REDACTBENCH: A FORMAL FRAMEWORK FOR LLM-BASED CONFIDENTIAL INFORMATION REDACTION

Saqif Ayaan Sudheer[1], Pedro Gonzalez[2], Aditya Singh[3]

1. UC Santa Barbara, 2. CSU San Bernardino, 3. Silver Creek High School

UC SANTA BARBARA

CALIFORNIA STATE UNIVERSITY SAN BERNARDINO

Introduction

Protecting sensitive information is critical for governments, companies, and institutions. **Redaction** enables sharing documents while concealing confidential content, balancing transparency with security.

Example: The redacted documents released by the U.S. Government under FOIA [1]

The challenge: Redaction is done manually by experts. The process is slow, costly, and error-prone [1]. Rules vary widely across domains such as finance, energy, and defense, making automation difficult.

Our contribution: We introduce **REDACTBENCH,** the first framework to:

A. Generate synthetic documents to benchmark redaction performance

B. Automate context-aware information redaction by using LLM-based agents.

Research Questions

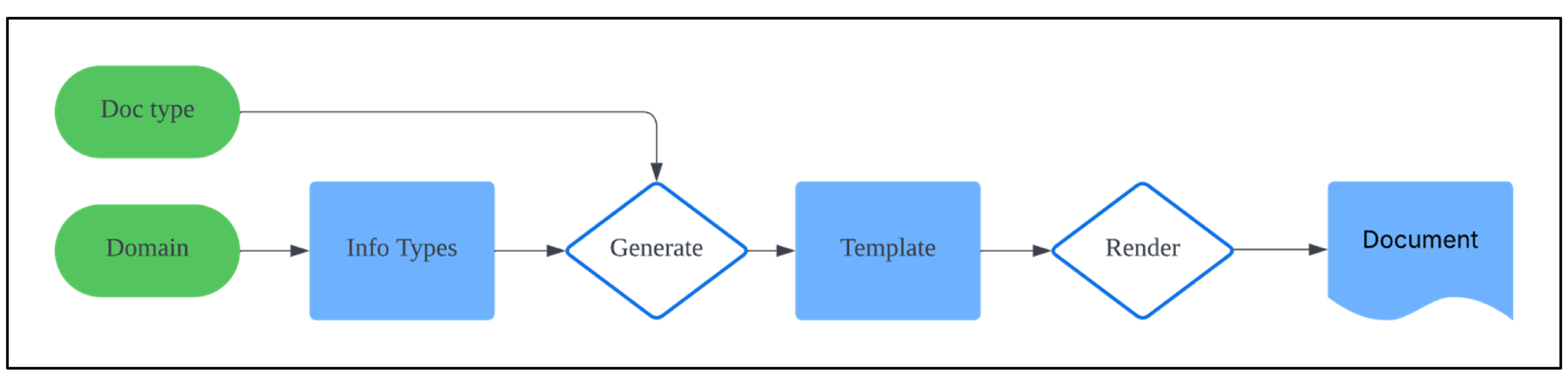
RQ1: Can LLMs perform accurate, domain-specific redactions beyond basic PII removal [2,3] while preserving document utility and minimizing leakage?

RQ2: How can synthetic document pipelines be designed to systematically evaluate and fine-tune LLM-based redaction across domains and information types?

RQ3: Which evaluation metrics best capture the *trade-off* between redaction accuracy, residual leakage, and retained utility?

OUR FRAMEWORK

A. Synthetic Document Generation Pipeline



Goals:

1. Generate synthetic documents using LLMs that contain both confidential (to be redacted) and non-confidential (to be preserved) information.
2. These documents should be **realistic** in content, structure, and complexity to serve as effective benchmarks for LLM-based redaction.

Approach:

Domain selection: Military, finance, energy,…

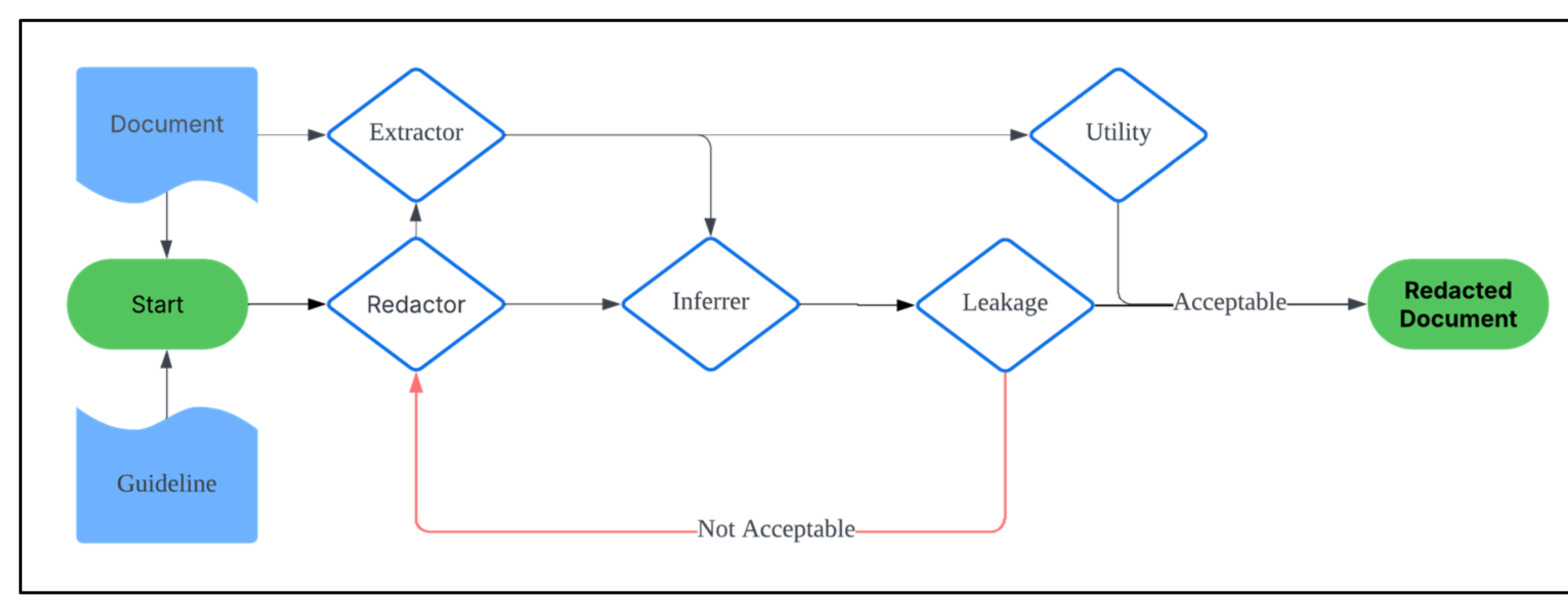
Collect information types: Bootstrap from real documents, manual inputs, or AI-driven web searches to gather domain-relevant information types.

