UKRAINIAN CATHOLIC UNIVERSITY

BACHELOR THESIS

Comparative Analysis and Framework Development of Anonymization Techniques for Large Language Models

Author:
Liliana HOTSKO

Supervisor:
Dr. Taras FIRMAN

A thesis submitted in fulfillment of the requirements for the degree of Bachelor of Science

in the

Department of Computer Sciences and Information Technologies
Faculty of Applied Sciences

Lviv 2025

"*There are no solutions. There are only trade-offs.*"

Thomas Sowell

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Bachelor of Science

Comparative Analysis and Framework Development of Anonymization Techniques for Large Language Models

by Liliana HOTSKO

Abstract

Large language models (LLMs) are becoming valuable tools for automating text-based data workflows. However, organizations operating under data protection laws like the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) often cannot adopt these systems due to concerns that input data and model outputs might be stored or reused by third-party providers. This risk limits organizations to using only local LLMs, which are often less efficient. To address this, we built a modular framework that anonymizes sensitive data before it reaches external models. The Internal pipeline includes multiple layers: pattern-based masking, transformer-based named entity recognition (NER), local adversarial models, context-aware masking, secure vector storage, and post-processing. In addition to the framework, we developed an interactive interface that allows the modification and testing of the pipeline configuration before integration. We propose various pipeline setups and analyze their effectiveness in masking private information, the Retrieval-Augmented Generation (RAG) quality, and trade-offs involved in applying masking steps, both in English and Ukrainian languages. Our results show that the fine-tuned general-purpose models can achieve 82% masking accuracy for Ukrainian (using RoBERTa) and 92% for English (using DistilBERT). Context-aware masking significantly improves the quality of generated answers over standard categorical masking — raising the baseline accuracy from 40–50% to nearly 98% in culturally dependent cases. These findings provide the necessary information to create the best anonymization setup for custom needs. Organizations or individual users can further integrate our framework, open-sourced at https://github.com/NaniiiGock/LLMAnonymizationThesis, into their applications, use LLMs while remaining compliant with privacy standards, and benefit from text processing automation.

Acknowledgements

I would like to express my gratitude to my supervisor, Dr. Taras Firman, for his guidance, support, and insights throughout this work. I am also thankful to ELEKS for an amazing journey that deepened my interest in LLMs and motivated me to set the highest personal goals. Finally, I am grateful to my friends, family, and all people who accompanied me during my years of study, supporting my growth and helping me make important decisions.

Contents

List of Figures

List of Tables

List of Abbreviations

LLM	**L**arge **L**anguage **M**odel
GenAI	**G**enerative **A**rtificial **I**ntelligence
PII	**P**ersonally **I**dentifiable **I**nformation
NER	**N**amed **E**ntity **R**ecognition
RAG	**R**etrieval-**A**ugmented **G**eneration
HE	**H**omomorphic **E**ncryption
CAM	**C**ontext-**A**ware **M**asking
ANN	**A**pproximate **N**earest **N**eighbor
HNSW	**H**ierarchical **N**avigable **S**mall **W**orld
GDPR	**G**eneral **D**ata **P**rotection **R**egulation
HIPAA	**H**ealth **I**nsurance **P**ortability and **A**ccountability **A**ct

List of Symbols

x_t	Token at position t in an input sequence
y_t	Label assigned to token x_t
$\mathcal{E}$	Set of entity labels (e.g., PERSON, LOCATION)
$\mathcal{X}$	Set of all possible token sequences
$\mathcal{D}$	Corpus used for training the LLM
T	Length of the token sequence x
θ	Parameters of the language model
f_θ	Neural network function (e.g., Transformer) with parameters θ
P_θ	Conditional probability distribution modeled by the LLM
x^*	Memorized sequence extracted from training data
q	Prompt that triggers memorized content in a model
$\text{Risk}_{\text{reid}}$	Probability of successful re-identification
$\mathcal{U}$	Set of users or individuals in the data
$A(x)$	Anonymization function applied to input x
$D(x)$	Detection function for sensitive spans
$M(\cdot)$	Masking or pseudonymization function
$E(\cdot)$	Enhancement function (e.g., synthetic text generation)
$\mathbf{v}_x$	Embedding vector of the original input sequence x
$\mathbf{v}_{A(x)}$	Embedding vector of the anonymized sequence $A(x)$
$\text{Sim}(x, A(x))$	Cosine similarity between original and anonymized embeddings
α	Trade-off parameter between privacy and utility loss
$\mathcal{L}_{\text{total}}$	Combined loss function
$\mathcal{L}_{\text{privacy}}$	Loss associated with privacy leakage
$\mathcal{L}_{\text{utility}}$	Loss associated with task performance degradation
$\vec{e}_i$	Embedding vector of the i-th document
$\vec{q}$	Embedding vector of the query
$\text{Enc}(\cdot)$	Homomorphic encryption function
$\mathcal{E}(\cdot)$	Encryption scheme (e.g., CKKS)
$\mathcal{D}(\cdot)$	Decryption function
O	Label indicating a non-entity token

