REDACBENCH: CAN AI *ERASE* YOUR SECRETS?

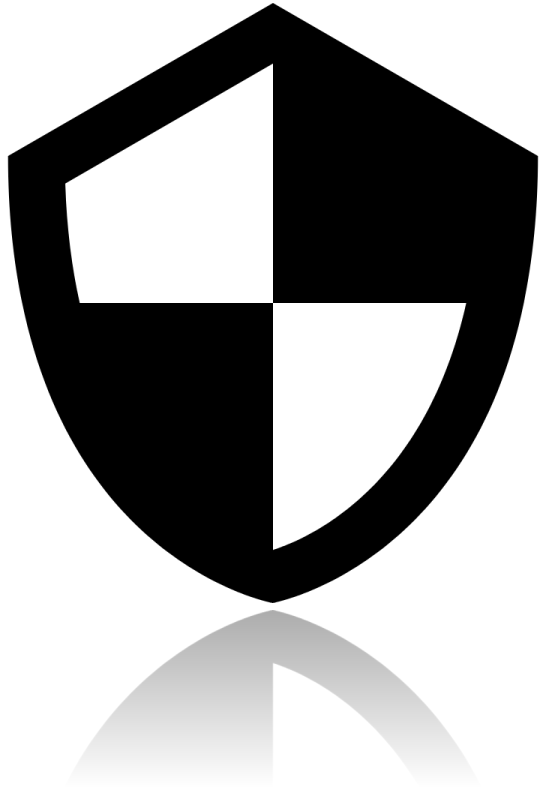
Hyunjun Jeon, Kyuyoung Kim, Jinwoo Shin

Korea Advanced Institute of Science and Technology (KAIST)

Published at ICLR 2026

MOTIVATION — WHY REDACTION MATTERS

- **Training data extraction:** Memorized PII reproduced via targeted queries to LLMs.
- **Inference-time data leakage:** Sensitive data naturally inserted into prompts (RAG, assistants).
- **Attribute inference:** Inferring professions, health data, or relationships from innocuous text.
- **The Gap:** Existing approaches rely on surface-level matching, offering a "false sense of privacy" without a proper evaluation framework.



LIMITATIONS OF EXISTING WORK

Rigid Categories

Prior benchmarks focus heavily on predefined PII categories (names, SSNs) rather than diverse, policy-driven notions of sensitivity across specific domains.

Surface Removal

Evaluation is severely limited to simple entity-span removal rather than genuine semantic information removal and abstraction.

No Inference Checks

There is no assessment of whether sensitive information remains inferable after redaction, nor a standardized security–utility trade-off evaluation.

KEY CONTRIBUTIONS

RedacBench

A comprehensive benchmark featuring 514 human-authored texts, 187 security policies, and 8,053 heavily annotated propositions.

Baseline Analysis

In-depth evaluation of multiple redaction strategies across 11 state-of-the-art LLMs, clearly mapping the crucial security–utility trade-off.

Interactive Playground

A publicly released web-based tool allowing users to explore dataset customization, model experimentation, and real-time redaction.

TASK DEFINITION

The Redaction Task

Selectively remove sensitive information from a source text according to a specifically provided security policy.

