Domain-Agnostic Text Redaction: Using Natural Language Rules and Instruction-Tuned Smaller Language Model

Whitepaper by

Dr. Aravindhan Arunagiri, Senior Manager, Mphasis NEXT Labs | Solution Architect
Ayaan Khan, Lead Data Scientist, Mphasis NEXT Labs | Solution Development
Sai Barath Sundar, Senior Manager, Mphasis NEXT Labs | Advisor
Dr. Udayaadithya Avadhanam, Principal & Vice President, Mphasis NEXT Labs | Advisor

Mphasis
The Next Applied

Contents

With the increasing digitization of personal and corporate communication, the automatic sanitization of textual data has become a crucial component of data privacy and compliance frameworks. Traditional text sanitization solutions are majorly suitable for obscuring sensitive data with regular patterns/structure such as Personal Identifiable Information (PII). These solutions are rule-based and do not provide an explanation for their redaction, which makes it difficult to audit them. This paper introduces Domain-Agnostic text redaction, an SLM-based natural language-driven text sanitization solution designed to identify and redact sensitive information in documents. Unlike traditional text sanitization, this method enables a user to conveniently define sensitive information in natural language which may be regular structured (E.g., PII) or non-regular structured information (E.g., legal terms and conditions). The solution reasons the rules of redaction from the user definition and redacts the sensitive content, while providing a transparent explanation for each redaction, highlighting specific rules that triggered the redaction decision. The solution comprises an SLM-based custom rule-engine that generates or augments NL rules of redaction (based on user definition of sensitive information) and use them for step-by-step contextual reasoning on any given document to identify and redact respective sensitive content. It generates NL explanations to support human reviewers and auditors in understanding why specific content is redacted. The solution comprises the evaluation of the redaction performance using a controlled set of sample documents with manually tagged sensitive information for quantifying the correctness of redaction during the testing phase. The reconstruction error metric shows the probability of reconstructing the sensitive information from the redacted document quantifying coverage of redaction. The solution shows high reconstruction error, signifying the best coverage of redaction and higher recall, ensuring the precision of redaction. The solution ensures privacy in critical applications that involve legal discovery, medical documentation and corporate information governance.

1. Introduction

As organizations increasingly rely on large-scale textual data for analytics, legal discovery and digital services, the need to protect sensitive information is a primal requirement. Data privacy laws such as the General Data Protection Regulation (GDPR), Health Insurance Portability and Accountability Act (HIPAA) and the California Consumer Privacy Act (CCPA) mandate strict controls over sensitive information (such as PII, commercial confidential information) residing in financial records, health data, legal document and other classified documents. In a document, sensitive information may exist along with non-sensitive information, which requires contextual knowledge (expert personal skill) to identify and sanitize them. Current text sanitization methods (either ML-based or LLM-based) are only capable of automatically identifying the sensitive content which has a regular pattern/structure such as email, telephone numbers, SSN and address. Microsoft (2023) Chen (2023). These methods cannot identify sensitive content that is rich in semantic information and has no regular pattern ((E.g., Medical – diagnosis, treatment; Business confidential information). Current methods firstly do not support the user to define the criteria of sensitivity of any content, and secondly, such criteria cannot be inferred on-the-fly by the redaction model to identify and redact that sensitive information. Such kind of context-dependent content sensitivity cannot be handled with existing methods. For example, in a contract document, the terms of the contract are sensitive in a specific business context. The user needs to define the criteria of sensitivity, say, “Do not show the terms of contract in the contract document” (Note: Though they are not per se sensitive information, for specific business requirement it is deemed sensitive). Existing solutions do not allow users to define on-the-fly, such content sensitivity, cannot automatically infer this definition and redact the content from the document. Owing to these limitations, in this white paper, a redaction solution is discussed that can allow users to custom define the content sensitivity (on-the-fly), and the model automatically infers the definition and redacts the content from data.