Dedicated to the future of secure and ethical AI

Chapter 1

Introduction

Large language models (LLMs) are becoming more common for automating the processing of textual data. Since the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) restrict organizations from sharing personal data, they are limited to using less effective local models. This work aims to develop a solution that benefits such organizations by anonymizing personally identifiable information (PII) before sending it to external models.

1.1 Motivation

PII Detection and Masking

LLMs are widely used to optimize data processing. PII, such as names, dates, locations, and job titles can be used by third parties when raw text is processed by generative models or stored in cloud vector storage. This raises concerns about data protection, compliance with privacy regulations (e.g., GDPR, HIPAA), and ethical AI usage. Therefore, there is a need for reliable and adaptable PII detection mechanisms for anonymization pipelines and methods that support effective contextually-aware masking.

Framework development

In addition to the need for effective PII detection, there is a growing demand for end-user tools that integrate customizable anonymization techniques with LLM-based applications. Most publicly available LLM anonymization applications do not provide wide privacy-aware configurations or support local deployments, which makes them unsuitable for use in organizations with strict data compliance requirements. In addition, privacy workflows often lack transparency and control. This thesis aims to address these gaps by developing a configurable and modular framework that allows users to process sensitive documents safely, train custom NER models, and apply anonymization strategies according to their needs through web-based UI and code-level interfaces.

1.2 Contributions

Comparison of PII masking techniques. We present an evaluation of multiple PII anonymization approaches, including regex-based masking, transformer-based NER models, LLM-based NER, and LLM-based context-aware replacements. Additionally, we analyze how context-aware masking influences bias in LLM outputs in culturally dependent fields. The methods are compared by the accuracy scores of the

NER models and RAG evaluation metrics. We evaluate our proposed pipelines using textual data in both English and Ukrainian.

Development of the Framework for anonymized LLM usage. We provide a modular framework for using LLMs on sensitive texts. We use locally hosted models for LLM-based NER with a two-model adversarial setup and for implementing three levels of context-aware anonymization. We also implement homomorphic encryption to perform private document retrieval in the cloud vector database. The framework includes a user interface for configuration of the masking parameters and pipeline testing, as well as scripts for custom model training. Interactive UI and integration options into existing applications make our system flexible and easy to use.

1.3 Structure Of The Thesis

- **Chapter 2: Related Work** — This chapter reviews previous research and public tools for PII anonymization, NER methods, LLMs bias in culturally dependent responses, the use of LLMs in sensitive domains, gaps in existing works, and the contribution of our research.
- **Chapter 3: Theoretical Background** — This chapter provides an overview of the technical foundations relevant to this work. Additionally, it outlines our work's used evaluation metrics and the regulatory context of our work.
- **Chapter 4: Proposed Solution** — This chapter describes the design and structure of the proposed framework, including its components, configuration logic, and user interface.
- **Chapter 5: Experiments and Results** — This chapter presents quantitative and qualitative evaluations of the anonymization techniques and the framework's performance across multiple configurations.
- **Chapter 6: Conclusions** — This chapter summarizes the work done, discusses its limitations, and outlines possible improvements and directions for future research.

Chapter 2

Related Work

Due to the increasing importance of data protection regulations, organizations and researchers are developing various anonymization techniques to use LLMs securely in their textual data processing. This chapter provides an overview of existing anonymization methods and supporting technologies, highlighting their strengths and limitations, and identifying gaps in the literature that this research aims to address.

2.1 Related approaches

Public Anonymization Tools

Several open-source and public anonymization frameworks are available to support privacy-preserving text processing, and Microsoft's Presidio [1] is one of the most widely adopted among them. It offers a modular and customizable pipeline for detecting and anonymizing personally identifiable information (PII). Recent work has shown that despite its flexibility, Presidio and similar tools often lack contextual awareness. Comparisons with large language model-based anonymizers demonstrate that LLMs can offer more context-sensitive masking strategies [2].

