RECAST: CONTRACT-PRESERVING PROTECTION FOR FIXED-INTERFACE MULTIMODAL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Remote multimodal models offer strong numerical reasoning capabilities over charts and speech, but sending private inputs risks exposing sensitive content. Text-only sanitization cannot directly satisfy fixed media interfaces, while identity anonymization leaves the underlying task content exposed. We introduce RECAST, an agentic plug-in framework that replaces source-specific content while preserving task-relevant relations and the required input modality. RECAST locally converts inputs into a shared textual evidence–query record, jointly rewrites entities and topics with a distilled 4B model, and substitutes values through a locally invertible, role-aware numerical map. A reconstruction agent generates and validates the required media from the protected record. The remote solver returns a program whose protected operands are restored locally before execution. On 4,000 held-out ChartQA and NMSQA examples, RECAST achieves 75.10% accuracy, retaining 92.43% of unprotected remote accuracy, while a model-based audit flags source-content retention in 7.95% of solver-bound requests. It outperforms all evaluated local baselines, preserving the benefit of remote reasoning while reducing source-content exposure under existing media interfaces.

1 INTRODUCTION

Local-remote collaboration allows users to keep private data on their devices while leveraging a more capable remote model. This setting has been studied extensively for text. Before a request is released, a local model can rewrite it through differentially private rewriting (Meisenbacher et al., 2024; Thareja et al., 2025), adversarial anonymization (Duan et al., 2026), or privacy-conscious delegation (Hartmann et al., 2024; Siyan et al., 2025). For numerical reasoning, Zhang et al. (2026) shift the topic of a textual instance, substitute its numbers, and then recover the original answer locally from a program returned by the remote model.

However, much private evidence is not textual. Business reports present their figures in charts, while meetings and customer calls convey theirs in recordings. Reasoning over such evidence requires strong multimodal perception and numerical computation, areas in which on-device models still lag far behind remote ones. These remote models are typically accessed through service interfaces with fixed request schemas: a chart-analytics endpoint expects an image, and a spoken question-answering endpoint expects a waveform. Protection must therefore extend beyond text to the media these services consume.

Existing multimodal protections fall short of this need. Visual instruction rewriting (Mishra et al., 2025) and symbolic scene representations (Bara et al., 2022) convert the input into text, which media interfaces do not accept. Media-level protections preserve the modality but target different threats: speaker anonymization conceals voice identity (Meyer et al., 2022), adversarial perturbation prevents unauthorized speech recognition (Jin et al., 2025), and learned image transformations hide visual privacy attributes from multimodal models (Hui et al., 2026a). In each of these media-level methods, the content itself, such as which entity has which value in a chart or in speech, reaches the remote model intact. Numerical tasks impose a further requirement: evidence and query must be transformed consistently, and the transformed quantities must still support both the requested computation and the recovery of the original answer.

RECAST closes this gap by separating what a task needs from what the source reveals. The task needs relations, namely which quantity belongs to which entity, which pair the query compares, and which

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

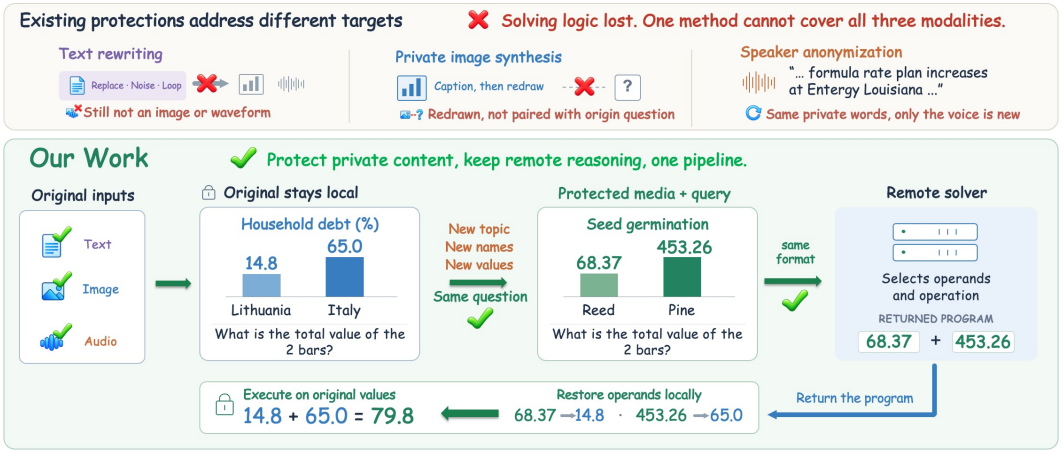


Figure 1: **RECAST: replacing source content while preserving task structure. Top:** The illustrated alternatives address different targets: text rewriting does not directly provide the required media, dataset-level image synthesis is not paired with a specific query, and speaker anonymization retains spoken content. **Bottom:** RECAST jointly replaces topics, entities, and values across evidence and query while retaining task-relevant relations and the required modality. The remote solver returns a program over protected operands; the client restores the original operands locally before executing the program to recover the answer. The chart example illustrates a unified protection framework for text, charts, and speech.

operation it requests, whereas the source contributes entities, topics, wording, and values. RECAST is an agentic plug-in pipeline built around a shared textual evidence–query record (Fig. 1). *Evidence capture* plugs in a module for the input modality and locally converts the chart or recording into this record. A *rewrite module* jointly replaces source-specific content across evidence and query while preserving their relations, and maps the numbers under a per-request map whose inverse remains local. A *modality reconstruction agent* plugs in a generator for the interface contract and renders the protected record into the required modality through a generate-review-repair loop before anything reaches the solver. The remote solver returns a Python program over marked protected operands, which the local side restores and executes on the original values, extending the recovery scheme of Zhang et al. (2026) to media inputs.

The main contributions of this work are as follows:

- We propose RECAST, an agentic plug-in pipeline that links evidence capture, rewriting, and modality reconstruction through a white-box, model-independent evidence–query record, so that per-modality and per-interface modules plug in without retraining.
- We design a rewrite module in which a distilled local 4B Rewriter jointly changes names and domain while fixing relations, and a role-aware numerical map with local inversion enables exact answer recovery.
- On 4,000 held-out ChartQA and NMSQA examples, RECAST keeps 92.43% of unprotected accuracy with only 7.95% source retention. Distillation adds 6.40 accuracy points and cuts retention from 39.75% to 7.95%, and on ChartQA the text release is 10.60 points above chart re-rendering.

2 RELATED WORK

2.1 TEXT PRIVACY

Text privacy methods range from word-level sanitization to document rewriting. SanText and CusText replace words with semantically related alternatives under differential privacy (Yue et al., 2021; Chen et al., 2023), while DP-Prompt and DP-MLM use language models to produce coherent protected text (Utpala et al., 2023; Meisenbacher et al., 2024). DP-Fusion instead limits the influence of

sensitive tokens by combining generation distributions conditioned on different versions of the input (Thareja et al., 2025). Beyond explicit identifiers, contextual clues can also reveal personal attributes (Staab et al., 2024). Adversarial anonymization uses such inference attacks to guide text revisions, with RLAA adapting this approach to local models (Staab et al., 2025; Duan et al., 2026).

Recent methods also account for task utility and semantic structure. STAMP considers both task relevance and privacy sensitivity when allocating privacy budgets, while DP-ST uses semantic triples to guide private document generation (Tian et al., 2026; Meisenbacher et al., 2025). These developments are relevant to our goal of preserving useful relations while replacing source content. Our setting adds two coupled requirements: evidence and query must remain aligned after numerical substitution, and the result must be usable through a fixed media interface. Text protection alone does not provide the required chart or waveform. RECAST therefore uses text as an editable intermediate shared across modalities, then reconstructs the protected record in the format the interface accepts.

2.2 LOCAL–REMOTE COLLABORATION

Local–remote collaboration uses a local model to control what a stronger remote model sees. Hartmann et al. (2024) study how a model can seek help from another without revealing private information. CoGenesis and PAPILLON coordinate local and remote generation while limiting the private context exposed to the remote model (Zhang et al., 2024; Siyan et al., 2025), whereas InferDPT combines local perturbation, remote generation, and local extraction (Tong et al., 2025). More recent work makes delegation sensitive to the task and its privacy risks: Privacy-R1 learns which text to transmit, Veilmind conditions sanitization on user intent, and privacy-risk estimation accounts for contextual disclosures beyond direct identifiers (Hui et al., 2026b; Liu et al., 2026; Siyan et al., 2026).