The model provides explanations in natural language that help the Responsible AI and Audit practice team to understand the decision of redaction. The performance of the model is evaluated with respect to correctness and coverage of redaction. The correctness refers to the capability of models to identify and redact correct sensitive information. The coverage shows the model’s capability to redact actual sensitive information while leaving non-sensitive information intact, thereby maintaining privacy and ensuring the utility of data Pilan (2025). The recent solution for de-identification of patients’ information in medical records requires the documents to be shared with the public API of GPT 4 and 3.5 API which affects privacy Altalla et al (2025). Hence, in this solution, a smaller language model (SLM) Phi 3.5 is instruct fine-tuned, that step-by-step understands the definition of sensitivity of any content,

infers rules of redaction, and later parses the given document to identify and redact the sensitive content. The usage of dedicated SLM improves the privacy of information and avoids costly GPT 4 API calls to redact information. This solution helps to maintain the privacy of the data subjects while giving complete control to the user to define and manage the privacy subjects. This feature gives users more flexibility to use the solution for various domain documents with little or no re-training overheads. The next section discusses the building blocks and functional operation of the solution highlighting the key performance indices maintained.

2. Functional Design and Operation of the Solution

The design involves a data ingestion process, Instruct fine-tuning of a Smaller Language Model (SLM) and deploying the same to make redaction on the input documents. The figure shows the basic building blocks of the redaction solution and presents their operations.

2.1 Functional blocks of solution

Data ingestion involves documents from various domains like business, medical, legal, etc. The user can input the data document and scenario, specifying the context and requirements with respect to the document. The data processing module involves cleaning the document for various formatting issues (such as spaces, blanks, page breaks, etc.) and chunking the content into fixed-length (small paragraph) chunks that are easily consumable by the SLM for inference.

The synthetic data generation module is used to create the training dataset, which comprises document content and the corresponding redaction rules. The module receives the document and scenario as input and creates synthetic content and rules of redaction that are applicable to the specific content and whole document in accordance with the scenario. The documents from different domains, such as legal, business, medical and scenarios, are used to generate training data sets, which are used to fine-tune the SLM for inference (redaction).

The Instruct fine-tuning process involves an SLM Phi 3.5, and the training data are generated using the Claude and GPT-4 models. The training data sets include domain documents and the set of rules of redaction that are applicable to each document. These rules of redaction and processed document content (with short paragraphs) are used to fine-tune the SLM. The tuned SLM can infer the rules of redaction, reason the rules on the input content and redact the relevant sensitive content while keeping intact the non-sensitive content. The SLM outputs the redacted document and the explanation of redaction in natural language.

