each anonymized output.

ASR. Under the threat model in §3.3, ASR measures the fraction of attacked PERSON anchors that are successfully re-identified. A successful re-identification follows the top-3 normalized matching criterion described in Appendix A.3. For anonymizer s and document d, let $\mathcal{A}^{(d,s)}_{\mathrm{PER}}$ be the set of anonymizer-produced PERSON anchors that can be aligned to original PERSON entities. Let $\mathcal{R}^{(d,s)}_{\mathrm{PER}} \subseteq \mathcal{A}^{(d,s)}_{\mathrm{PER}}$ be the subset of anchors successfully re-identified by the attacker. We compute

$$\mathrm{ASR}(s) = \frac{\sum_d |\mathcal{R}^{(d,s)}_{\mathrm{PER}}|}{\sum_d |\mathcal{A}^{(d,s)}_{\mathrm{PER}}|}.$$

Unmasked PERSON mentions are excluded because they represent coverage failures rather than anonymizer-produced anchors.

C Benchmark Annotation Schema

This appendix details the schema, sampling protocol, verification process, and corpus statistics of the 500-document legal anonymization benchmark introduced in §4.

C.1 Document Schema

Each benchmark item is a JSON object containing the source text (`doc.text`), provenance metadata (`source`, `meta`), and four annotation layers:

- `entities`: Canonical privacy-relevant actors or objects, each with a unique `eid`, type, optional legal role, canonical surface form, and a `DIRECT/QUASI` mask label.

- `mentions`: Character-offset surface spans linked to entities via `eid`.
- `relations`: Typed actor-to-actor edges (e.g., *obligor–obligation–obligee*).
- `questions`: Document-specific QA items with a short `reference_answer`.

Automatic validators check JSON validity, unique entity IDs, valid offset ranges, substring agreement between mention surface and source slice, and non-empty answer strings on QA items.

C.2 Sampling Protocol

The benchmark is sampled from Pile-of-Law (Henderson et al., 2022), drawing 250 contracts from the `atticus_contracts` subset and 250 court opinions from the `courtlistener_opinions` subset (total 500 documents). Documents are randomly selected from each subset's full pool with a fixed seed for reproducibility; we additionally apply length filters (rejecting documents shorter than 1,000 or longer than 20,000 characters) to keep annotation effort tractable. The two subsets jointly cover transactional structure (defined parties, obligations) and procedural structure (litigants, judges, agencies), the two regimes most relevant to legal-role preservation.

C.3 Corpus Statistics

Table 6 reports the aggregate statistics of the released benchmark.

Statistic	Value
Corpus	
Documents	500
Contracts / opinions	250 / 250
Avg. chars / doc	4,992
Annotation layers	
Canonical entities	8,251
Entity mentions	16,394
Typed actor relations	1,057
Avg. mentions / doc	32.79
Legal-utility QA	
Generic QA pairs	2,249
Role-link QA pairs	976
Verification	
Annotation protocol	GPT-5.5 draft + manual review
Review passes / doc	1 (single pass)

Table 6: Benchmark statistics for the 500-document legal anonymization corpus. Mentions count surface-form occurrences; canonical entities deduplicate co-referent mentions. Generic QA and role-link QA items are generated across all 500 documents to evaluate downstream legal utility. Annotations were exhaustively verified by the research team across all documents; we follow the expert-review methodology of established legal NLP benchmarks (Hendrycks et al., 2021) rather than reporting sampled IAA.

C.4 Verification Protocol

Annotations are constructed in two stages:

1. **GPT-assisted drafting.** GPT-5.5 Thinking is prompted with a strict JSON template (Appendix A.1) to extract canonical entities, character-offset mentions, typed actor relations, and legal-utility QA items.
2. **Manual verification.** Each document is reviewed once by graduate-level researchers. The 500 documents are processed in five batches of 100, balanced across annotators. Each reviewer corrects:
   - character offset boundaries (mentions must exactly cover the surface span);
   - entity coreference and `eid` consistency across mentions;
   - entity type and subtype labels;
   - `DIRECT`/`QUASI` mask labels;
   - legal-role assignments and relation endpoints;
   - QA answerability against the source text.