The closest work is that of Zhang et al. (2026), who shift the topic of a textual numerical-reasoning instance, substitute its numbers, and recover the original answer locally from a program returned by the remote model. Separating program generation from execution is also central to program-aided reasoning (Gao et al., 2023). RECAST extends this recovery scheme to media inputs and outputs. A shared textual evidence–query record connects modality-specific evidence capture, joint rewriting, and reconstruction. The record makes cross-field references explicit, while reconstruction checks whether the released chart or speech still represents the protected evidence. The distinction is not simply the use of a multimodal solver: the released artifact must satisfy the media interface and retain the relations needed for local answer recovery.

2.3 MULTIMODAL AND MEDIA PRIVACY

Privacy risks also arise from information embedded in media: vision–language models can infer personal attributes even from images without a direct depiction of the individual (Tömekçe et al., 2024). One response is to replace the raw input with an abstract representation. Symbolic scene representations support privacy-preserving visual question answering (Bara et al., 2022), and ReVision rewrites image-dependent instructions into self-contained text (Mishra et al., 2025). Outside privacy protection, DePlot similarly separates chart perception from language reasoning through plot-to-table translation (Liu et al., 2023). These approaches motivate an explicit intermediate representation, but their textual or symbolic outputs cannot directly replace an image or waveform required by a fixed media interface.

Media-level methods retain the modality but target different aspects of privacy. Speaker anonymization changes voice identity while retaining linguistic content (Meyer et al., 2022), SpeechShield masks sensitive speech segments before cloud recognition (Benazir & Lin, 2025), and learned image transformations protect visual privacy attributes from black-box multimodal models (Hui et al., 2026a). SPTI is particularly related in using private textual intermediaries to synthesize differentially private images (Wang et al., 2025). It releases a synthetic dataset, however, rather than a protected media–query pair tied to a particular computation. RECAST instead replaces source-specific entities, topics, and values for each request while retaining the evidence–query relations needed for numerical reasoning. Protection comes from transforming the shared record; reconstruction provides interface compatibility, and the inverse numerical map remains local for answer recovery.

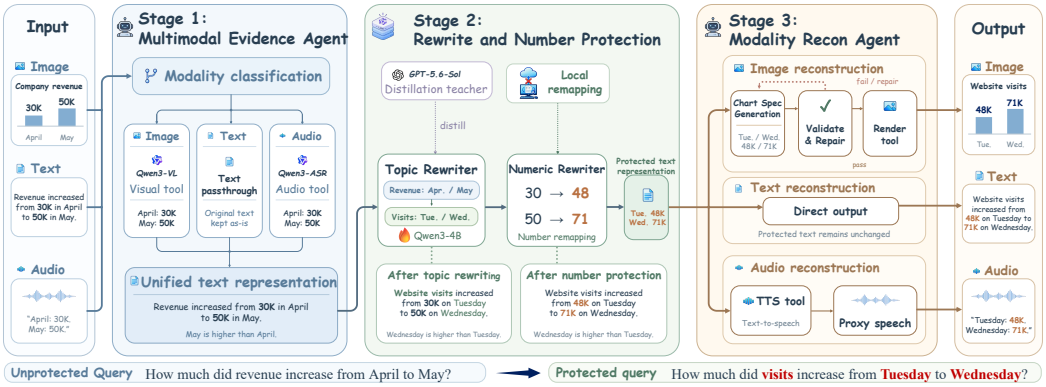


Figure 2: **Overview of RECAST.** The pipeline first converts multimodal inputs into a unified textual representation (Sec. 3.2), then applies topic rewriting abd local numerical remapping to protect entities and numbers (Sec. 3.3). The protected text is subsequently reconstructed into a surrogate in original modality, which regenerates charts through ChartSpec-based rendering and validation, passes through text directly, synthesizes audio as proxy speech (Sec. 3.4). An example in the figure shows this pipeline.

3 METHOD

As shown in Fig. 2, RECAST separates source-specific content from task-relevant relations through a shared textual evidence–query record. It captures evidence locally (Sec. 3.2), jointly rewrites its semantic and numerical content (Sec. 3.3), and reconstructs the protected record in the required modality (Sec. 3.4). The remote solver returns a program whose operands are restored locally before execution.

3.1 TASK AND FRAMEWORK OVERVIEW

Given private evidence X and a textual query q, let $m, r \in \{\text{text}, \text{image}, \text{audio}\}$ denote the source and accepted release modalities, respectively. We primarily consider $r = m$, using another release modality only when the interface permits it. The goal is to reduce exposure of source-specific entities, topics, wording, and values while retaining the evidence–query associations needed for local answer recovery.

Let E_m be the local evidence-capture operator, $D = E_m(X, q)$ its normalized textual record, R_θ the semantic rewriter with trainable adapter parameters θ, and H_h the numerical transformation induced by a request-specific injective map h. For requests passing release validation,

$$\widetilde{D} = H_h(R_\theta(D)), \qquad (\widetilde{X}, \widetilde{q}) = G_r(\widetilde{D}). \tag{1}$$

Here, $\widetilde{D}$ is the protected record; $\widetilde{X}$ and $\widetilde{q}$ are the released evidence in modality r and its protected textual query. G_r denotes reconstruction and validation; its local validation inputs (X, q, D) and inverse map h^{-1} are implicit arguments. During deployment, these inputs and the inverse map remain local. Remote reconstruction services receive only protected content and derived artifacts; the solver receives $(\widetilde{X}, \widetilde{q})$. The retention audit covers solver-bound requests, not all reconstruction-service inputs.

3.2 MULTIMODAL EVIDENCE CAPTURE

Evidence capture makes task-bearing content explicit without answering the query. For charts, local, question-aware Qwen3-VL-8B-Instruct (Bai et al., 2025) extracts category–value associations, legend–series correspondences, units, ordering, and queried visual attributes. Qwen3-ASR-1.7B (Shi et al., 2026) transcribes speech, while a deterministic parser normalizes text and preserves field boundaries and numerical spans.

All branches share five fields: `id`, `modality`, `context`, `question`, and `choices`, with empty choices for numeric-answer tasks. Numerical expressions are normalized before rewriting to match repeated quantities across evidence and query. This explicit, model-independent interface makes content inspectable and allows capture plug-ins to be replaced independently, provided their outputs remain schema-compatible.

3.3 Joint Rewriting and Numerical Protection

As shown in Fig. 3, this stage protects the shared evidence–query record through two complementary transformations. A distilled local rewriter jointly changes entities and topics while maintaining evidence–query alignment; a request-specific numerical map then substitutes values under explicit relational constraints. The inverse map remains local for restoring operands in the program returned by the remote solver.

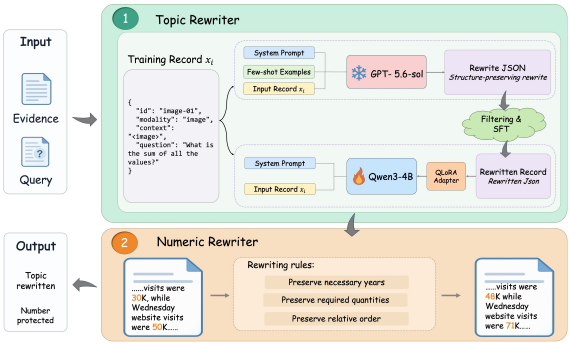


Figure 3: **Joint rewriting and numerical protection.** A distilled local rewriter changes entities and topics consistently across evidence and query while retaining numerical literals. A separate request-specific map replaces numerical values under the supported relational constraints. Its inverse remains local for operand recovery.

Joint semantic rewriting. Evidence and query are rewritten together into a coherent, semantically distinct domain, retaining quantity assignments, cross-references, the requested operation, and answer type. Chart rewrites maintain category–value and legend–series correspondences, including queried visual attributes; speech rewrites maintain entity–quantity assignments, coreference, event order, and temporal or comparative links. Numerical literals remain unchanged, separating semantic rewriting from numerical substitution.

Distilling a local rewriter. We fine-tune Qwen3-4B (Yang et al., 2025) with QLoRA (Dettmers et al., 2023) on offline GPT-5.6-Sol (OpenAI, 2026) rewrites of training-side chart descriptions and speech transcripts. Schema and per-field numerical-literal filters yield 15,795 accepted pairs, split into training and development sets disjoint from evaluation data. Deployment uses the local student for rewriting. Let $\mathcal{T}$ be the training split and Y the token sequence of the serialized teacher rewrite of D, before numerical substitution. We minimize

$$\mathcal{L}_{\text{rewrite}}(\theta) = -\sum_{(D,Y)\in\mathcal{T}} \sum_{t=1}^{|Y|} \log p_\theta(Y_t \mid D, Y_{<t}). \tag{2}$$

Here, Y_t is the t-th target token, $|Y|$ the sequence length, $Y_{<t}$ its preceding prefix, and p_θ the student's conditional next-token distribution. Loss applies only to output tokens. Inference outputs failing schema or numerical-literal checks are regenerated; semantic consistency is checked separately before release. Appendix A.4 gives filtering and split details.