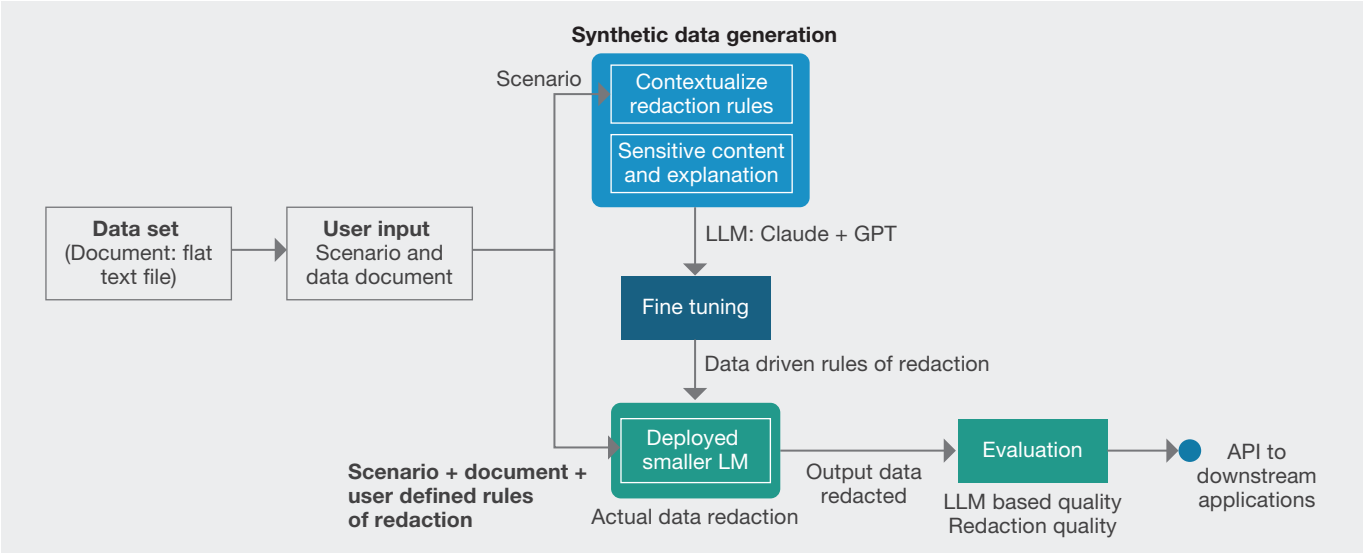

Figure 2.1: Framework for text sanitization using fine-tuned LLM

2.2 Scheme of operation of redaction solution

The fine-tuned SLM is deployed in the inference framework to make real-time redaction on the input documents. The user inputs the scenario which includes information regarding the document domain, business context, criteria of sensitivity and actual document. The user can give explicitly the rules of redaction, or the rules will be inferred from the criteria of sensitivity by SLM. The SLM outputs the redacted document devoid of the sensitive information requested by the user. The solution is completely automatic, and the user can evaluate the results in real time.

The solution is backed by evaluation metrices that quantify the correctness and coverage of the redaction. The reconstruction accuracy gives the coverage of redaction where the simulated adversarial attacks cannot reconstruct the actual document from the redacted document. This implies the document is completely sanitized of sensitive content. In testing, the reconstruction accuracy is less than 0.1 for the redacted document ensuring high resilience of the SLM towards Claude LLM-based inference attacks. This shows that Instruct fine-tuned Phi 3.5 model has highly scalable redaction skill, while performing equally in cross-domain document instances.

The solution performs correct redaction which is evaluated using the redaction recall metrics. The solution shows high recall in the testing with the ground truth document (synthetically generated using LLM). This metric captures the **correctness** of the model's redaction capability and is essential for determining whether the model is omitting any content that should have been removed owing to privacy requirement. A high Redaction Recall indicates that the model is effectively capturing the full scope of sensitive content that needs to be redacted. Conversely, a low value reflects under-redaction and raises concerns about residual privacy risk due to missed entities (content).

These metrices are particularly important in real-world settings, where missing even a single sensitive token can lead to a privacy breach. The fine-tuned SLM can redact sensitive content over 95% of the sensitive instances in the input document. In this article, various measures are used to ensure the correctness and coverage of redaction. This ensures the data privacy preservation productivity of the SLM, especially in large volume documents.

3. Summary

The solution offers state-of-the-art capability to handle (identify and redact) non-regular structured information (E.g., legal clauses, medical information) in contrast to current text sanitization solutions that can handle only regular structured information (E.g., PII). The critical contribution of this solution includes enabling the user to custom define any sensitive content which may be a legal clause, medical therapy, business contract deals and the LLM engine automatically generates rules of redaction to redact the sensitive content from the document. The SLM engine reasons the rules of redaction step-by-step on any given document to identify and redact the sensitive content. The SLM engine is trained with a synthetically generated sample data set from various domains such as Legal, Business, Medical. The usage of synthetic data sets to train the SLM engine mitigates any privacy attack (inference attacks) on the SLM engine whereby maintaining the data privacy during training and inference stages. The SLM engine shows a higher degree of redaction performance with the well-formatted document. In cases of ill-formatted documents (with large blanks and spaces) the solution comprises the data preprocessing pipeline that makes the data amenable for inferencing. The solution uses two dedicated metrices to evaluate the correctness and coverage of the redaction. In this article, customized manually redacted data sets are used to validate the testing performance of redaction. The reconstruction metrics give the coverage performance of the SLM engine through entire document. The reconstruction metric uses the reconstruction attack on the redacted document evaluating if any part of the redacted content can be constructed. The solution shows high reconstruction errors which indicate good redaction coverage. The SLM engine can also be used to redact the other domain documents (those not used in training) with a minimal number of few shots learning and/or fine-tuning. The solution is versatile for redacting any free-text sensitive content that can be custom defined by the user on-the-fly and the SLM engine can automatically redact it. The automated data redaction improves the data privacy productivity of sanitizing large documents with less manual intervention, high accuracy and utility.