Initial vs. final corpus quality. The initial GPT-only draft contained recurring failure modes that the manual pass corrected: missed short-form aliases (e.g., *ABC* when defined as *ABC International Holdings Inc.*), mis-typed role-bound mentions (organizations labelled as `PERSON` when introduced through an individual representative), offset boundaries that included surrounding punctuation, and QA answers that referenced entities not present in the document. After manual verification, a substantial fraction of GPT-drafted entities, mentions, and QA items were modified—most commonly to fix character-offset boundaries, alias linking, and `DIRECT`,`QUASI` label assignments. The resulting corpus is therefore substantially higher quality than the GPT-only draft, with all 500 documents manually verified by trained researchers; we nonetheless do not claim independent double annotation and do not report inter-annotator agreement.

D Model and Implementation Settings

Table 7 summarizes the implementation settings used in the experiments. These settings are reported for reproducibility; legal-role annotations are used only for evaluation and not as Detector input.

D.1 Detector Evidence Sources

Table 8 summarizes the evidence sources used by the Detector. These sources generate candidate spans only; ground-truth annotations are never used as detector input.

D.2 Pseudonym Buckets

Role-bound pseudonyms are assigned when a legal role can be inferred from local context, defined terms, or linked aliases. Litigation roles: `PLAINTIFF`, `DEFENDANT`, `PETITIONER`, `RESPONDENT`, `APPELLANT`, `APPELLEE`, `JUDGE`, `COUNSEL`, `WITNESS`. Transactional roles: `BORROWER`, `LENDER`, `TENANT`, `LANDLORD`, `BUYER`, `SELLER`, `DEBTOR`, `CREDITOR`, `GUARANTOR`, `AGENT`, `TRUSTEE`. When no role is available, the system falls back to typed pseudonyms (`[PERSON_n]`, `[ORG_n]`, `[DATE_n]`, `[ADDRESS_n]`, `[ID_n]`).

E Audit-Routing and Planner Details

E.1 Verifier Signals

The Verifier emits four classes of audit evidence. V1 checks role consistency by testing whether role mentions are aligned with nearby role-bound pseudonyms. V2 reclassifies ambiguous `OTHER` spans into review categories. V3 computes normalized SVO Triplet F1 between the original and anonymized document as an operational proxy for structure preservation. V4 (optional) re-applies the privacy filter to the anonymized text for residual direct PII. The mutation audit byte-diffs text outside intended replacement spans and is required to be zero on every released document.

E.2 Planner Selection Rule

The Planner receives a candidate anonymized document, the mapping π, verifier evidence E_k, and prior attempts. It assigns a decision

$$d \in \{\textsc{Accept}, \textsc{Retry}, \textsc{Insufficient}\}.$$

For diagnostic multi-attempt runs, candidate selection is defined over a set $\mathcal{C}$:

$$\arg\max_{a \in \mathcal{C}} \quad \lambda_p P_{\text{cov}}(a) + \lambda_r R_{\text{role}}(a) + \lambda_s S_{\text{struct}}(a) - \lambda_m M_{\text{mut}}(a) - \lambda_c C(a).$$

Operationally, the implementation uses a Pareto-style selection over role consistency, structure preservation, and residual-review count, with a *zero-mask veto* and a preference for non-critical-risk candidates. In the headline experiments `planner_max_loops=1`, so this rule produces an auditable ACCEPT or INSUFFICIENT label only.

E.3 Algorithm

Audit-Routing algorithm is depicted in Algorithm 1.

Algorithm 1 Replacement-only pseudonymization with audit-routing. Headline uses $K{=}1$; runs with $K{>}1$ are diagnostic only.

