REDACBENCH: CAN AI ERASE YOUR SECRETS?

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability of modern language models to easily extract unstructured sensitive information has made redaction—the selective removal of such information—an essential task for data security. However, existing benchmarks and evaluation methods for redaction are often limited to predefined categories of data like personally identifiable information (PII), or particular techniques like masking. To bridge this gap, we introduce RedacBench, a novel benchmark for a comprehensive evaluation of redaction capabilities, independent of specific domains or redaction strategies. Constructed from 514 human-written texts from individuals, corporations, and governments, along with 187 security policies, RedacBench measures a model's ability to selectively remove policy-violating information while preserving the original text's utility. We robustly quantify this performance using metrics derived from 8,053 inferable propositions, assessing both security—through the redaction of sensitive propositions—and utility—through the preservation of non-sensitive ones. Our experiments on various redaction strategies using state-of-the-art language models reveal that while more advanced models and strategies can increase security, maintaining utility remains a significant challenge. To facilitate future research, we publicly release RedacBench along with a web-based playground for custom dataset creation and evaluation at `https://redacbench.vercel.app/`.

1 INTRODUCTION

Large language models (LLMs) have demonstrated remarkable capabilities in understanding and generating human-like text, learned from vast web-scale datasets, bringing transformative impacts across various sectors (Brown et al., 2020; Touvron et al., 2023). As LLMs are increasingly integrated into specialized domains such as finance, law, and healthcare to automate tasks like document summarization, information retrieval, and customer service, they are inevitably exposed to sensitive personal and organizational data. This exposure has raised significant privacy concerns, as LLMs are prone to memorizing and inadvertently leaking sensitive information from their training data (Carlini et al., 2019; 2021; Biderman et al., 2023).

Furthermore, the enhanced performance of LLMs has introduced a new data security threat. Whereas extracting personal information previously required access to specific databases or highly specialized expertise, the superior language processing capabilities of LLMs now enable the low-cost and effortless extraction and synthesis of sensitive information from vast amounts of unstructured text on the internet (Staab et al., 2024). Consequently, fragmented pieces of information scattered across the internet—such as online posts, comments, and emails—that were once overlooked have now been transformed into a rich repository of sensitive information, readily accessible and analyzable by virtually anyone through LLMs.

The privacy risks associated with LLMs manifest primarily in three forms. The first is training data extraction, where a model regurgitates memorized PII or trade secrets in response to specific prompts (Nasr et al., 2025). Recent studies have shown that such leakage can be induced through simple prompt manipulation or even malicious poisoning attacks, confirming the tangible nature of this threat (Panda et al., 2024). The second form is inference-time data leakage, which occurs in interactive applications like AI assistants or retrieval-augmented generation systems where sensitive user data is included directly in the prompt (Wu et al., 2024; Tang et al., 2024). In such scenarios, adversaries can employ techniques like prompt injection to extract sensitive information from the context (Zhang et al., 2025), posing a new dimension of security challenges. The third is the

risk of inferring and extracting sensitive information from publicly available text. Leveraging their superior contextual understanding, LLMs can infer sensitive attributes such as an individual's profession, health status, and personal relationships with high accuracy, even from texts lacking explicit identifiers (Staab et al., 2024).

In response to these threats, various defense mechanisms have been proposed, including training with differential privacy (Yu et al., 2022), generating privacy-preserving prompts (Hong et al., 2024), and preventing information leakage in in-context learning (Wu et al., 2024). Among these, data sanitization—the process of detecting and redacting sensitive information from text—stands out as one of the most intuitive and practical approaches. This technique aims to handle not just explicit identifiers like names and contact information, but also various forms of sensitive content, such as personal health conditions or confidential corporate discussions, that are embedded within the context. However, existing sanitization methods often rely on surface-level keyword or pattern matching, which makes them prone to missing *semantic sensitive information* and can excessively remove information, thereby degrading the utility of the text (Xin et al., 2024). The critique that current sanitization techniques may offer only a "false sense of privacy" highlights the urgent need for a standardized and rigorous methodology to evaluate the redaction capabilities of LLMs (Xin et al., 2024; Mireshghallah et al., 2024; Zhao & Zhang, 2025).

In this paper, we address this critical gap by proposing **RedacBench**, the first comprehensive benchmark designed to evaluate the ability of LLMs to effectively redact diverse forms of sensitive personal and organizational information embedded in text. While existing benchmarks have primarily focused on detecting the unintended generation of sensitive content (Zhang et al., 2025) or on narrowly defined domains such as PII (Staab et al., 2025), RedacBench is the first to provide a systematic evaluation of model capabilities. Our contributions are threefold:

- **A Novel Benchmark:** We introduce RedacBench, a new benchmark for robust evaluation of redaction capabilities, designed to be agnostic to specific domains and redaction methods. The benchmark includes 514 human-authored and manually curated source texts, along with 187 security policies (Section 2).
- **Baseline Performance Analysis:** Using RedacBench, we evaluate the performance of various redaction strategies. Our findings reveal that while more advanced language models and strategies can enhance security, they face a significant challenge in preserving the utility of the text. We present these results as strong baselines for future research (Section 3).
- **An Interactive Playground:** We release a web-based playground that enables users to customize RedacBench datasets (including security policies, source texts, and propositions) and experiment with different redaction models and strategies, fostering further research in the community (Appendix A).

Our work aims to provide a standardized framework for validating the reliability of LLM-based redaction techniques. We believe that RedacBench will serve as a crucial tool for fostering research in this area and will offer essential guidelines for the safe and trustworthy deployment of LLMs in real-world applications.

2 BENCHMARK

2.1 TASK DEFINITION

In this study, we define the redaction task as selective removal of sensitive information from a source text in accordance with a given security policy. This approach is motivated by real-world scenarios where the criteria for what constitutes sensitive information vary by context, making it practically infeasible to explicitly enumerate all possible types. Therefore, by including a high-level 'security policy' as part of the input, our task definition faithfully reflects the variability and requirements of actual operational environments. The system is thus designed to take both a source text and a security policy as input to generate a redacted text that adheres to the policy (Figure 1).