Numerical rewriting. Let $\mathcal{V}$ contain the distinct normalized values in the evidence and query. Each request uses an injective map $h : \mathcal{V} \to \widetilde{\mathcal{V}}$, where $\widetilde{\mathcal{V}} = h(\mathcal{V})$, with local inverse $h^{-1} : \widetilde{\mathcal{V}} \to \mathcal{V}$. Let $\mathcal{B}_k \subseteq \mathcal{V}$ be the k-th magnitude band of remapped ordinary quantities and $\mathcal{Y}_g \subseteq \mathcal{V}$ the g-th contextual year group. Collision-free replacements are sampled and sorted within each band, while years in a group share an offset δ_g:

$$v < v' \implies h(v) < h(v'), \qquad v, v' \in \mathcal{B}_k, \tag{3}$$

$$h(y) = y + \delta_g, \qquad y \in \mathcal{Y}_g. \tag{4}$$

Here, v, v' are ordinary values and y is a year value. Repeated values receive the same replacement; non-year integers from 0 to 15 use identity assignments $h(v) = v$. The map preserves within-band order and within-group year intervals, but not general differences, ratios, or cross-class comparisons. Construction details appear in Appendix A.3.

Local operand recovery. Following Zhang et al. (2026), the solver returns a Python program $\widetilde{P}$ with observed numerical operands marked as `protected(<number>)`. Let $\mathcal{I}_{h^{-1}}$ replace each marked operand `protected`$(h(v))$ with its original value v. The client computes

$$\widehat{a} = \mathrm{Exec}\Big(\mathcal{I}_{h^{-1}}[\widetilde{P}]\Big), \tag{5}$$

where $\mathrm{Exec}(\cdot)$ denotes local program execution and $\widehat{a}$ is the predicted answer to q on X. Recovery acts on operands, not on an answer computed from protected values. Correct recovery requires faithful evidence and correct operand and operation selection. Decisions based on unpreserved differences or ratios need not transfer unless deferred to program execution after restoration.

3.4 Modality Reconstruction and Release Validation

As shown in Fig. 4, the reconstruction agent converts $\widetilde{D}$ into an interface-compatible artifact through a generate–review–repair loop. It aims to represent protected evidence faithfully, not to reproduce the original chart appearance or speaker identity.

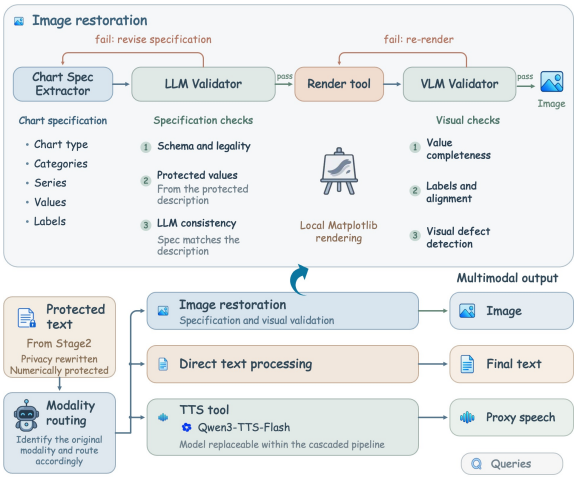


Figure 4: **Modality reconstruction and release validation.** The interface selects the reconstruction plug-in. Charts use specification generation, local rendering, and review-guided repair; speech uses synthesis followed by transcription-based checks; text is serialized directly. Release requires both protected-record consistency and a local source-consistency check.

Chart reconstruction. A remote code model translates protected evidence into a declarative ChartSpec containing the chart type, categories, series, values, labels, units, and colors. After local structure and value checks, fixed Matplotlib templates (Hunter, 2007) render the specification. A remote vision model reviews the image against the protected record, providing structured feedback for at most four repairs. Neither remote model receives the original chart. Failed repairs trigger a validated table fallback only when accepted by the interface; otherwise, the request is withheld.

Speech and text reconstruction. A remote text-to-speech (TTS) service synthesizes protected passages. The controller transcribes each candidate waveform, checks intelligibility, protected quantities, and entity references against the passage, and regenerates failures. For text-compatible interfaces, protected evidence is serialized directly with the protected query.

Two complementary release checks. Let $V_{\mathrm{protected}}, V_{\mathrm{source}} \in \{0, 1\}$ indicate whether a candidate passes the protected-record and local source-consistency checks, respectively, with 1 for pass and 0 for failure. The former checks rendered values, labels, units, associations, and queried attributes against $\widetilde{D}$ for charts, and transcribed quantities, references, and query-relevant content for speech. The latter aligns rewritten entities with source roles, restores mapped values, and compares assignments and dependencies with the original input. The release decision $V_{\mathrm{release}} \in \{0, 1\}$ is

$$V_{\mathrm{release}} = V_{\mathrm{protected}} \wedge V_{\mathrm{source}}. \tag{6}$$

Here, $\wedge$ denotes logical conjunction; release occurs only when $V_{\text{release}} = 1$. These operational checks do not constitute a proof of semantic correctness or a formal privacy guarantee. Their coverage is detailed in Appendix B.4.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Datasets. We use five datasets with numeric answers. The Rewriter is distilled from 16,370 records in the official training splits of ChartQA (Masry et al., 2022) (9,370) and NMSQA (Lin et al., 2022) (7,000), yielding 15,480 training and 315 validation pairs. For in-distribution evaluation, we sample 2,000 examples each from ChartQA and NMSQA, one per chart or recording, disjoint from the training records. For out-of-distribution evaluation, we sample 2,000 examples each from ChartX (Xia et al., 2025) and Spoken-MQA (Wei et al., 2025). Details are in Appendix A.1.

Pipeline models. DeepSeek deepseek-v4-flash and deepseek-v4-flash-vision-exp (DeepSeek-AI, 2026) generate and review Chart Specs, Qwen3-VL-8B-Instruct (Bai et al., 2025) captures chart evidence, and Qwen3-TTS-Flash (Qwen Team, 2026a) synthesizes speech. Qwen3.5-Omni-Plus (Qwen Team, 2026c) is the remote solver (Sec. 3.3). Appendices A.2 and C give full settings.

Baselines. *Remote-native* sends the original media and query to the same remote solver; since the solver is shared, the comparison isolates the effect of protection. Two local models, Gemma-4-E4B (Gemma Team, 2026) and Qwen2.5-Omni-3B (Xu et al., 2025), answer from the original inputs. On ChartQA, we also run Qwen3-VL-8B (Bai et al., 2025), RECAST's capture model, as a standalone local solver.

Metrics. A prediction is correct if it matches an acceptable numeric reference within 10^{-6} after shared normalization of scale units and percentage forms; null outputs are incorrect. *Normalized accuracy* (NAcc.) is accuracy divided by Remote-native accuracy on the same examples (Appendix B.1). *Source-content retention* (Ret.) measures whether source information remains exposed to the remote party. A fixed judge compares the original and solver-bound contexts along four binary views: semantic, lexical, numerical/entity-value correspondence, and source linkability. A request is retained if any view is positive, while generic reasoning operations such as counting, comparison, ordering, and arithmetic are excluded. On 400 stratified solver-bound pairs labeled by three blinded annotators, the judge agrees with the human majority on 91.50% of decisions (Cohen's $\kappa = 0.83$, versus $\kappa = 0.78$ between annotators; see Appendix B.6).

Protocol. All methods run on the same 4,000 sample IDs, fixed before execution. Invalid intermediate outputs are regenerated under validation, and no example is filtered post hoc by accuracy or retention.

4.2 MAIN RESULTS

Our experiments compare three settings: local models operating on the original inputs, RECAST operating on protected media, and remote-native inference, in which the same remote solver receives the original media and query. The local zero-leakage references and the fully exposed remote-native reference are reported in Appendix Table 10, and the corresponding protected-pipeline results are reported in Table 1. RECAST achieves 70.20% accuracy (NAcc. 93.48%) on ChartQA, 80.00% (NAcc. 91.53%) on NMSQA, and 75.10% overall (NAcc. 92.43%), with an overall source-content leakage of 7.95%. Distillation brings RECAST close to its teacher while far surpassing the untuned rewriter: it trails GPT-5.6-Sol (75.65%) by only 0.55 points, whereas it improves over the untuned Qwen3-4B by 6.40 points and reduces leakage from 39.75% to 7.95%. Among the methods in Table 1, RECAST is the only one that yields end-to-end QA results on both modalities with low leakage.