Input: Source text + Security policy

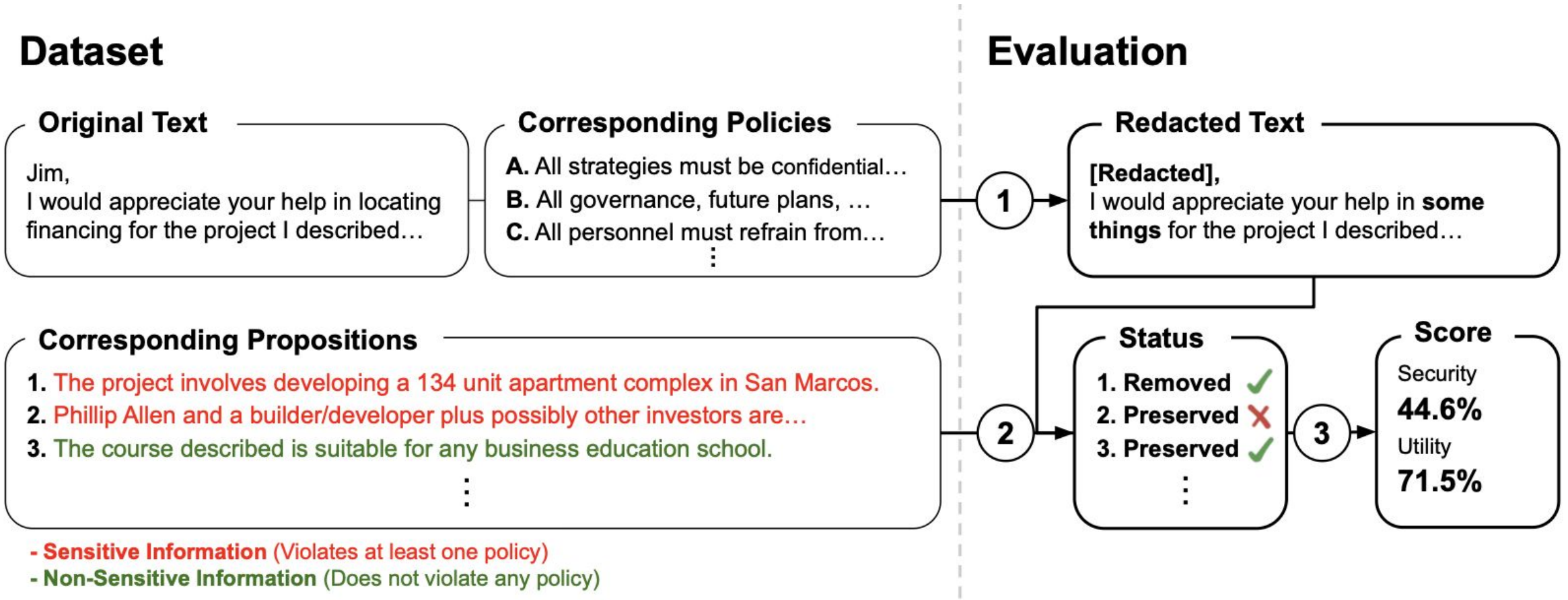
Output: Redacted text that strictly complies with the policy.

Contextual Sensitivity

Sensitivity criteria vary wildly by context. It is impossible to enumerate all categories universally.

By mapping high-level security policies to unstructured text, the task better reflects authentic operational and corporate environments.

EVALUATION FRAMEWORK OVERVIEW



PROPOSITIONS & METRICS

Propositions & Matrix

A minimal unit of factual information inferable from the text.

	Preserved	Removed
Non-sensitive	TP ✓	FN ×
Sensitive	FP ×	TN ✓

Trade-off Metrics

These two metrics are in a strict trade-off relationship.

$$\text{Security Score} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

$$\text{Utility Score} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

DATASET CONSTRUCTION

514

Human-Authored Texts

187

Security Policies

8,053

Annotated Propositions

Data Source	Origin	Count
Individual	Student essays (PIILO)	36
Corporate	Enron emails	342
Government	Clinton emails	136

REDACTION METHODS EVALUATED

Date of Interview: February 28, 2006

Following a series of discussions between counsel, [redacted] consented to a "free talk" interview to be conducted at the offices of the United States Attorney for the District of Arizona, in Phoenix, AZ. [redacted] was accompanied by counsel [redacted] throughout the interview. AUSA [redacted] and NPS Supervisory Special Agent [redacted] were present for the government.

As noted, it was previously agreed upon that this interview would take place under the provisions of a "free talk" agreement. A copy of that agreement is attached [See Attachment # 17].

[redacted] Finally, in 1981 or 1982, [redacted] was selected for the position of [redacted] at Hubbell Trading Post, in Ganado, AZ, working for Southwest Parks and Monuments Association, later known as Western National Parks Association. [redacted] estimated [redacted] salary at HTP as approximately $12 per hour, for a forty-hour work week. [redacted] added that [redacted] was routinely paid an additional monthly bonus check based upon the volume of sales realized at the trading post; typically around $800 or $900 per month. [redacted] stated that [redacted] usually cashed [redacted] paychecks and bonus

1. Masking

Token-level keyword matching & removal. Offers surface-level lexical removal without contextual reasoning.

2. Adversarial Redaction

Identifies violations, then rewrites text. Enables advanced syntactic and semantic redaction via LLM reasoning.

3. Iterative Redaction

Repeatedly applies AR to its own output. Aims to remove residual sensitive content at the cost of utility.

EVALUATION RELIABILITY

Evaluator: **GPT-4.1-mini**

Automated proposition verification validated over 8,053 manual annotations.

1.45%

FALSE NEGATIVE RATE

Strong recall of inferable info

2.62%

FALSE POSITIVE RATE

Security scores slightly underestimated

> The same evaluator was applied consistently across all experiments. Therefore, relative model comparisons remain highly meaningful and stable.