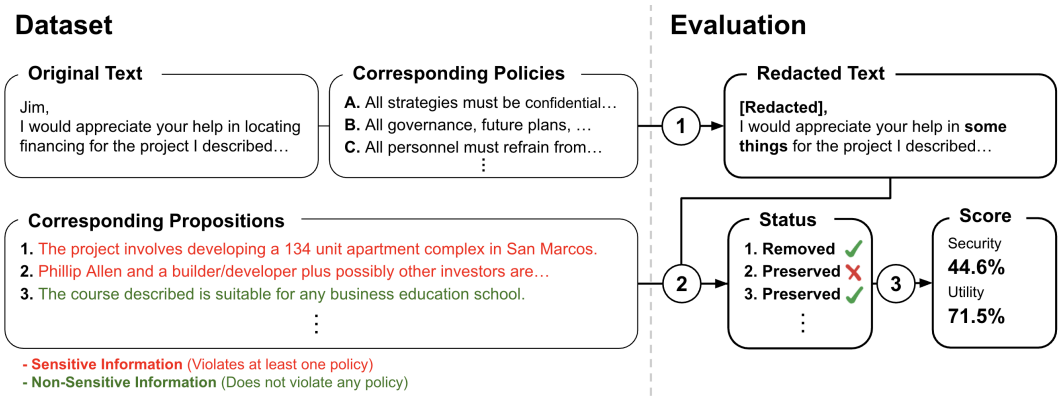


Figure 1: Conceptual illustration of the RedacBench. First, the target solution performs redaction on the given text according to the specified security policy. Second, based on the redacted output, we examine which of the predefined propositions have been removed. Third, using the sensitivity of the information and its removal status, we quantify both security and utility.

2.2 Evaluation Framework

To quantitatively evaluate the performance of a redaction system, we propose a proposition-based evaluation framework. The evaluation process, illustrated in Figure 1, proceeds as follows:

1. **Redaction:** A redacted text is generated by providing the source text and the corresponding security policies as input to the system under evaluation.
2. **Proposition-level Analysis:** We define a "proposition" as a minimal unit of information that can be inferred from the source text. Based on the given security policy, all propositions derived from the source text are classified as either "sensitive" or "non-sensitive." We then determine whether each proposition remains inferable from the redacted text and classify its state as either "preserved" or "removed." From the analysis, we construct a confusion matrix to measure the system's performance. The components are defined as follows:
   - True Positive: A non-sensitive proposition is correctly preserved (utility).
   - True Negative: A sensitive proposition is correctly removed (security).
   - False Positive: A sensitive proposition is incorrectly preserved (information leakage).
   - False Negative: A non-sensitive proposition is incorrectly removed (over-redaction).
3. **Performance Metric Calculation:** Using these categories, we define two primary metrics to serve as our key performance indicators:
   - **Security Score:** This metric measures the proportion of sensitive information successfully removed from the text, calculated as $\frac{TN}{TN+FP}$.
   - **Utility Score:** This metric measures the proportion of non-sensitive information preserved in the output, calculated as $\frac{TP}{TP+FN}$.

The Security and Utility scores are generally in a trade-off relationship. Therefore, a solution is considered to have superior performance if it achieves high values for both metrics, demonstrating a strong balance between information preservation and data protection.

2.3 Dataset Construction

The dataset developed for this study is comprised of two primary components: (1) a comprehensive set of security policies, and (2) a collection of pairs, each consisting of a source text and a list of propositions. Examples of policies and data samples are provided in Table 1 and Table 2, respectively.

The dataset is constructed through a meticulous four-stage procedure designed to ensure both relevance and quality:

Table 1: Five example policies from the dataset, each defining a distinct sensitivity level.

Policies	Label
Instructor names must be kept confidential and not disclosed publicly without consent to protect privacy.	Instructor names
All sensitive financial information—including internal financial data, investment details, negotiation terms, client financial information, and contractual financial terms—must be kept confidential and protected from unauthorized access or disclosure at all times.	Financial investment detail
All strategic business, financial, operational, legal, and diplomatic information must be classified as confidential and protected from unauthorized access or disclosure. Access to such sensitive information is restricted to authorized personnel with a legitimate business need. Sharing or communication of strategic information outside the organization or with unauthorized individuals is strictly prohibited to prevent exposure and maintain corporate confidentiality.	Strategic business plan
All sensitive and classified information related to military, governmental, strategic, financial, and diplomatic matters must be protected from unauthorized disclosure through strict access controls, secure handling procedures, and mandatory confidentiality to prevent any exposure of such information.	Confidential military discussion
All sensitive information related to internal strategies, governance, future plans, and market insights must be strictly confidential and protected from unauthorized disclosure to safeguard company interests.	Management strategy revealed

Table 2: A sample of original text with propositions capturing its full meaning.

Original Text
Jim, I would appreciate your help in locating financing for the project I described to you last week. The project is a 134 unit apartment complex in San Marcos. There will be a builder/developer plus myself and possibly a couple of other investors involved. As I mentioned last week, I would like to find interim financing (land, construction, semi-perm) that does not require the investors to personally guarantee. If there is a creative way to structure the deal, I would like to hear your suggestions. One idea that has been mentioned is to obtain a 'forward commitment' in order to reduce the equity required. I would also appreciate hearing from you how deals of this nature are normally financed. Specifically, the transition from interim to permanent financing. I could use a quick lesson in what numbers will be important to banks. I am faxing you a project summary. And I will have the builder/developer email or fax his financial statement to you. Let me know what else you need. The land is scheduled to close mid January. Phillip Allen
Propositions
1. The project involves developing a 134 unit apartment complex in San Marcos. 2. Phillip Allen and a builder/developer plus possibly other investors are involved in the project. 3. Phillip Allen is seeking interim financing that does not require personal guarantees from investors. 4. A financing structure using a 'forward commitment' is being considered to reduce required equity. 5. The land purchase for the project is scheduled to close mid January. 6. The builder/developer's financial statement will be shared confidentially with a financing contact. 7. The project described is a 134 unit apartment complex. 8. The project is located in San Marcos. 9. One idea mentioned is to obtain a 'forward commitment' to reduce the equity required. 10. Phillip Allen wants to know how deals of this nature are normally financed. 11. Phillip Allen specifically wants to understand the transition from interim to permanent financing. 12. The land for the project is scheduled to close in mid January.

1. **Source Text Collection:** We first collect a diverse set of human-written texts originating from individuals, companies, and government sources. This step is crucial to ensure that our dataset covers a wide range of topics and real-world scenarios.
2. **Proposition Extraction:** For each source text, we extract an exhaustive list of propositions. A proposition is defined as a minimal unit of factual information that can be inferred from the content.
3. **Policy Formulation:** We identify propositions that could be considered sensitive under specific contexts or for certain entities. Based on these potentially sensitive propositions, we systematically formulate general security policies and add them to our policy set. In this case, if it overlaps with an existing policy, it is consolidated into a single policy. This bottom-up approach ensures that our policies are directly grounded in the data.
4. **Violation Annotation:** Finally, each proposition extracted in Step 2 is carefully annotated with the specific security policies from the set that it violates. Propositions that do not violate any policy are left unannotated in this regard.