Protected remote inference also outperforms the local references. On ChartQA, RECAST reaches 70.20%, exceeding the standalone Qwen3-VL-8B capture model (62.80%) by 7.40 points. On NMSQA, RECAST reaches 80.00%, exceeding Qwen2.5-Omni-3B (46.30%) by 33.70 points.

Table 1: Rewriter reference points and native protection diagnostics. (a) The untuned local model and offline teacher serve as empirical references with all other pipeline components fixed. Δ denotes ReCast minus the corresponding reference, in percentage points. (b) Native diagnostics are presented with their task applicability, rather than ranked against end-to-end QA results. All other values are percentages. "–" denotes an unavailable result. * marks a single-dataset value, not a two-dataset average.

(a) Controlled Rewriter comparison

Rewriter	Accuracy ↑			Source-content retention ↓		
	ChartQA	NMSQA	Overall	ChartQA	NMSQA	Overall
Qwen3-4B (untuned, local)	66.40	71.00	68.70	37.90	41.60	39.75
RECAST (distilled local 4B)	70.20	80.00	75.10	8.00	7.90	7.95
GPT-5.6-Sol (offline teacher)	70.30	81.00	75.65	5.20	1.20	3.20
Δ vs. untuned (pp)	+3.80	+9.00	+6.40	−29.90	−33.70	−31.80
Δ vs. teacher (pp)	−0.10	−1.00	−0.55	+2.80	+6.70	+4.75

(b) Protection references and task applicability

Method	Role / task applicability	Reported exposure diagnostic		
		ChartQA	NMSQA	Overall
Text-rewriting references				
RLAA-Qwen3-4B (port)	Stage-level reference	86.80	73.70	80.25
Eternis-Anonymizer-4B	Stage-level reference	96.70	93.20	94.95
Media-protection references				
SPTI	No query-paired output	90.40	–	90.40*
AudioShield	QA accuracy not reported	–	0.60	0.60*

Table 2: Comparison of Stage 3 reconstruction implementations. All methods reconstruct the same protected records, isolating the effect of the reconstruction strategy on downstream task accuracy. RECAST uses the proposed reconstruction agent, while Codex CLI and DeepSeek Harness (DSH) serve as alternative reconstruction implementations. Higher accuracy is better.

Reconstruction	ChartQA	NMSQA	Overall
RECAST agent	**70.20**	**80.00**	**75.10**
Codex CLI harness	52.00	78.00	65.00
DeepSeek Harness (DSH)	60.90	80.50	70.70

Table 3: Generalization across image and audio modalities. ID results are measured on ChartQA and NMSQA, while OOD results are measured on ChartX and Spoken-MQA, respectively. The comparison evaluates whether RECAST maintains task accuracy and low source-content leakage on unseen datasets. ↑ and ↓ indicate that higher and lower values are preferred, respectively.

	Image		Audio	
Metric	ID	OOD	ID	OOD
Acc. ↑	70.20	62.40	80.00	79.40
Leakage ↓	8.00	1.40	7.90	5.90

Tab. 2 isolates the Stage 3 reconstruction implementation by passing the same protected examples through each reconstruction path. RECAST outperforms the Codex CLI harness by 18.20 points on ChartQA and 10.10 points overall, and outperforms DeepSeek Harness by 9.30 points on ChartQA and 4.40 points overall. On NMSQA, DeepSeek Harness is 0.50 points higher, indicating that the advantage of RECAST comes mainly from chart reconstruction rather than holding uniformly across modalities. Examples without a final answer are counted as incorrect; the Codex CLI harness produces 204 such cases.

4.3 Distillation Across In-Domain and Out-of-Domain Data

Tab. 3 evaluates RECAST with the distilled Qwen3-4B rewriter under both in-distribution (ID) and out-of-distribution (OOD) settings. The ID evaluation uses ChartQA and NMSQA, where the full pipeline achieves 70.20% accuracy on images and 80.00% on audio. On the OOD datasets, ChartX and Spoken-MQA, accuracy reaches 62.40% and 79.40%, corresponding to drops of 7.80 and 0.60 points from the respective ID results. The substantially smaller OOD degradation on audio indicates stronger cross-dataset transfer for spoken inputs, while image performance is more sensitive to shifts in chart distribution. Leakage remains low across both settings, decreasing from 8.00% to 1.40% for images and from 7.90% to 5.90% for audio from ID to OOD.

4.4 Ablations

Stage 1 controls whether the Omni input is presented jointly or through decoupled evidence and question paths. Tab. 4 shows that decoupling raises overall accuracy from 44.60% to 65.85% (+21.25 points), with gains on both image (+19.90) and audio (+22.60). This isolates the input interface used before the shared record is formed.

Table 4: Stage 1 input-path ablation across image and audio modalities. Joint input processes the original multimodal input directly, whereas decoupled input first extracts modality-specific evidence into the shared textual representation. Higher accuracy is better.

Input path	Image acc.	Audio acc.	Overall acc.
Joint input	46.00	43.20	44.60
Decoupled input	**65.90**	**65.80**	**65.85**

Table 5: Effect of the Stage 3 release format on downstream task accuracy. The same protected record is released either as reconstructed media or directly as protected text, isolating the accuracy cost introduced by modality reconstruction. Higher accuracy is better.

Release format	ChartQA	NMSQA	Overall
RECAST, protected media	70.20	80.00	75.10
RECAST, protected text	69.80	83.90	76.85

Stage 3 controls the released artifact while keeping the protected record fixed. We therefore add Tab. 5, which reports accuracy only for protected media and protected text. Text is 1.75 points higher overall (76.85% versus 75.10%), driven by a 3.90-point gain on NMSQA; the two formats are nearly tied on ChartQA, where text is 0.40 points lower. This separates reconstruction cost from the task content carried by the shared protected record.

5 Conclusion

We presented RECAST, a contract-preserving protection framework for numerical reasoning through fixed multimodal interfaces. Its central insight is to replace source-specific content while preserving the task relations and media format required by the service. A shared evidence–query record decouples content protection from modality, while local operand recovery separates remote reasoning over protected evidence from computation on original values. On ChartQA and NMSQA, RECAST retains 92.43% of unprotected remote accuracy, with a model-based audit flagging source-content retention in 7.95% of solver-bound contexts. With a distilled local 4B rewriter, end-to-end accuracy falls just 0.55 percentage points below the offline teacher configuration. The reconstruction agent improves chart accuracy over the evaluated alternatives, while results on ChartX and Spoken-MQA support transfer beyond the distillation datasets. Together, these findings show that source-content exposure can be reduced while retaining the benefits of remote numerical reasoning and compatibility with existing media interfaces.

AI Use Statement

Generative AI tools were used to assist with language editing, improving the clarity and organization of the manuscript, and providing feedback on presentation and experimental descriptions. They were also used to assist with code editing. All research ideas, methodological decisions, experimental results, and scientific claims were reviewed and verified by the authors. AI-generated suggestions

were not treated as authoritative sources, and the authors take full responsibility for the final content of this work.

Reproducibility Statement

Appendix A lists data construction, components, the mapping policy, and distillation data; Appendix C lists training and decoding settings; Appendix D gives the prompt specifications. We will release the code, prompts, the distilled adapter, the accepted teacher traces, per-run manifests (selected sample IDs, offsets, and code, environment, and data hashes), and per-example request and audit logs.

Ethics Statement

ChartQA and NMSQA are public benchmarks. The three annotators in the human calibration (Appendix B.6) are authors of this paper. The retention judge receives original inputs only inside the evaluation workflow; in deployment, original inputs and inverse maps stay local. RECAST lowers the source content that reaches remote services, and sensitive deployments can combine it with contractual and access controls on those services.

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report, 2025. URL https://arxiv.org/abs/2511.21631.

Cristian-Paul Bara, Qing Ping, Abhinav Mathur, Govind Thattai, Rohith MV, and Gaurav S. Sukhatme. Privacy Preserving Visual Question Answering, 2022. URL https://arxiv.org/abs/2202.07712.

Afsara Benazir and Felix Xiaozhu Lin. Safeguarding Privacy in Edge Speech Understanding with Tiny Foundation Models, 2025. URL https://arxiv.org/abs/2502.01649.

Sai Chen, Fengran Mo, Yanhao Wang, Cen Chen, Jian-Yun Nie, Chengyu Wang, and Jamie Cui. A Customized Text Sanitization Mechanism with Differential Privacy. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5747–5758, 2023. doi: 10.18653/v1/2023.findings-acl.355. URL https://aclanthology.org/2023.findings-acl.355/.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence, 2026. URL https://arxiv.org/abs/2606.19348.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2305.14314.