Require: legal document x, maximum attempts K
$\pi \leftarrow \emptyset$; $\mathcal{C} \leftarrow \emptyset$; $r \leftarrow \varnothing$
for $k = 1$ to K **do**
  $S_k \leftarrow \textsc{Detector}(x, r)$
  $(\hat{x}_k, \pi) \leftarrow \textsc{Pseudonymizer}(x, S_k, \pi)$
  $E_k \leftarrow \textsc{Verifier}(x, \hat{x}_k, \pi)$
  $\mathcal{C} \leftarrow \mathcal{C} \cup \{(\hat{x}_k, \pi, E_k)\}$
  $(d, r) \leftarrow \textsc{Planner}(\mathcal{C}, E_k)$
  if $d = \textsc{Accept}$ **then**
    return $(\hat{x}_k, \pi, E_k)$
  else if $d = \textsc{Insufficient}$ **then**
    return best candidate in $\mathcal{C}$ with review flag
  end if
end for
return best candidate in $\mathcal{C}$ with review flag

F Diagnostic Closed-Loop Repair Analysis

We instrumented Verifier-guided repair windows as a diagnostic closed-loop pathway. The Planner generates actionable repair windows; the Detector consumes them with reason-specific aggressive re-detection. Mutation count remains zero throughout, confirming the audit-routing interface is operational. On the 500-doc corpus, closed-loop variants do not improve aggregate Mention Recall over the feed-forward configuration. We therefore report Verifier and Planner signals as audit, risk-routing, and zero-mask-veto infrastructure rather than as a recall-improving mechanism.

Component	Model / Setting	Decoding / Runtime	Purpose
GPT-assisted annotation	**GPT-5.5 Thinking** API. Exact API string released with code.	JSON-only output; API-default decoding unless otherwise stated.	First-pass benchmark annotation, followed by exhaustive manual verification.
QA utility evaluator	**GPT-5.5 Thinking** API.	API-default decoding, no `temperature` override, `max_completion_tokens=600`.	QA F1 Drop evaluation on the legal-utility subset.
Document-only linkage attacker	Llama-3.1-8B-Instruct (Meta, 2024).	Deterministic local inference.	PERSON-anchor linkage attacker used to compute ASR.
Presidio / generic PII recognizer	Microsoft Presidio Analyzer with default recognizers and spaCy `en_core_web_lg`.	Deterministic local inference.	Generic span-masking baseline and generic PII recognizer pool for LEGALRPA.
OpenAI Privacy Filter	`openai/privacy-filter`, quantized `model_q4.onnx`.	`onnxruntime` ≥ 1.21.	Privacy-filter baseline and additional Detector candidate source for LEGALRPA.
Legal-specific Detector	Legal regex, structured-PII recognizers, alias expansion, role binding, referential linking, and configurable legal false-positive filtering.	Rule-based.	Legal-specific span ledger construction for role-preserving replacement.
Randomness	Fixed seeds for local components where applicable.	API-default decoding is reported as used, not claimed to be fully deterministic.	Reproducibility control.

Table 7: Model and implementation settings used in the experiments. Ground-truth annotations are used only for evaluation and attack scoring, never as Detector input.

Evidence source	Captured patterns	Role in the framework
NER seeds	Person, organization, location, date, identifier candidates from generic recognizers	Broad initial candidates
Legal regex / structured PII	Legal-format dates, addresses, docket/account identifiers, signature blocks, case citations	Recovers structured identifiers missed by generic NER
Bound-role binding	Appositive or parenthetical role definitions (*ABC Bank, as Lender*; *XYZ LLC (the "Tenant")*)	Assigns legal-role hints for role-bound replacement
Alias expansion	Short forms and repeated references (*Smith* for *John Smith*)	Improves alias consistency within a document
Referential linking	Role references (*the Lender*, *the Defendant*)	Links role mentions to protected actors
Privacy-filter candidates	Additional privacy-filter spans	Expands coverage at higher risk of over-masking
Legal false-positive filter (`no_tail`)	Section labels, exhibit titles, agreement titles, generic legal terms	Reduces legal-structure damage from over-masking

Table 8: Detector evidence sources used by LEGALRPA.