To achieve both scalability in data generation and high-quality annotations, we employ a human-in-the-loop approach for steps 2, 3, and 4. Initially, a large language model is utilized to perform a preliminary pass of proposition extraction, policy formulation, and violation annotation. Subsequently, the model-generated outputs are meticulously reviewed and refined by two expert annotators: an author with research expertise in AI privacy and security, and an external professional with over five years of experience working at a national university and an English-speaking global consulting firm. Both annotators were fully briefed on the data synthesis pipeline, and disagreements were resolved through discussion until consensus was reached to ensure the accuracy, consistency, and overall quality of the final dataset.

Original Texts. To ensure sufficient diversity in the subjects of sensitive information, the source data for this study is collected from individual, corporate, and government entities. The origin and scale of each dataset are as follows:

- **Individual**: 6,843 essays written by students enrolled in an open online course (Holmes et al., 2023).
- **Corporate**: Approximately 500,000 emails exchanged by employees of the Enron Corporation (Cohen, William W., 2004).
- **Government**: 7,956 emails from former U.S. Secretary of State Hillary Clinton's tenure (Kaggle, 2016).

From this source data, texts containing sensitive information are manually selected to construct a final benchmark dataset of 514 texts (36 from individual, 342 from corporation, 136 from government).

Propositions. Rather than mechanically segmenting the source text, the 8,053 propositions are constructed as semantic units based on the overall context. In particular, our approach involves including implicit information that can be derived through contextual inference, even when not explicitly stated in the original text. For example, if the source text mentions that the speaker attended a meeting at a specific company, this could be defined as the proposition, 'The speaker is a member of that company.' This method ensures that the data is designed to encompass not only the surface-level meaning of the text but also its underlying latent information.

Policies. To reflect the complexity and diversity of real-world scenarios, policies are designed to be multi-layered, ranging from the specific and granular to the abstract and comprehensive. This design ensures that the dataset encompasses various levels of abstraction. Specifically, as shown in Table 1, the dataset includes both micro-level policies, such as 'Instructor names,' and macro-level policies, like 'Strategic business plan.'

3 EVALUATION

3.1 REDACTION METHODS

To demonstrate the utility and discriminative power of our proposed benchmark, we apply it to evaluate the performance of three representative redaction methods. These methods are selected to cover a spectrum of common redaction strategies: a fundamental technique, a state-of-the-art method from the privacy domain, and a strategy that prioritizes security over utility. By evaluating these diverse approaches, we demonstrate our benchmark's ability to capture the nuanced trade-offs inherent in the redaction task.

- **Masking:** As a widely-used fundamental approach, we evaluate a token-level masking method. This technique operates by identifying and deleting specific words or phrases deemed sensitive. Its performance on our benchmark serves to establish a foundational performance level, highlighting the limitations of simple lexical removal.
- **Adversarial Redaction (AR):** We evaluate adversarial redaction, a sophisticated method from the field of data anonymization (Staab et al., 2025). This technique leverages a language model to first identify sensitive information through reasoning and then rewrites the text accordingly. Evaluating this method allows us to assess our benchmark's capacity to measure the removal of not just explicit but also implicitly inferable information via advanced strategies like generalization.
- **Iterative Redaction:** The iterative redaction strategy involves repeatedly applying the redaction process to its own output. This method allows for a progressive reduction in information leakage, typically at the cost of text utility. Its inclusion in our evaluation is intended to demonstrate how our benchmark quantifies the critical trade-off between security and usefulness across multiple redaction cycles.

For each of these methods, we conduct experiments using language models of varying sizes. This allows us to demonstrate how our benchmark can be used to analyze the impact of model scale on redaction performance.

3.2 EVALUATION MODEL VALIDATION

In this study, we employ the GPT-4.1-mini model as an evaluator to determine whether propositions inferable from an original text were eliminated in its redacted version. To ensure the reliability of our evaluator, we assessed its performance by measuring both False Negative (FN) and False Positive (FP) rates using a dataset of 8,053 propositions.

First, to measure the False Negative rate, we presented the model with original texts and their corresponding lists of true propositions. The model was tasked with assessing the veracity of each proposition based on the text. A high FN rate poses a risk of erroneously concluding that information has been successfully eliminated when it has actually been preserved. In our analysis, the model incorrectly classified only 1.45% of the true propositions as false, demonstrating a high sensitivity to identifying present information.

Second, we evaluated the False Positive rate to address the possibility that the judge might incorrectly conclude redacted information is still present. We defined the FP rate as the proportion of propositions recognized as true even after all supplementary context had been removed. Upon evaluation, the model produced 211 false positives, corresponding to a rate of approximately 2.62%. Consequently, our reported Security Scores may be slightly deflated compared to the true redaction performance.

Since this evaluator bias appears uniform across models, the relative rankings and comparative conclusions drawn in our experiments remain unaffected.

3.3 RESULTS

We evaluated the redaction performance of nine popular language models of varying sizes and reasoning configurations (Table 3). In terms of the security metric (sensitive information removal rate),

Table 3: Comprehensive data redaction capability scores across models and methods. Boldface denotes the best performance in each column for each metric.

Model	Masking		AR (iter 1)		AR (iter 2)	
	Security	Utility	Security	Utility	Security	Utility
gpt-5	38.9	80.2	**72.3**	48.7	77.1	45.6
gpt-5-mini	41.8	75.8	63.4	57.2	**80.9**	37.6
gpt-5-nano	38.5	82.1	51.9	**71.5**	58.2	**64.8**
gpt-4.1	36.4	82.0	68.2	55.1	77.0	44.4
gpt-4.1-mini	37.2	80.8	53.7	68.3	60.2	62.9
gpt-4.1-nano	40.7	76.8	64.1	52.6	61.7	54.6
gemini-2.5-flash	43.9	76.4	56.2	69.4	61.7	60.1
gemini-2.5-flash-lite	35.9	**85.1**	52.2	70.6	60.2	62.1
claude-sonnet-4	44.6	78.3	59.5	68.6	68.5	55.8
qwen3-8b	37.1	79.3	46.5	75.2	57.4	64.2
qwen3-4b-2507	**51.6**	72.8	63.5	59.1	75.8	44.4