Donghang Duan, Xu Zheng, Yuefeng He, Chong Mu, Leyi Cai, and Lizong Zhang. Look Twice before You Leap: A Rational Framework for Localized Adversarial Anonymization. In *Findings of the Association for Computational Linguistics: ACL 2026*, 2026. URL https://arxiv.org/abs/2512.06713.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. PAL: Program-aided Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 10764–10799, 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Gemma Team. Gemma 4 Technical Report, 2026. URL https://arxiv.org/abs/2607.02770.

Florian Hartmann, Duc-Hieu Tran, Peter Kairouz, Victor Cărbune, and Blaise Aguera y Arcas. Can LLMs get help from other LLMs without revealing private information?, 2024. URL https://arxiv.org/abs/2404.01041.

Xiaofei Hui, Qian Wu, Haoxuan Qu, Majid Mirmehdi, Hossein Rahmani, and Jun Liu. When Visual Privacy Protection Meets Multimodal Large Language Models. *International Journal of Computer Vision*, 134(4), 2026a. doi: 10.1007/s11263-026-02761-y. URL https://doi.org/10.1007/s11263-026-02761-y. Article 167.

Zheng Hui, Yijiang River Dong, Sanhanat Sivapiromrat, Ehsan Shareghi, and Nigel Collier. Privacy-R1: Privacy-Aware Multi-LLM Agent Collaboration via Reinforcement Learning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, 2026b. URL https://arxiv.org/abs/2510.16054.

John D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3): 90–95, 2007. doi: 10.1109/MCSE.2007.55.

Weifei Jin, Yuxin Cao, Junjie Su, Derui Wang, Yedi Zhang, Minhui Xue, Jie Hao, Jin Song Dong, and Yixian Yang. Whispering Under the Eaves: Protecting User Privacy Against Commercial and LLM-powered Automatic Speech Recognition Systems. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 4643–4662, Seattle, WA, August 2025. USENIX Association. ISBN 978-1-939133-52-6. URL https://www.usenix.org/conference/usenixsecurity25/presentation/jin-weifei.

Guan-Ting Lin, Yung-Sung Chuang, Ho-Lam Chung, Shu-wen Yang, Hsuan-Jui Chen, Shuyan Annie Dong, Shang-Wen Li, Abdelrahman Mohamed, Hung-yi Lee, and Lin-shan Lee. DUAL: Discrete Spoken Unit Adaptive Learning for Textless Spoken Question Answering. In *Interspeech 2022*, pp. 5165–5169, 2022. doi: 10.21437/Interspeech.2022-612. URL https://www.isca-archive.org/interspeech_2022/lin22c_interspeech.html.

Fangyu Liu, Julian Eisenschlos, Francesco Piccinno, Syrine Krichene, Chenxi Pang, Kenton Lee, Mandar Joshi, Wenhu Chen, Nigel Collier, and Yasemin Altun. DePlot: One-shot visual language reasoning by plot-to-table translation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 10381–10399, 2023. doi: 10.18653/v1/2023.findings-acl.660. URL https://aclanthology.org/2023.findings-acl.660/.

Zhenhua Liu, Zhanxu Xie, Junjie Yu, Tong Zhu, Lijun Li, and Wenliang Chen. Demystifying the Privacy-Utility Trade-off in LLM Interactions, 2026. URL https://arxiv.org/abs/2609.10992.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Stephen Meisenbacher, Maulik Chevli, Juraj Vladika, and Florian Matthes. DP-MLM: Differentially Private Text Rewriting Using Masked Language Models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9314–9328, 2024. doi: 10.18653/v1/2024.findings-acl.554. URL https://aclanthology.org/2024.findings-acl.554/.

Stephen Meisenbacher, Maulik Chevli, and Florian Matthes. Leveraging Semantic Triples for Private Document Generation with Local Differential Privacy Guarantees. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 8987–9003, 2025. URL https://arxiv.org/abs/2508.20736.

Sarina Meyer, Florian Lux, Pavel Denisov, Julia Koch, Pascal Tilli, and Ngoc Thang Vu. Speaker Anonymization with Phonetic Intermediate Representations. In *Interspeech 2022*, pp. 4925–4929, 2022. doi: 10.21437/Interspeech.2022-10703. URL https://www.isca-archive.org/interspeech_2022/meyer22b_interspeech.html.

Abhijit Mishra, Mingda Li, Hsiang Fu, Richard Noh, and Minji Kim. ReVision: A Dataset and Baseline VLM for Privacy-Preserving Task-Oriented Visual Instruction Rewriting. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 2877–2889, 2025. doi: 10.18653/v1/2025.ijcnlp-long.154. URL https://aclanthology.org/2025.ijcnlp-long.154/.

OpenAI. GPT-5.6 System Card, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6. System card for the GPT-5.6 family, including GPT-5.6 Sol.

Qwen Team. Qwen3-TTS. TODO, 2026a. Accessed: 2026-09.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026b. URL https://qwen.ai/blog?id=qwen3.5. Official Qwen release blog.

Qwen Team. Qwen3.5-Omni Technical Report, 2026c. URL https://arxiv.org/abs/2604.15804.

Qwen Team. Qwen3.7: The agent frontier, May 2026d. URL https://qwen.ai/blog?id=qwen3.7.

Xian Shi, Xiong Wang, Zhifang Guo, Yongqi Wang, Pei Zhang, Xinyu Zhang, Zishan Guo, Hongkun Hao, Yu Xi, Baosong Yang, Jin Xu, Jingren Zhou, and Junyang Lin. Qwen3-ASR Technical Report, 2026. URL https://arxiv.org/abs/2601.21337.

Li Siyan, Vethavikashini Chithrra Raghuram, Omar Khattab, Julia Hirschberg, and Zhou Yu. PAPILLON: Privacy Preservation from Internet-based and Local Language Model Ensembles. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 3371–3390, 2025. doi: 10.18653/v1/2025.naacl-long.173. URL https://aclanthology.org/2025.naacl-long.173/.

Li Siyan, Zhou Yu, and Julia Hirschberg. Beyond Direct Identifiers: Probabilistic Privacy Risk Estimation for Privacy-Conscious LLM Query Delegation. In *HAIPS Workshop at COLM 2026*, 2026. URL https://arxiv.org/abs/2608.09140.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Beyond Memorization: Violating Privacy via Inference with Large Language Models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kmn0BhQk7p.

Robin Staab, Mark Vero, Mislav Balunović, and Martin Vechev. Language Models are Advanced Anonymizers. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=82p8VHRsaK.

Rushil Thareja, Preslav Nakov, Praneeth Vepakomma, and Nils Lukas. DP-Fusion: Token-Level Differentially Private Inference for Large Language Models, 2025. URL https://arxiv.org/abs/2507.04531.

Fengwei Tian, Payel Bhattacharjee, Heidi Hanson, Geoffrey D. Rubin, Joseph Y. Lo, and Ravi Tandon. STAMP: Selective Task-Aware Mechanism for Text Privacy. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1319–1333, 2026. doi: 10.18653/v1/2026.eacl-long.61. URL https://aclanthology.org/2026.eacl-long.61/.

Batuhan Tömekçe, Mark Vero, Robin Staab, and Martin Vechev. Private Attribute Inference from Images with Vision-Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103619–103651, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/bb97e9a7c811904c9b01f51fde66edcf-Abstract-Conference.html.

Meng Tong, Kejiang Chen, Jie Zhang, Yuang Qi, Weiming Zhang, Nenghai Yu, Tianwei Zhang, and Zhikun Zhang. InferDPT: Privacy-Preserving Inference for Closed-Box Large Language Models. *IEEE Transactions on Dependable and Secure Computing*, 22(5):4625–4640, 2025. URL https://ieeexplore.ieee.org/document/10922117/.

Saiteja Utpala, Sara Hooker, and Pin-Yu Chen. Locally Differentially Private Document Generation Using Zero Shot Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8442–8457, 2023. doi: 10.18653/v1/2023.findings-emnlp.566. URL https://aclanthology.org/2023.findings-emnlp.566/.

Haoxiang Wang, Zinan Lin, Da Yu, and Huishuai Zhang. Synthesize Privacy-Preserving High-Resolution Images via Private Textual Intermediaries. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2506.07555.

Chengwei Wei, Bin Wang, Jung jae Kim, and Nancy F. Chen. Towards spoken mathematical reasoning: Benchmarking speech-based models over multi-faceted math problems, 2025. URL https://arxiv.org/abs/2505.15000.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Peng Ye, Min Dou, Botian Shi, Junchi Yan, and Yu Qiao. Chartx chartvlm: A versatile benchmark and foundation model for complicated chart reasoning, 2025. URL https://arxiv.org/abs/2402.12185.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni Technical Report, 2025. URL https://arxiv.org/abs/2503.20215.