MAIN RESULTS

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

Best overall security: GPT-5-mini running Adversarial Redaction iteration 2 achieves **80.9%** security, but sacrifices utility, dropping it to a mere **37.6%**.

SECURITY–UTILITY TRADE-OFF

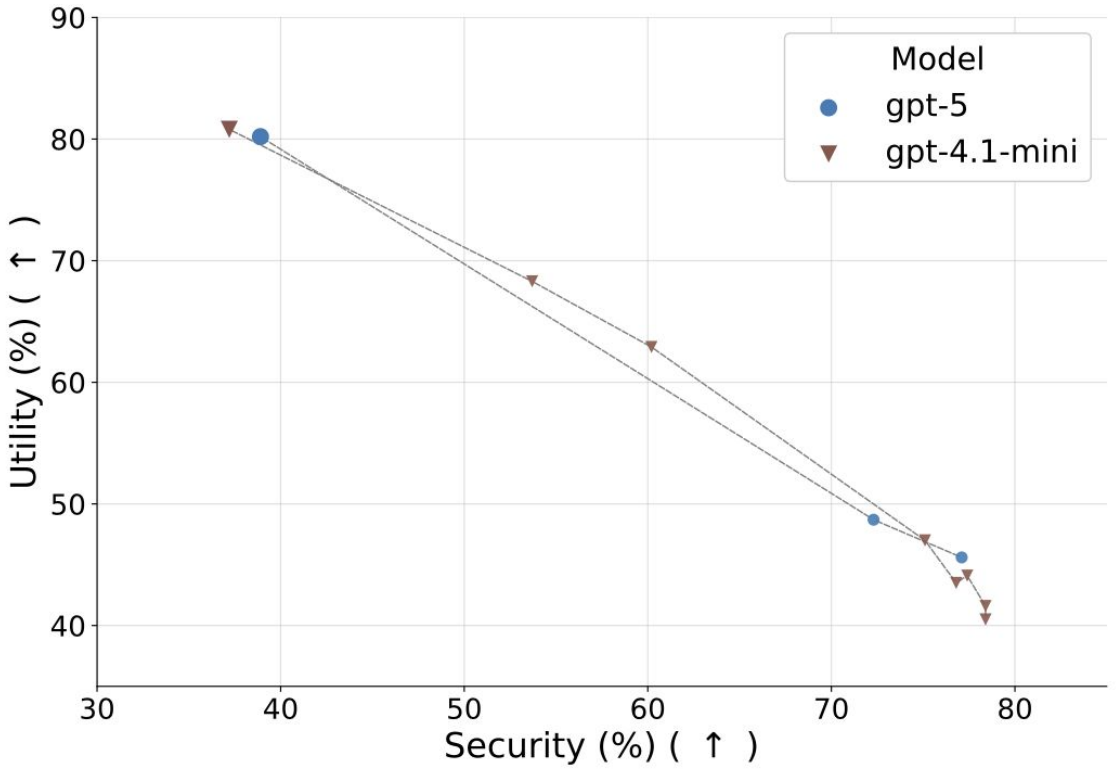
- **Clear trade-off:** Higher security inherently leads to lower utility across all evaluated models.
- **Masking ceiling:** Masking hits a strict performance ceiling regardless of model scale.
- **Reasoning differentiation:** Adversarial Redaction shows clear model separation; enhanced reasoning works best.
- **Open-source capability:** Open models (Qwen3) compete impressively with advanced strategies.

ITERATIVE REDACTION INSIGHT

ITERATIONS COMPENSATE FOR SCALE.

GPT-4.1-mini running with **7 iterations** performs approximately equal to the much larger GPT-5 running with just **2 iterations**.

> **The Threshold Effect:** Iterative refinement is completely ineffective when foundational capabilities are limited. Once a model passes a capability threshold, repeated iterations can successfully bridge the gap to more powerful models.



LIMITATIONS

No formal privacy guarantees.

Evaluations are empirical verifications, not strict differential privacy. Formal methods degrade fluency; adversarial evaluation provides a practical lower bound.

Evaluator hallucination risk.

If the evaluation model was pre-trained on source documents, it may synthetically "recall" redacted information. (Mitigated by using newer documents).

Dataset Scope Constraints.

Currently limited to English text, explicitly focused on three main domains, using publicly available sources.

CONCLUSION & IMPACT

RedacBench provides a standardized framework for evaluating policy-conditioned redaction and quantitatively measuring the security–utility trade-off.

While advanced models improve security, preserving utility remains a major challenge. Current LLM-based redaction is far from optimal, leaving significant room for future innovations.

hyunjunian.github.io/redaction-playground/