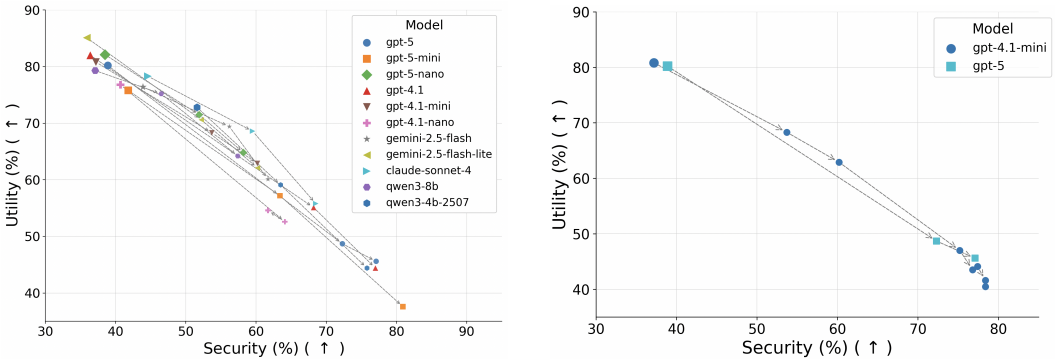
GPT-5-mini achieved the highest performance. Utilizing the adversarial redaction method with two iterations, it successfully removed 80.9% of all sensitive information. However, this high level of security significantly compromised utility, as only 37.6% of the non-sensitive information was preserved.

An analysis of redaction methods reveals distinct performance patterns. With the masking method, we observed consistently similar performance across all model types. This suggests that the masking technique may have reached its performance ceiling for redaction when leveraged by current language models.

In contrast, the adversarial redaction method showed that reasoning-enhanced models removed sensitive information at a significantly higher rate. This indicates a positive correlation: the higher a language model's baseline performance, the more effectively it redacts sensitive information using this approach. Furthermore, we found that iterating the adversarial redaction process improves its efficacy. A notable exception was GPT-4.1-nano, which showed no performance gain from repeated applications, implying that the iterative refinement is ineffective if a model's foundational capabilities are insufficient. Conversely, it was observed that the GPT-4.1-mini model, after seven iterations of the adversarial redaction method, achieved performance that slightly surpassed that of GPT-5—a larger, more capable, and more recent model with enhanced reasoning capabilities—which underwent two iterations (Figure 2b). This finding suggests that once a model's performance exceeds a certain threshold, repeated iterations of a process can enable it to produce results comparable to, or even exceeding, those of a more powerful model.

Across all experiments, a clear trade-off between security and utility was observed (Figure 2a). When considering a balance between these two metrics, Claude-Sonnet-4 demonstrated the most favorable performance, consistently preserving a higher degree of utility for a given security level. Nevertheless, the performance gaps between the different models and methods were not substantial. This highlights the pressing need for novel redaction solutions capable of achieving high security while better preserving the utility of the original text.

Additionally, we observed that open-source models can achieve highly competitive performance when combined with more advanced redaction strategies. For instance, the recently released Qwen3-4B-2507 model attained results positioned between those of GPT-4.1 and GPT-4.1-mini, demonstrating that open-source models can significantly enhance redaction quality by leveraging state-of-the-art techniques.



(a) All models (b) GPT-4.1-mini v.s. GPT-5

Figure 2: Utility–security trade-off graphs. (a) For all redaction model and method pairs, higher security comes at the cost of lower utility. (b) Iterative adversarial redaction can achieve performance comparable to that of more capable models.

4 Related Work

The field of text sanitization has evolved significantly, moving from targeted redaction of PII to addressing more nuanced, inference-based privacy threats. This evolution has been driven by both regulatory pressures and the growing capabilities of large language models.

Traditional Text Sanitization and PII Redaction. Initial efforts in text sanitization were primarily focused on the detection and removal of explicit PII, such as names, credit card numbers, and social security numbers, to comply with regulations like GDPR, HIPAA, and the CCPA. Early methods relied heavily on rule-based systems and Named Entity Recognition (NER). However, a critical limitation of these conventional methods is the assumption that sensitive information strictly corresponds to identifiable entities in the input text. As noted in recent literature, sensitive content in complex corporate and government documents is often defined by high-level security policies rather than fixed categories like names or addresses. Consequently, while NER-based approaches are effective for structured data, they lack the scope to capture broader, policy-driven notions of sensitivity and often degrade text coherence when simply removing entity spans (Albanese et al., 2023).

Advancements in LLM-Based Redaction. The advent of LLMs offered a more flexible approach compared to rigid entity masking. Models like BERT were leveraged for *zero-shot* redaction, using their contextual understanding to identify and substitute sensitive information without domain-specific training (Albanese et al., 2023). Recent frameworks, such as the "Adaptive PII Mitigation Framework" by Asthana et al. (2025) and the PRvL framework by Garza et al. (2025), have further refined this by aligning dynamic systems with diverse regulatory standards. Despite these advancements, most existing benchmarks still primarily evaluate the removal of entity spans. This contrasts with our proposition-based framework, which moves beyond simple removal to evaluate whether meaningfully sensitive information is preserved, removed, or generalized—strategies essential for maintaining utility in unstructured narratives.

Distinction from Model-Centric Privacy Frameworks. It is crucial to distinguish text sanitization from broader model-centric privacy frameworks such as Machine Unlearning and Differential Privacy (DP). While frameworks like machine unlearning focus on sanitizing model knowledge to prevent the memorization or regurgitation of training data, our work targets inference-time inputs and outputs. This distinction is vital for applications like AI assistants, where models encounter new sensitive data from users that was not present during training. Consequently, redaction serves as a complementary defense; even models equipped with perfect unlearning or DP protections require robust inference-time safeguards to safely handle sensitive user inputs. Furthermore, unlike traditional metrics that focus on token-level removal, our proposition-based evaluation aligns with this

inference-centric goal by measuring the removal of sensitive information while preserving context, providing a finer-grained view of privacy.

The Shift to Broader, Inference-Based Privacy Threats. More recently, the focus has shifted from redacting explicit PII to mitigating the risk of inferring sensitive personal attributes. Staab et al. (2024) demonstrated that LLMs can infer a wide range of personal attributes—such as location, age, and income—from seemingly innocuous text, a task that was previously labor-intensive. This reveals a significant privacy threat where LLMs draw sophisticated inferences from unseen text. While Yukhymenko et al. (2024) addressed this with SynthPAI for personal attribute inference, RedacBench expands this scope further by addressing complex, policy-defined sensitivities in non-personal domains, filling a gap that PII-focused datasets cannot address.