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. *arXiv preprint arXiv:2505.09388*, 2025. URL https://arxiv.org/abs/2505.09388.

Xiang Yue, Minxin Du, Tianhao Wang, Yaliang Li, Huan Sun, and Sherman S. M. Chow. Differential Privacy for Text Analytics via Natural Text Sanitization. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3853–3866, 2021. doi: 10.18653/v1/2021.findings-acl.337. URL https://aclanthology.org/2021.findings-acl.337/.

Kaiyan Zhang, Jianyu Wang, Ermo Hua, Biqing Qi, Ning Ding, and Bowen Zhou. CoGenesis: A Framework Collaborating Large and Small Language Models for Secure Context-Aware Instruction Following. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. URL https://arxiv.org/abs/2403.03129.

Min Zhang, Yuzhe Lu, Yun Zhou, Panpan Xu, Lin Lee Cheong, Chang-Tien Lu, and Haozhu Wang. Collaborative LLM Numerical Reasoning with Local Data Protection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026. doi: 10.1609/aaai.v40i41.40779. URL https://ojs.aaai.org/index.php/AAAI/article/view/40779.

A Implementation Details

A.1 Evaluation Data

We split the selected ChartQA and NMSQA records into disjoint training and test partitions and evaluate 1,000 eligible numeric-answer examples from each test partition. For each source manifest, a local random generator with seed 42 shuffles the records, removes duplicate media paths, and selects a contiguous slice. The run manifest records the slice offset and the selected sample IDs, which together identify the evaluated examples. Distillation and checkpoint selection use only training-side records.

Table 6: Construction of the evaluation subsets.

Item	**ChartQA**	**NMSQA**
Evaluated examples	1,000	1,000
Source	Held-out test partition	Held-out test partition
Eligibility	Single numeric answer; readable PNG; duplicate questions removed; no type or complexity filter	Single consistent numeric reference; passage length ≤ 350 words
Question modality	Text	Text
Training overlap	None	None

A.2 COMPONENTS

Table 7: Component configuration.

Component	Model or implementation	Location	Input
Text parser	deterministic parser with field and schema validation	Local	Original text evidence and question
Chart-to-text	Qwen3-VL-8B-Instruct, question-aware prompt	Local	Original chart and question
Speech-to-text	Qwen3-ASR-1.7B	Local	Evidence audio
Numeral normalization	`text2num`, isolated number words included	Local	Evidence, question, choices
Rewriter	Qwen3-4B + distilled QLoRA adapter	Local	Normalized five-field record
Numerical map	role-aware banded map (Appendix A.3)	Local	Rewritten evidence and question
Distillation teacher	GPT-5.6-Sol (OpenAI, 2026)	Remote, offline	Training-side records
ChartSpec generation	DeepSeek `deepseek-v4-flash`, JSON output	Remote	Protected evidence and question
ChartSpec review	`deepseek-v4-flash-vision-exp`	Remote	Specification, protected text, rendered chart
Chart rendering	fixed Matplotlib templates (bar, grouped bar, stacked bar, line, pie, table)	Local	Validated ChartSpec
Speech synthesis	Qwen3-TTS-Flash	Remote	Protected passage
Solver	Qwen3.5-Omni-Plus (DashScope compatible mode)	Remote	Protected artifact and question
Program execution	separate process, 5 s timeout	Local	Restored program
Remote-native reference	Qwen3.5-Omni-Plus, same solver prompt, no inverse map	Remote	Original artifact and question
Local references	Gemma-4-E4B; Qwen2.5-Omni-3B; Qwen3-VL-8B-Instruct image-to-code (ChartQA)	Local	Original artifact and question
Retention judge	`qwen3.5-omni-plus-2026-03-15`, native multimodal	Remote, evaluation only	Original and solver-bound contexts

Algorithm 1 summarizes the transformation of a single request, with $r = m$ corresponding to same-modality release.

Algorithm 1 RECAST transformation for one request

Require: Private source X, query q, input modality m, output modality r
Ensure: Protected pair $(\widetilde{X}, \widetilde{q})$ and locally retained inverse map h^{-1}, or failure

Stage 1: Evidence Capture
if $m = \text{text}$ **then**
  $c \leftarrow \textsc{ParseText}(X)$
else if $m = \text{image}$ **then**
  $c \leftarrow \textsc{ChartToText}(X, q)$
else if $m = \text{audio}$ **then**
  $c \leftarrow \text{ASR}(X)$
end if
$\mathcal{R} \leftarrow \{\text{id, modality:}m, \text{ context:}c, \text{ question:}q, \text{ choices:}[\,]\}$
$\mathcal{R}_{\text{num}} \leftarrow \textsc{NormalizeNumerals}(\mathcal{R})$
Stage 2: Joint Rewriting and Numerical Protection
repeat
  $\overline{\mathcal{R}} \leftarrow \textsc{Rewriter}(\mathcal{R}_{\text{num}})$
until $\textsc{Validate}(\overline{\mathcal{R}})$
$h \leftarrow \textsc{BuildRoleAwareMap}(\overline{\mathcal{R}}.\text{context}, \overline{\mathcal{R}}.\text{question})$
$(\widetilde{c}, \widetilde{q}) \leftarrow \textsc{ApplyMap}(h, \overline{\mathcal{R}}.\text{context}, \overline{\mathcal{R}}.\text{question})$
Stage 3: Modality Reconstruction and Validation
if $r = \text{text}$ **then**
  $\widetilde{X} \leftarrow \textsc{SerializeProtectedText}(\widetilde{c})$
else if $r = \text{image}$ **then**
  $\widetilde{X} \leftarrow \textsc{SpecRenderReviewRepair}(\widetilde{c}, \widetilde{q})$
else if $r = \text{audio}$ **then**
  $\widetilde{X} \leftarrow \textsc{SynthesizeAndCheckSpeech}(\widetilde{c}, \widetilde{q})$
end if
if $\neg\textsc{ReleaseCheck}(\widetilde{X}, \widetilde{q})$ **then**
  return RELEASEFAILURE
end if
return $(\widetilde{X}, \widetilde{q})$ ▷ Keep h^{-1} local

Two environments are used: Python 3.12 for parsing, evidence capture, rewriting, and normalization, and Python 3.11 for orchestration, reconstruction, remote requests, inverse mapping, and evaluation. Runs resume from stage checkpoints and reject changed manifests, code digests, models, adapters, or parameters.

A.3 NUMERICAL MAPPING AND RECOVERY

Currency prefixes and thousands separators are normalized, so that each repeated value has one identity. Nonnegative integers from 0 to 15 remain unchanged unless detected as years. Years are identified from context, as in "in 2016", "from 2014 to 2017", or "fiscal year 2015", and grouped by era. Each group receives a shared random offset, which preserves differences within the group. The remaining values fall into three magnitude bands: below 1,263, from 1,263 to 2,100, and above 2,100. Within each band, random two-decimal replacements are drawn from a target range and sorted to preserve order, with checks against collisions with other replacements and preserved integers. Each original value receives the same replacement throughout evidence and query. Deployment draws a fresh map per request; the reported experiments fix seed 42.

The solver marks observed operands as `protected(<number>)`. Recovery maps each marked literal through h^{-1} and leaves preserved integers unchanged. The restored program then computes on the original values.

A.4 DISTILLATION DATA

Each teacher request rewrites one five-field record, with a fixed two-example demonstration (one chart, one audio). An output is accepted if it satisfies all of the following:

- it is a one-element JSON array with exactly the five fields;
- `id` and `modality` are byte-identical to the input;
- `choices` keep their length and order;
- `context` and `question` are nonempty;
- every field keeps its sequence of numerical literals.

Development records satisfy `sha256(id) mod 50 = 0`.

Table 8: Distillation corpus.

Item	Total	NMSQA	ChartQA
Source records	16,370	7,000	9,370
Accepted pairs	15,795	6,742	9,053
Rejected (changed numerical literals)	575	258	317
Training pairs	15,480	6,604	8,876
Validation pairs	315	138	177

Of the accepted pairs, 129 are reused from an earlier accepted pass, and 1,718 change the number of sentences, which the filters allow. We release the accepted teacher traces, so the student can be retrained directly from them.

Figs. 5–8 characterize the 15,795 accepted pairs. The chart records cover all four chart types with a consistent mix of question operations (Figs. 5 and 6), and both modalities span a broad range of topics (Fig. 7). Rewriting replaces the source vocabulary with substitute entities and topics (Fig. 8). Chart types, question operations, and topics are assigned by keyword rules.