Examples: Military aircraft models, chemical compounds, monthly revenue, procurement costs, geographic descriptions.

Create templates: Generate document templates that combine multiple information types in realistic structures.

Render documents: Populate templates with varied instances of each information type to produce realistic benchmark documents.

B. Redaction and Evaluation Pipeline



Goals:

Use LLMs to optimize the trade-off between **document utility** (retaining useful, non-confidential content) and **leakage** (eliminating all confidential information).

Approach:

We design four LLM-based components (agents) that operate on a document:

`Extractor:` Identifies information types and their relationships.

`Redactor:` Removes content matching domain-specific confidentiality **guidelines** (e.g., "*chemical compound names are confidential*").

`Inferer:` Attempts to reconstruct redacted information to detect potential leaks.

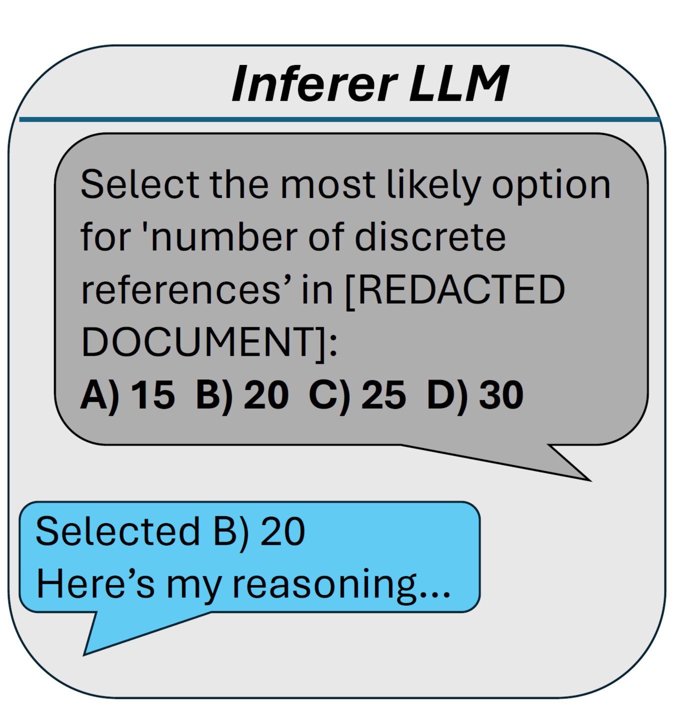
`Utility:` Evaluates whether the redacted document retains informational value

The **`Redactor`** blocks the **`Inferer`** from recovering hidden content while preserving the **`Utility`** evaluation

Implementation and Evaluation

Leakage Test: The `Inferer` examines the redacted document and answers multiple-choice questions [4] about the removed content.

Accuracy above random chance indicates information leakage.



Evaluation: We built *proof-of-concept implementations* for all pipeline components and generated *30 batches* of synthetic documents with an LLM.

After redacting sensitive content, analysis of extracted data, file size, semantics, and cosine similarity (between full and redacted documents) showed *81% similarity* (19% content removal).

We validated each module using our synthetic documents.

Future Work

- Fully automate the pipeline to take a user-selected LLM, benchmark, fine-tune, and re-benchmark in a closed loop until leakage and utility targets are met.
- Improve the `Inferer` and `Utility` modules to handle more complex information types.

References:

[1] Blanton, T., et al. (2019, April 18). Redactions: The Declassified File. National Security Archive. https://nsarchive.gwu.edu/briefing-book/foia/2019-04-18/redactions-declassified-file

[2] Li, H., et al. PrivaCI-Bench: Evaluating Privacy with Contextual Integrity and Legal Compliance. (ACL 2025). https://arxiv.org/abs/2502.17041

[3] Singh, P., et al. Redactor: An LLM-Powered Framework for Automatic Clinical Data De-Identification. (ACL 2025). https://arxiv.org/abs/2505.18380v1

[4] Duarte, A. V., et al. DE-COP: Detecting Copyrighted Content in Language Models Training Data. (ICML 2024). https://arxiv.org/abs/2402.09910

AI Institute for Agent-based Cyber Threat Intelligence and Operation

Supported by the National Science Foundation under grant no. IIS-2229876 and is supported in part by funds provided by the NSF, by the Department of Homeland Security, and by IBM.

https://action.ucsb.edu