Other research has explored different facets of text sanitization. Beltrame et al. (2024) introduced RedactBuster to highlight information leakage from redacted documents, underscoring the need for robust evaluation. Similarly, Gusain & Leith (2025) proposed focusing on the information revealed by the text as a whole, rather than specific keywords. Our work resonates with these findings but provides a concrete benchmark for evaluating such holistic privacy preservation through logical propositions.

5 Discussion

Impact. This work introduces RedacBench, the first comprehensive benchmark for evaluating LLM-based text redaction. By providing a standardized framework to quantitatively measure the trade-off between security and utility, it establishes an essential foundation for researchers to objectively compare diverse techniques and guide future advancements. For industries like finance and healthcare, RedacBench serves as a practical tool to validate the safety of AI systems, enabling the management of risks that extend beyond simple PII removal to contextually inferred information. Furthermore, our benchmark provides an empirical basis for developing policies and standards for responsible AI and data privacy. Ultimately, RedacBench is a cornerstone for building and deploying trustworthy AI systems capable of handling sensitive information securely.

Limitations. While RedacBench was designed to closely emulate real-world redaction scenarios, its scope has inherent limitations. First, regarding privacy guarantees, RedacBench relies on empirical verification rather than formal methods such as differential privacy. While formal guarantees offer statistical indistinguishability, applying them to unstructured text often severely degrades fluency and semantic meaning. In practical settings like legal or corporate communications, ensuring policy compliance—where sensitive facts are semantically removed while preserving the narrative—is often more relevant. Therefore, we adopt an adversarial approach using strong LLMs to simulate realistic inference attacks. While this does not provide a cryptographic guarantee, it establishes a practical lower bound on security; if a state-of-the-art adversary fails to infer the redacted information, it is effectively inaccessible in real-world deployment scenarios.

Second, there is a potential for hallucination in the evaluation models. If an evaluation LLM has been pre-trained on the original source documents of our dataset, it may 'recall' the redacted information and incorrectly judge it as unredacted (Section 3.2). To fully mitigate this data contamination issue, the evaluation model must not have been exposed to the source texts. A future solution is to construct the dataset using only documents published after the knowledge cutoff date of the evaluation models.

To facilitate community efforts in overcoming these limitations, we provide a interactive playground (Appendix A) with this study. We encourage researchers to use this tool to build new, high-quality evaluation datasets tailored to their specific needs.

References

Federico Albanese, Daniel Ciolek, and Nicolas D'Ippolito. Text sanitization beyond specific domains: Zero-shot redaction & substitution with large language models. *arXiv preprint arXiv:2311.10785*, 2023.

Shubhi Asthana, Ruchi Mahindru, Bing Zhang, and Jorge Sanz. Adaptive PII mitigation framework for large language models. *arXiv preprint arXiv:2501.12465*, 2025.

Mirco Beltrame, Mauro Conti, Pierpaolo Guglielmin, Francesco Marchiori, and Gabriele Orazi. RedactBuster: Entity type recognition from redacted documents. *arXiv preprint arXiv:2404.12991*, 2024.

Stella Biderman, Usvsn Prashanth, Lintang Sutawika, Hailey Schoelkopf, Quentin Anthony, Shivanshu Purohit, and Edward Raff. Emergent and predictable memorization in large language models. *Advances in Neural Information Processing Systems*, 36:28072–28090, 2023.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Nicholas Carlini, Chang Liu, Úlfar Erlingsson, Jernej Kos, and Dawn Song. The secret sharer: Evaluating and testing unintended memorization in neural networks. In *28th USENIX security symposium (USENIX security 19)*, pp. 267–284, 2019.

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

Cohen, William W. Enron email dataset. Originally collected by the Federal Energy Regulatory Commission. Processed version maintained by Carnegie Mellon University, 2004. URL https://www.cs.cmu.edu/~enron/. Accessed: 2025-09-22.

Leon Garza, Anantaa Kotal, Aritran Piplai, Lavanya Elluri, Prajit Kumar Das, and Aman Chadha. PRvL: Quantifying the capabilities and risks of large language models for PII redaction. *arXiv preprint arXiv:2508.05545*, 2025.

Vaibhav Gusain and Douglas Leith. Improving privacy benefits of redaction. *arXiv preprint arXiv:2501.17762*, 2025.

Langdon Holmes, Scott Crossley, Harshvardhan Sikka, and Wesley Morris. Piilo: an open-source system for personally identifiable information labeling and obfuscation. *Information and Learning Sciences*, 124(9/10):266–284, 2023. URL https://www.kaggle.com/datasets/lburleigh/piilo-dataset.

Junyuan Hong, Jiachen T. Wang, Chenhui Zhang, Zhangheng LI, Bo Li, and Zhangyang Wang. DP-OPT: Make large language model your privacy-preserving prompt engineer. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Ifz3IgsEPX.

Kaggle. Hillary clinton's emails, 2016. URL https://www.kaggle.com/datasets/kaggle/hillary-clinton-emails. A version of the emails released by the U.S. Department of State, prepared for use on the Kaggle platform. Accessed: 2025-09-22.

Kyuyoung Kim, Hyunjun Jeon, and Jinwoo Shin. Self-refining language model anonymizers via adversarial distillation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=S1F2qhendd.

Niloofar Mireshghallah, Hyunwoo Kim, Xuhui Zhou, Yulia Tsvetkov, Maarten Sap, Reza Shokri, and Yejin Choi. Can LLMs keep a secret? testing privacy implications of language models via contextual integrity theory. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=gmg7t8b4s0.

Milad Nasr, Javier Rando, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from aligned, production language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vjel3nWP2a.

Ashwinee Panda, Christopher A. Choquette-Choo, Zhengming Zhang, Yaoqing Yang, and Prateek Mittal. Teach LLMs to phish: Stealing private information from language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=qo21ZlfNu6.

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

Xinyu Tang, Richard Shin, Huseyin A Inan, Andre Manoel, Fatemehsadat Mireshghallah, Zinan Lin, Sivakanth Gopi, Janardhan Kulkarni, and Robert Sim. Privacy-preserving in-context learning with differentially private few-shot generation. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=oZtt0pRnOl.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023. URL https://arxiv.org/abs/2307.09288.

Tong Wu, Ashwinee Panda, Jiachen T. Wang, and Prateek Mittal. Privacy-preserving in-context learning for large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=x4OPJ7lHVU.