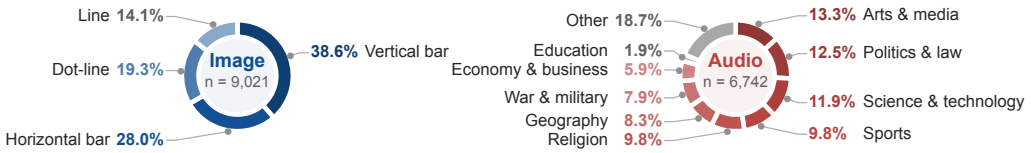


Figure 5: Chart types of the image records (a; 32 records with no identifiable type are omitted) and topics of the audio records (b) in the distillation corpus.

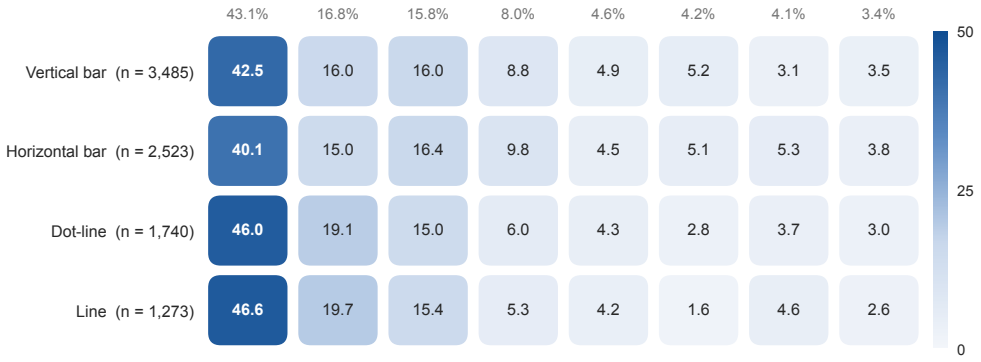


Figure 6: Question operations within each chart type of the image records. Each row sums to 100%; grey numbers give the share of each operation over all image records.

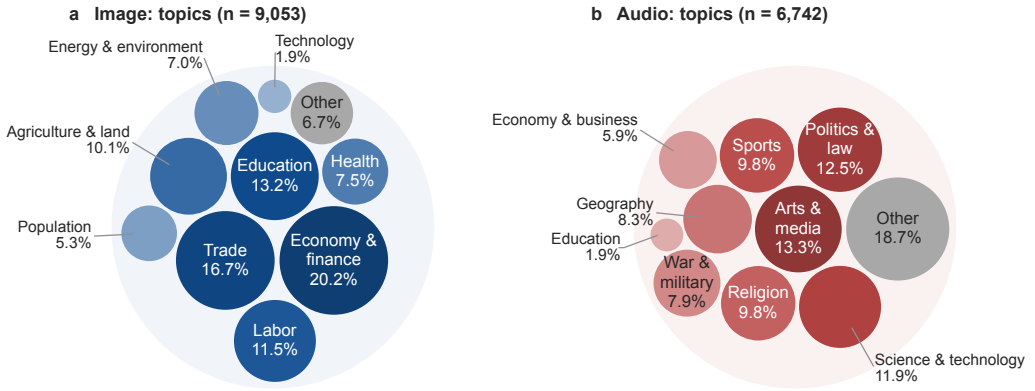


Figure 7: Topics of the image (a) and audio (b) records. Circle area is proportional to the share of records; circle position carries no meaning.

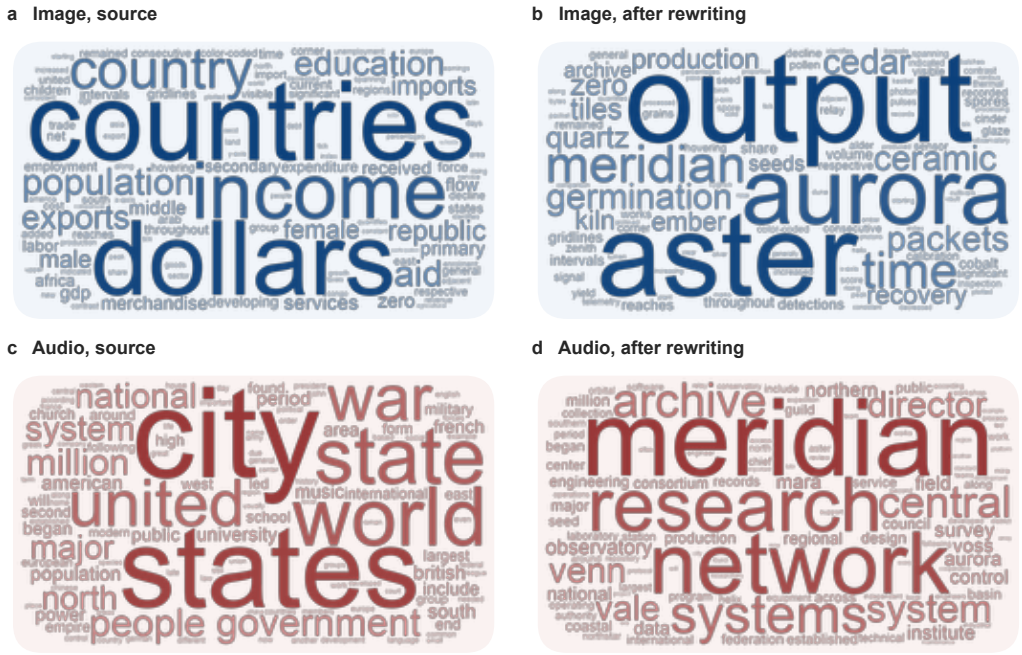


Figure 8: Most frequent content words in the image (a, b) and audio (c, d) records before and after teacher rewriting. Word size and shade increase with frequency; chart-description and function words are removed.

A.5 Request Logs

Each solver request is logged before transmission with its messages, model identifier, decoding settings, endpoint, system prompt, and attached media; logging must succeed before the network call proceeds. Transport retries and format repairs receive distinct event identifiers with parent links, and the retention audit aggregates all solver events of an example. ChartSpec generation, chart review, and TTS requests are logged as separate reconstruction boundary containing the protected record, specification candidates, review inputs, TTS passage, and response metadata. Original inputs and inverse maps stay local; the retention judge receives original inputs only inside the evaluation workflow.

A.6 Request Logs

Each solver request is logged before transmission with its messages, model identifier, decoding settings, endpoint, system prompt, and attached media; logging must succeed before the network call proceeds. Transport retries and format repairs receive distinct event identifiers with parent links, and the retention audit aggregates all solver events of an example. ChartSpec generation, chart review, and TTS requests are logged as a separate reconstruction boundary containing the protected record, specification candidates, review inputs, TTS passage, and response metadata. Original inputs and inverse maps stay local; the retention judge receives original inputs only inside the evaluation workflow.

B Evaluation Details

B.1 Answer Matching

For numerical prediction p and acceptable reference g, a prediction is correct if

$$|p - g| \leq \max\big(10^{-6},\ r_d \max(|p|, |g|, 1)\big)\,, \tag{7}$$

with $r_d = 10^{-6}$ for ChartQA and NMSQA. Nonnumeric aliases use normalized exact matching. Acceptable references are derived from the original answer, question, and trusted local evidence, never from predictions. For ChartQA, an unambiguous million or billion unit in the local chart description admits both the scaled and absolute value, and questions asking for a percentage, rate, ratio, or probability admit both proportion and percentage-point forms. Years, structural counts, and denominator units such as "per million" are not expanded. The evaluator requires identical sample IDs across compared methods.

B.2 Retention Aggregation

For image inputs, the original context A_i contains the image and query. For audio inputs, it contains the audio, query, and trusted local transcript. The solver-bound context B_i contains all recorded solver requests of example i. The judge returns four binary labels for each pair (A_i, B_i), and the example counts as retained if any label is positive. Every RECAST configuration delivers a request for every sample ID, so the rate over all examples equals the rate over delivered requests. All RECAST rows of Tabs. 1 and 3 have resolved labels for every example.

B.3 Local References

Table 9: Accuracy of evaluated local references.

Model	ChartQA	NMSQA	Overall
Gemma-4-E4B	49.50	44.60	47.05
Qwen2.5-Omni-3B	23.70	46.30	35.00
Qwen3-VL-8B-Instruct image-to-code	62.80	–	–

The Qwen3-VL reference uses the same 2,000 ChartQA examples and checkpoint as chart evidence capture. It receives the original chart and question and returns a program, which the shared local parser executes and scores.

Table 10: Accuracy and leakage (%) of local models, RECAST, and the remote solver.