4. References

1. Microsoft. (2023). Presidio: An open-source data protection and anonymization tool. Retrieved from https://microsoft.github.io/presidio/
2. Ildikó Pilán and Pierre Lison and Lilja Øvrelid and Anthi Papadopoulou and David Sánchez and Montserrat Batet (2022). The Text Anonymization Benchmark (TAB): A Dedicated Corpus and Evaluation Framework for Text Anonymization. arXiv. Eprint 2202.00443. https://arxiv.org/abs/2202.00443.
3. Daniel Lopresti and A. Lawrence Spitz (2004). Quantifying information leakage in document redaction. In Proceedings of the 1st ACM workshop on Hardcopy document processing (HDP '04). Association for Computing Machinery, New York, NY, USA, 63–69. https://doi.org/10.1145/1031442.1031452
4. Ruisi Zhang, Seira Hidano, Farinaz Koushanfar (2022). Text Revealer: Private Text Reconstruction via Model Inversion Attacks against Transformers. arXiv.eprint 2209.10505, https://arxiv.org/abs/2209.10505
5. Altalla', B., Abdalla, S., Altamimi, A. et al. (2025). Evaluating GPT models for clinical note de-identification. Sci Rep 15, 3852. https://doi.org/10.1038/s41598-025-86890-3
6. Carol El-Hayek, Siamak Barzegar, Noel Faux, Kim Doyle, Priyanka Pillai, Simon J. Mutch, Alaina Vaisey, Roger Ward, Lena Sanci, Adam G. Dunn, Margaret E. Hellard, Jane S. Hocking, Karin Verspoor, Douglas IR. Boyle. 2023. An evaluation of existing text de-identification tools for use with patient progress notes from Australian general practice, International Journal of Medical Informatics, Volume 173, 105021, ISSN 1386-5056, https://doi.org/10.1016/j.ijmedinf.2023.105021.
7. Ildikó Pilán, Benet Manzanares-Salor, David Sánchez, Pierre Lison (2025). Truthful text sanitization guided by inference attacks, Applied Soft Computing, Volume 185, Part B, 2025, 114013, ISSN1568-4946, https://doi.org/10.1016/j.asoc.2025.114013. (https://www.sciencedirect.com/science/article/pii/S1568494625013262).
8. Veena Vasudevan, Ansamma John (2014). A Review on Text Sanitization. International Journal of Computer Applications. 95, 25 (June 2014), 14-17. DOI=10.5120/16749-6916

About Mphasis

At Mphasis, engineering has been in our DNA since inception.

Mphasis is an AI-led, platform-driven company with human-in-the-loop intelligence, helping global enterprises modernize, infuse AI, and scale with agility. The Mphasis.ai unit and Mphasis AI-powered 'Tribes' are focused on client outcomes and embed artificial intelligence and autonomy into every layer of the enterprise technology and process stack. Mphasis built NeoIP™, a breakthrough AI platform which orchestrates a powerful pack of AI platforms and solutions to deliver impactful outcomes across the entire enterprise IT value chain, because we believe 'AI Without Intelligence Is Artificial™.' NeoIP™ is powered by the Ontosphere, a dynamic and ever-evolving knowledge base, delivering continuous and constant innovation through perpetual intelligent engineering - driving end-to-end enterprise transformation.

At the heart of our approach is customer-centricity—reflected in our proprietary Front2Back™ transformation framework, which uses the exponential power of cloud and cognitive to deliver hyper-personalized digital experiences ($C = X2C^{2}_{TM} = 1$) and build strong relationships with marquee clients. Our Service Transformation solutions enable enterprises pivot from legacy systems and operations to secure, adaptive, cloud-first operating models with minimal disruption. Continuous investments in platforms, such as the Neo series, enable enterprises to stay efficient, relevant, and ahead in a dynamic AI-first world. Mphasis is a Hi-Tech, Hi-Touch, Hi-Trust company, rooted in a learning and growth culture. Click here to know more. (BSE: 526299; NSE: MPHASIS)

For more information, contact: marketinginfo.m@mphasis.com

USA
Mphasis Corporation
41 Madison Avenue
35th Floor, New York
New York 10010, USA
Tel: +1 (212) 686 6655

UK
Mphasis UK Limited
1 Ropemaker Street, London
EC2Y 9HT, United Kingdom
T : +44 020 7153 1327

INDIA
Mphasis Limited
Bagmane World Technology Center
Marathahalli Ring Road
Doddanakundhi Village, Mahadevapura
Bangalore 560 048, India
Tel.: +91 80 3352 5000

www.mphasis.com

VAS 22/02/26 US LETTER BASIL 9883