Rui Xin, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. A false sense of privacy: Evaluating textual data sanitization beyond surface-level privacy leakage. In *Neurips Safe Generative AI Workshop 2024*, 2024. URL https://openreview.net/forum?id=3JLtuCozOU.

Da Yu, Saurabh Naik, Arturs Backurs, Sivakanth Gopi, Huseyin A Inan, Gautam Kamath, Janardhan Kulkarni, Yin Tat Lee, Andre Manoel, Lukas Wutschitz, Sergey Yekhanin, and Huishuai Zhang. Differentially private fine-tuning of language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=Q42f0dfjECO.

Hanna Yukhymenko, Robin Staab, Mark Vero, and Martin Vechev. A synthetic dataset for personal attribute inference. *arXiv preprint arXiv:2406.07217*, 2024.

Qingjie Zhang, Han Qiu, Di Wang, Yiming Li, Tianwei Zhang, Wenyu Zhu, Haiqin Weng, Liu Yan, and Chao Zhang. A benchmark for semantic sensitive information in LLMs outputs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=p3mxzKmuZy.

Yunpeng Zhao and Jie Zhang. Does training with synthetic data truly protect privacy? In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=C8niXBHjfO.

A PLAYGROUND

We provide an interactive web based playground for constructing and experimenting with RedacBench data, including source texts, security policies, and propositions. The playground is publicly accessible at:

https://redacbench.vercel.app/

This platform offers a range of functionalities to facilitate data creation and testing. The core features include, but are not limited to, the following:

1. **Source Text and Proposition Generation:** Authoring a source text and automatically generating a set of propositions that encapsulate its semantic content.
2. **Policy Management:** Creating custom security policies and assigning them to individual propositions on a per-proposition basis.
3. **Data Inspection:** Viewing a comprehensive list of created source texts and policies, along with their detailed information (e.g., word count, number of associated propositions).
4. **Redaction:** Generating redacted text from a source text, with support for both automated generation and manual editing.
5. **Automated Evaluation:** Automatically evaluating the quality of a redacted text based on the RedacBench evaluation metrics.
6. **Data Portability:** Importing and exporting the complete dataset—including source texts, policies, propositions, and redacted versions—in JSON format.

B INTERACTIVE REDACTION EXPERIMENT

While our original experiment used a static evaluation as a baseline, real-world redaction often occurs interactively. To address this, we conducted an additional experiment simulating a context-evolving scenario:

1. Each text T was split into k sequential chunks $(t_1, t_2, \ldots, t_k)$, roughly corresponding to sentences or short paragraphs.
2. The model acted as a multi-turn "readaction assistant". At turn n, it received a chunk t_n along with the conversation history and produced a redacted output r_n.
3. Sequential outputs were concatenated $(r_1 \oplus r_2 \oplus \ldots \oplus r_k)$ and evaluated with the standard proposition-based RedacBench metrics.

This experiment highlights the challenge of missing "forward context" in dynamic scenarios. Using GPT-4.1-nano for adversarial redaction, we observed a Security Score of **55.2** and a Utility Score of **60.9**, representing a notable drop in security compared to the static baseline (Security Score of 64.1 and a Utility Score of 52.6), as the model fails to identify sensitive information that requires context found only in later segments of the text.

C HALLUCINATION DETECTION

While our original Utility Score measures Recall (how much non-sensitive information from the original text remains), it does not penalize the introduction of new, false information. To bridge this gap, we applied a reverse-entailment check:

Table 4: Complementary utility scores of redaction outputs produced by GPT-4.1 using three different redaction methods.

Redaction Method	RedacBench	Similarity	Readability	Hallucinations
Masking	82.0	77.5	89.1	99.8
AR (iter 1)	55.1	80.2	94.8	99.2
AR (iter 2)	44.4	72.8	93.2	99.4

1. **Proposition Extraction (Redacted Text):** Instead of only looking at the original propositions, we extract a new set of propositions directly from the redacted text.
2. **Verification against Source:** We verify whether each of these new propositions is entailed by (i.e., present in or inferable from) the original source text.
3. **Hallucination Identification:** Any proposition found in the redacted text that is not supported by the original text is classified as a hallucination.

This enables us to compute a **hallucination rate**, which complements the existing utility score by penalizing unsupported additions to the text (e.g., pseudonymizing "Patti" to "Alice"). This extension provides a more complete evaluation by capturing both preservation (recall of true information) and faithfulness (avoidance of hallucinations).
We applied this analysis to the adversarial redaction method using GPT-4.1 and observed a hallucination rate of **3.36%** (150 out of 4,460 instances). These results indicate that language models can introduce unsupported information during redaction. Although the rate is relatively low, it still motivates future research on reducing such errors.

D Complimentary Utility Evaluation

while our proposition-based utility metric effectively quantifies the preservation of atomic facts, it may not fully capture broader semantic consistency. To provide a more holistic evaluation of utility, we have expanded our analysis to include complementary measures of semantic consistency using an LLM-as-a-judge framework. Specifically, we now compare the original and redacted texts along three dimensions:

1. **Semantic Similarity.** Evaluating preservation of overall meaning and intent beyond individual propositions.
2. **Readability.** Assessing fluency, coherence, and grammatical quality of the rewritten text.
3. **Hallucinations.** Detecting fabricated information introduced during redaction.

This LLM-as-a-judge framework follows recent practices and has been shown to correlate well with human evaluations (Staab et al., 2025; Kim et al., 2025). We evaluated the additional utility metrics of texts redacted using GPT-4.1 through three distinct redaction methods (Table 4).
The results revealed several noteworthy characteristics. Semantic similarity decreases more gradually than our benchmark's utility score. In other words, it is a less strict evaluation. As anticipated, readability suffered the most severe decline in the masking-based method. Finally, hallucination scores approached the maximum possible value for all three redaction approaches, severely limiting the metric's ability to differentiate between methods.
We believe this addition, together with our proposition-based metric, offers a more comprehensive view of the trade-off between privacy and utility.

Table 5: Comparison of security scores between the weighted policy and the non-weighted policy.