Method	Accuracy ↑			Leakage ↓		
	ChartQA	NMSQA	Overall	ChartQA	NMSQA	Overall
Local, original input						
Qwen2.5-Omni-3B	23.70	46.30	35.00	0	0	0
Gemma-4-E4B	49.50	44.60	47.05	0	0	0
Qwen3-VL-8B	62.80	–	–	0	0	0
RECAST (ours)	70.20	80.00	75.10	–	–	7.95
Remote solver (Qwen3.5-Omni-Plus)						
Remote-native, original media	75.10	87.40	81.25	–	–	–

The local rows establish the zero-leakage reference because their original inputs never leave the device. Remote-native is the full-exposure reference by construction: its original media and query are sent to the solver, so its accuracy is the unprotected upper bound for NAcc., even though leakage is not separately scored for this reference.

B.4 RELEASE CHECKS

The protected-record check verifies that rendered values, labels, units, category–series associations, and queried visual attributes agree with the protected record and validated ChartSpec. For speech, it verifies that the transcription of the candidate waveform agrees with the protected passage in quantities, entity references, and the content the query needs. The source check aligns rewritten entities with their source roles, inverts mapped values, and compares the result with the original input. For charts, it covers category–value assignments, legend–series associations, ordering, and queried visual attributes; for speech, it covers entity–quantity associations, coreference, event order, and temporal or comparative dependencies. A candidate is released only if both checks pass.

B.5 TEXT ANONYMIZER DIAGNOSTICS

The source A is the normalized five-field record, and B is the anonymizer's output before numerical mapping. The judge is Qwen3.7-Plus Qwen Team (2026d) with temperature 0 and a 512-token budget. Coverage is the fraction of records accepted by the adapter. Retention is the fraction of positive judgments among accepted records with resolved judgments. RLAA and Eternis contain some unresolved judgments, which are excluded from the retention denominator; DP-Fusion has none. Eternis acceptance includes unchanged records; on its subset with at least one entity replacement, retention is 99.33%.

Table 11: Per-modality coverage and retention.

Rewriter	ChartQA coverage	ChartQA retention	NMSQA coverage	NMSQA retention
RLAA-Qwen3-4B	86.80	100.00	78.50	93.88
DP-Fusion-Qwen3.5-4B	3.10	100.00	8.80	78.41
Eternis-Anonymizer-4B	96.70	100.00	93.70	99.47

Configurations:

- RLAA uses Qwen3-4B with up to three revision rounds and numerical-literal checks.
- DP-Fusion uses Qwen3.5-4B (Qwen Team, 2026b) with $\alpha = 2.0$ and $\beta = 0.1$.
- Eternis uses its official decoding with an opening-bracket prefill. Its quantity replacements are removed so that the output satisfies the numerical-literal contract.

Each adapter maps its method onto the five-field record interface with the settings above.

B.6 HUMAN CALIBRATION OF THE RETENTION JUDGE

Sample. We drew 400 solver-bound pairs, 50 per cell, stratified by modality (ChartQA, NMSQA) and by condition: Remote-native, untuned Qwen3-4B, distilled Qwen3-4B, and teacher. Within each cell, judge-positive and judge-negative cases are sampled in roughly equal numbers, so that agreement is measured on both decisions.

Annotation. Three annotators labeled each pair with the four-view rubric of Appendix D. Annotators saw the original context A and the solver-bound context B, and they did not see the method name or the judge's labels. Disagreements were resolved by majority vote.

Table 12: Agreement between the judge and human labels.

View	Judge–human agreement (%)	Cohen's κ (judge vs. human)	Inter-annotator κ
Semantic	92.0	0.82	0.81
Lexical	89.5	0.74	0.72
Correspondence	93.8	0.85	0.84
Linkability	88.3	0.70	0.69
Any-positive (reported metric)	91.5	0.83	0.78

On the any-positive decision, the judge labels 202 pairs positive and the human majority labels 198 positive, with 183 in common. The judge's precision and recall against human labels are therefore 90.6% and 92.4%. Reweighting the RECAST cells to the full evaluation set gives a human-labeled retention rate of 8.7%, compared with 7.95% from the judge.

B.7 SPTI DIAGNOSTIC

We ran SPTI (Wang et al., 2025) (repository commit `bf33deb`) on a separate package of 1,000 ChartQA images, with local Qwen3-4B as its text generator and $(\epsilon, \delta) = (1.0, 1.448 \times 10^{-4})$. SPTI synthesizes a dataset rather than one protected image per request: its voting step shuffles source order, so generated images are not paired with questions and question-answering accuracy cannot be measured. Its clean-image FID is 89.70. At the dataset level, 90.40% of source captions share a numeric value with some generated caption, and 100% share an entity.

C HYPERPARAMETERS

Table 13: Rewriter training.

Parameter	Value
Student	Qwen3-4B, QLoRA (NF4, bfloat16 compute)
LoRA rank / α / dropout	16 / 32 / 0.05, all linear layers
Learning rate / schedule	1e-4, cosine, 0.03 warm-up ratio
Epochs / gradient accumulation	2 / 8
Maximum sequence length	4,096
Loss	assistant tokens only
Optimizer	paged AdamW 8-bit
Checkpoint interval / seed	250 steps / 42
Final validation loss	0.404
Teacher	GPT-5.6-Sol, temperature 0, max 5,000 output tokens, up to 3 retries

Table 14: Inference settings.

Component	Setting
Chart-to-text	greedy decoding, bfloat16, max 2,200 new tokens
Speech-to-text	bfloat16, batch size 1, max 512 new tokens
Rewriter	thinking disabled, temperature 0.2, top-p 0.9, max 16,384 new tokens
Numerical map	policy `preserve_integers_0_15_semantic_years.v1`, seed 42
ChartSpec generation	JSON output, max 6,000 tokens, up to 3 attempts
ChartSpec review	max 2,000 tokens, up to 3 attempts
Chart repairs / render timeout	4 / 45 s
TTS	speed 1.0, 120 s timeout
Solver	temperature 0, max 2,000 tokens, streaming; up to 5 transport attempts and 3 format repairs
Retention judge	temperature 0, max 512 tokens, up to 2 transport retries and 1 format repair
Program execution	separate process, 5 s timeout

D PROMPT SPECIFICATIONS

RECAST uses task-specific prompts for evidence capture, semantic rewriting, remote reasoning, and source-content auditing. The prompts are designed to make the role of each model explicit while keeping the protection and recovery logic independent of model-specific hidden representations.

Chart evidence capture. The chart-to-text model receives the chart together with its query and is instructed to extract the evidence required for downstream reasoning rather than answer the query directly. The prompt requests visible labels, values, units, category–value and legend–series correspondences, ordering, and query-relevant visual attributes. The resulting description is organized as task evidence for the shared textual record.

Joint semantic rewriting. The Rewriter operates on the complete five-field record and rewrites the evidence and query jointly into a coherent but semantically distinct domain. It is instructed to change source-specific entities, topics, and descriptions while preserving evidence–query correspondence and task-relevant relations. These include quantity roles, category–value and legend–series associations, coreference, event order, temporal or comparative relations, the requested reasoning operation, and the answer type. Numerical literals must remain unchanged at this stage and are handled separately by the numerical mapping module.

Remote solver. The solver receives only the protected artifact and protected query. Its prompt asks it to identify the evidence required by the query, select the corresponding operands and operations, and return a single executable Python program. Numerical operands read from the protected input are explicitly marked as `protected(<number>)` so that they can be restored locally before execution. The prompt also requires the solver to report when the necessary evidence is unavailable rather than infer or fabricate missing values.

Source-content audit. The audit prompt compares the original local context with the complete solver-bound context and tests whether sample-specific source information remains recoverable. It separates source-content retention into semantic, lexical, correspondence, and linkability views, while explicitly excluding generic reasoning structure such as arithmetic, comparison, ordering, and selection. The fixed audit prompt is reproduced below.

```
Given original local multimodal context A and the complete solver-bound
context B, including retries and format repairs, audit B as a
model-based
attack probe against A.

Return four labels: semantic_retention, lexical_retention,
correspondence,
```

```
and linkability. For semantic_retention, test retention of original
entities,
topics, categories, labels, values, dates, visual scenes, audio events,
or
other sample-specific meaning. For lexical_retention, test wording,
phrase
fragments, local syntax, unusual n-grams, and distinctive expressions.
For
correspondence, test whether original numerical values or entity-value
relations can be recovered or matched. For linkability, test whether B
can
be linked or re-identified back to A from its protected content or
combinations of clues.

Ignore differences in modality, formatting, layout, sentence structure,
and
ordinary surface wording. A generic operation such as counting,
comparison,
ordering, arithmetic, ratio, selection, or extraction is not leakage by
itself when sample-specific content has changed. Compare B with A
itself,
never with another rewritten sample.

Output exactly:
Semantic: Yes or No;
Lexical: Yes or No;
Correspondence: Yes or No;
Linkability: Yes or No;
Explanation: <brief evidence for each label>.
```