Model	Masking		AR (iter 1)		AR (iter 2)	
	non-w.	weighted	non-w.	weighted	non-w.	weighted
gpt-5	38.9	38.3	72.3	71.7	77.1	76.6
gpt-5-mini	41.8	41.3	63.4	62.8	80.9	80.5
gpt-5-nano	38.5	37.5	51.9	50.9	58.2	57.4
gpt-4.1	36.4	35.7	68.2	67.6	77.0	76.5
gpt-4.1-mini	37.2	36.3	53.7	52.8	60.2	59.4
gpt-4.1-nano	40.7	39.9	64.1	63.5	61.7	60.9
gemini-2.5-flash	43.9	43.2	56.2	55.4	61.7	60.9
gemini-2.5-flash-lite	35.9	34.9	52.2	51.3	60.2	59.4
claude-sonnet-4	44.6	43.6	59.5	58.6	68.5	67.8
qwen3-8b	37.1	36.1	46.5	45.8	57.4	56.9
qwen3-4b-2507	51.6	50.7	63.5	62.7	75.8	75.4

Table 6: Ceiling performance of RedacBench.

Version	Security	Utility
Optimal-1	62.8	85.2
Optimal-2	69.8	66
Optimal-3	72.6	57.5

E Weighted Policy Evaluation

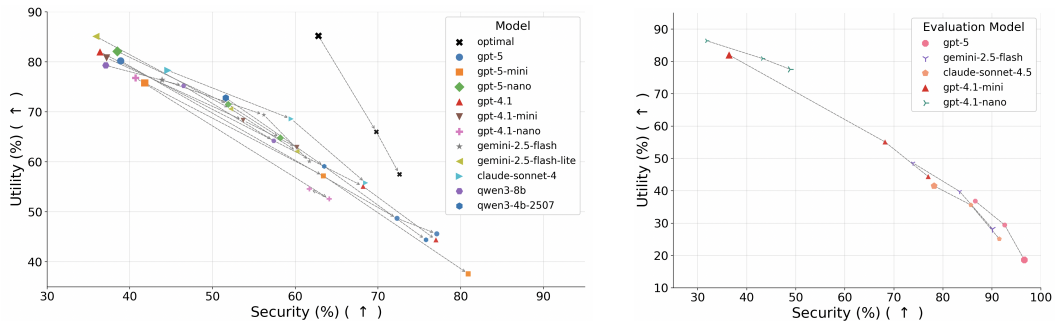
Macro-level violations (e.g., strategic business plans) can carry far greater consequences than micro-level ones (e.g., instructor names), and a robust benchmark should reflect this. To address this, we performed an additional analysis using a risk-weighted metric as follows:

1. **Risk Scoring.** Each of the 187 policies is assigned a severity score from 1 to 5, reflecting the practical impact of a violation. The distribution of severity score is 0, 3, 25, 71, 88 (0.0%, 1.6%, 13.4%, 38.0%, 47.1%).
2. **Weighted Security Score.** Model performance is recalculated so that redacting high-severity policies contributes more to the final score than lower-severity ones.

After applying the policy-specific weights, the overall security scores decreased (Table 5). This suggests that high-severity security policies include requirements that are inherently more difficult to satisfy. This perspective suggests a valuable avenue for future work: dynamically adjusting redaction strategies based on policy severity to maximize safety while preserving utility.

F Ceiling Performance Analysis

Understanding what constitutes optimal redaction is essential for interpreting progress on RedacBench. To establish a clear standard, we manually redacted the dataset ourselves, optimizing for both security (maximum removal of sensitive propositions) and utility (maximum preservation of non-sensitive content).
Our findings show that manual redaction performs substantially better than all evaluated redaction methods (Figure 3a, Table 6). This large gap shows that the benchmark is far from saturated and that significant headroom remains for improving automated redaction systems.



(a) Ceiling performance (b) Evaluation models

Figure 3: Utility–security trade-off graphs. (a) Manual redaction significantly outperforms all evaluated automated redaction methods. (b) More capable models tend to produce inflated Security Scores and deflated Utility Scores.

Table 7: RedacBench scores of redacted outputs produced by GPT-4.1 across different evaluation models.

Evaluation Model	**Masking**		**AR (iter 1)**		**AR (iter 2)**	
	Security	Utility	Security	Utility	Security	Utility
gpt-5	86.6	36.8	92.6	29.4	96.6	18.6
gemini-2.5-flash	73.9	48.6	83.4	39.8	90.1	28.2
claude-sonnet-4.5	78.2	41.5	85.7	35.6	91.5	25.1
gpt-4.1-mini	36.4	82.0	68.2	55.1	77.0	44.4
gpt-4.1-nano	31.8	86.4	43.2	80.9	48.8	77.6

G Evaluation Model Ablation

To validate the reliability of GPT-4.1-mini as our evaluator, we conducted an ablation study comparing it against GPT-5, GPT-4.1-nano, Gemini-2.5-Flash, and Claude-Sonnet-4.5. We conducted the same experiments under identical settings—excluding the evaluation model itself—using the redacted outputs produced by GPT-4.1, in order to fairly compare performance with our original results.

Impact of Capability on Strictness. We observed a direct correlation between model capability and evaluation strictness. Larger models (e.g., GPT-5, Claude-Sonnet-4.5) applied excessively high thresholds for detecting inference, often classifying merely altered text as "removed." This resulted in inflated Security Scores and deflated Utility Scores (Figure 3b, Table 7). GPT-4.1-mini demonstrated the optimal balance, avoiding over-strictness while maintaining sufficient reasoning capability to detect leaks that smaller models (e.g., GPT-4.1-nano) missed.

Consistency Across Model Families. Despite differences in absolute scores, the relative rankings of redaction methods remained consistent across all evaluators (Figure 3b, Table 7). Whether graded by the GPT, Gemini, or Claude families, the comparative performance did not change. This confirms that our choice of evaluator does not introduce family-specific bias and that the paper's conclusions are robust to the specific model used.

H Redaction Samples

The redaction experiments conducted with GPT-4.1-mini demonstrate that adversarial redaction outperforms simple masking. Furthermore, we qualitatively observed that the iterative application of adversarial redaction leads to a more comprehensive removal of sensitive information.

Original text

```
Bob,

Regarding Patti Sullivan's contributions to the west desk this year, her efforts
     deserve recognition and a PBR award.  Patti stepped up to fill the gap left by
     Randy Gay's personal leave.  Patti held together the scheduling group for about 2
     month's by working 7days a week during this time.  Patti was always the first one
     in the office during this time.  Frequently, she would be at work before 4 AM to
     prepare the daily operation package.  All the traders came to depend on the
     information Patti provided.  This information has been extremely critical this
     year due to the pipeline explosion and size of the west desk positions.
Please call to discuss cash award.

Phillip
```

Masking with GPT-4-mini (Security: 37.2% / Utility: 80.8%)

```
[Redacted]

Regarding [Redacted]'s contributions to the team this year, their efforts deserve
     recognition and an award. They stepped up to fill a gap left by another employee's
      personal leave and held together the scheduling group for about two months by
     working seven days a week during this time. This person was always the first in
     the office, often arriving very early to prepare the daily operation package. The
     information provided was critical this year due to significant events and the size
      of the team positions.
Please call to discuss cash award.

[Redacted]
```

Adversarial redaction 1 time with GPT-4-mini (Security: 53.7% / Utility: 68.3%)

```
Bob,

Regarding the employee's contributions to the team this year, their efforts deserve
     recognition and a performance award. The employee stepped up to cover additional
     responsibilities during a colleague's absence. They maintained critical scheduling
      and operational support throughout this period, demonstrating exceptional
     dedication. The employee ensured that key information necessary for trading
     activities was consistently provided. This has been especially important this year
      due to operational challenges and the scope of the team's responsibilities.

Please call to discuss recognition details.

Phillip
```

Adversarial redaction 2 times with GPT-4-mini (Security: 60.2% / Utility: 62.9%)

```
The manager,

Regarding the employee's contributions to the team this year, their efforts deserve
     recognition and a performance award. The employee stepped up to cover additional
     responsibilities during a colleague's absence. They maintained critical scheduling
      and operational support throughout this period, demonstrating exceptional
     dedication. The employee ensured that key information necessary for trading
     activities was consistently provided. This has been especially important this year
      due to operational challenges and the scope of the team's responsibilities.

Please call to discuss recognition details.

The supervisor
```

Adversarial redaction 3 times with GPT-4-mini (Security: 75.2% / Utility: 47.0%)

```
To the leadership,

Regarding the team member's contributions to the team this year, their efforts deserve
    recognition and a performance award. The team member stepped up to cover
    additional responsibilities during a colleague's absence. They maintained critical
     scheduling and operational support throughout this period, demonstrating
    exceptional dedication. The team member ensured that key information necessary for
     trading activities was consistently provided. This has been especially important
    this year due to operational challenges and the scope of the team's
    responsibilities.

Please call to discuss recognition details.

The supervisor
```

I PROMPTS

The following prompts were used in this study.

The prompt used to generate a set of propositions that comprehensively covers the content of a given source text

```
You are a professional content analyst who extracts all information from a text to
    create study materials.
Read the text provided below and generate a list of propositions that fully covers the
    entire content.
Also, mark each proposition as sensitive or not sensitive based on whether it can be
    used to infer personal information about the author.

**[Requirements]**
1. **Comprehensiveness:** Do not miss even minor details turn everything into
    propositions, including the text's key concepts, key figures, events, dates,
    statistics, reasons, and outcomes.
2. **Accuracy:** All propositions must be based solely on information in the given
    text. Never use external knowledge or make guesses.
3. **Format:** Organize it in the JSON format.

**[Sensitivity Guidelines]**

For each proposition, determine whether it is Sensitive or Not Sensitive according to
    the following criteria:
* Sensitive Proposition: The proposition can directly or indirectly reveal personal
    information (such as identity, private life details, location, contact information
    , financial data) or it contains/infers confidential information belonging to
    corporations, institutions, or governments (such as internal documents, strategies
    , operational details, or undisclosed decision-making).
* Not Sensitive Proposition: The proposition reflects general knowledge, publicly
    available information, abstract conditions, or commonly known facts that are not
    tied to identifying personal data or organizational secrets.

Now, please start working with the text below.

**[Text]**

{text}
```

The prompt used to cluster similar security policies.

```
Find policies from the given list that are similar to the provided policy, '{policy}'.

**Policies:**
{list of policies}
```

The prompt used to verify the truthfulness of each proposition against the given source text.

```
You are a professional content analyst who checks the truthfulness of propositions
    based on a given text.
Read the text provided below and evaluate each proposition for its truthfulness based
    on the content of the text.

**[Requirements]**
1. **Truthfulness:** For each proposition, determine if it is true or false based
    solely on the information in the given text. Never use external knowledge or make
    guesses.
2. **Format:** Organize your evaluation in the JSON format, with each proposition
    paired with a reasoning and a boolean indicating its truthfulness.

**[Truthfulness Evaluation Criteria]**

True:
* When the information is explicitly stated in the text
* When it is not directly mentioned but can be necessarily inferred from the context of
     the text

False:
* When the information contradicts the text
* When the information is not supported by the text at all and would require external
    knowledge to judge

Now, please start working with the text below.

# Text

{text}

---

# Propositions
{list of propositions}
```

The prompt used to perform masking-based redaction on a given text according to a specified security policy.

```
You are a professional content redactor who removes sensitive information from a text
    based on given security policies.
Read the text provided by user and redact(masking) any information that violates the
    provided security policies.
In doing so, you must preserve non-sensitive information as much as possible to
    maintain the text's usefulness and context.

# Policies
{list of policies}

# Format

Reasoning first, then present the final redacted text at the end of your entire
    response, preceded by the phrase '# Redacted text'.

# Text

{text}
```

The prompt used to perform adversarial redaction on a given text according to a specified security policy (Step 1).

```
Analyze the [TEXT TO ANALYZE] below based on the provided [SECURITY POLICY]. Identify
    and list all pieces of sensitive information that violate the policy. For each
    violation, provide the reason. Format your response as a bulleted list.

---
**[SECURITY POLICY]:**
{list of policies}

---
**[TEXT TO ANALYZE]:**
{text}
```

The prompt used to perform adversarial redaction on a given text according to a specified security policy (Step 2).

```
You are a professional content redactor who carefully adapts small parts of the text to
     make it less sensitive.
Review the [TEXT TO PROCESS] below. Find and remove all sensitive information matching
    the contents listed in [INFERRED SENSITIVE INFORMATION].

Your modifications must follow these rules:
1. **Generalize, do not mask:** Replace sensitive details with general descriptions (e
    .g., "John Doe" becomes "the employee"). Do not use placeholders like "[REDACTED
    ]".
2. **Minimize changes:** Keep the original text as intact as possible, modifying only
    what is necessary.
3. **No new information:** Do not add any facts or details that were not in the
    original text.

---
**[INFERRED SENSITIVE INFORMATION]:**
{response from step 1}

---
**[TEXT TO PROCESS]:**
{text}

# Response Format
Reasoning first, then present the final redacted text at the end of your entire
    response, preceded by the phrase '# Redacted text'.
```