Retrieval-Augmented Generation(RAG) vs Fine-Tuning

Recent studies have compared the trade-offs between fine-tuning LLMs and augmenting them with external knowledge through RAG to adapt them to specific fields. While it has been shown that the combined approach performs best, RAG alone gives nearly the same results [3]. Other research works highlight that there is no best universal solution [4, 5], but RAG mostly outperforms fine-tuning for small models [6]. Considering these benchmarks, we use RAG in our framework.

Transformer-Based Named Entity Recognition

Transformer architectures such as BERT and RoBERTa significantly outperform traditional sequence labeling methods in both accuracy and generalization [7]. Transformer-based NER is also recognized as a better approach than non-transformer baselines in multiple domains [8], and it achieves high performance for various entity types and tasks [9].

LLM-Based NER

There is growing interest in using general-purpose LLMs for NER and de-identification tasks. When NER is a single step to recognize the private entities within the text, de-identification ensures that no PII can be restored from the masked text based on NER. Evaluation on clinical data demonstrates that even the best domain-specific transformer-based models leave 9% of patient information that can be re-identified by LLM [10]. Therefore, adversarial anonymization was introduced to ensure that the final LLM cannot restore the contents from processed text [2, 11, 12]. To improve the efficiency of adversarial threats, another approach adds a k-anonymity step, which significantly shortens the processing time [13].

OpenAI Dominance and Bias Concerns

A survey [14] shows that GPT models dominate enterprise use cases (Figure A.1) due to their strong performance, developer accessibility, and seamless API integrations. Our framework aims to replicate the performance of commonly used LLMs. Therefore, we compare our results to a baseline using pure GPT-4o.

However, other studies [15, 16, 17, 18] find that large language models — including those by OpenAI — tend to align with Western cultural norms and values, even if they perform well on multilingual tasks. These biases are particularly relevant to our anonymization pipelines, where cultural assumptions may affect OpenAI responses while processing masked texts.

Context-Aware Masking

To make the real-world anonymization of personal data more effective and less dependent on OpenAI cultural bias, we observed context-aware masking. Some works use pseudonymization techniques, where context sensitivity improves masking accuracy [19]. Alternative sentence versions can also be generated and evaluated using MSE scoring to select the best candidate for further usage [20]. Synthetic data generation with LLMs is effective, as no specific fine-tuning is needed to provide realistic replacements, and in our research, we use this approach to handle potential biases. Masked texts can be further applied in downstream tasks, such as classification and information retrieval [21], which we implement in our framework.

Homomorphic Encryption

We explored Homomorphic Encryption(HE) as a solution for privacy-preserving operations on language model embeddings. One approach uses CKKS encryption to perform classification directly over encrypted BERT embeddings, demonstrating that secure computation is feasible in practical NLP tasks [22]. In our work, we similarly apply encryption to store text embeddings in a cloud-based vector database and support the similarity search, which is required for RAG. The high computational complexity of HE can lead to increased processing times and costs, making it less suitable for real-time applications [23, 24].

RAG Evaluation with RAGAS

We integrate a benchmark framework, RAGAS [25], to evaluate our retrieval-augmented generation pipelines based on retrieved context, and to determine how different anonymization strategies affect RAG output quality. When using RAGAS, we can validate whether anonymized inputs maintain downstream utility based on context relevance and factual consistency metrics.

2.2 Gaps and Inconsistencies in Existing Research

Existing studies focus on individual anonymization techniques, such as NER, synthetic data generation, adversarial anonymization, and differential privacy, without integrating them into a modular framework suitable for flexible applications. First, they also prioritize PII detection but lack context sensitivity analysis. Although context-aware masking approaches have been proposed [19, 20], they were not integrated widely into the anonymization pipelines to mitigate domain-specific and cultural bias of LLMs [16, 17]. In addition, when RAG is considered as a strong alternative to fine-tuning LLMs, there is a gap in studying the impact of anonymization techniques on RAG performance metrics such as faithfulness and answer relevancy. Finally, studies like DeID-GPT [11] and GPT-4-based de-identification rely on closed-source APIs, making them unsuitable for adversarial threats in regulated environments (e.g., GDPR, HIPAA). There is limited work using local, open-weight models (like LLaMA, Mistral) for de-identification and adversarial risk analysis [26, 27]

Chapter 3

Theoretical Background

In this chapter, we introduce the key concepts used in our work. First, we discuss what personally identifiable information is and how third parties can expose it. Second, we outline categories of techniques implemented in our framework, including pipeline composition and evaluation metrics. Third, we explain the regulatory context and compare the features of several global privacy regulations.

3.1 Personally Identifiable Information (PII)

PII refers to any data that can be used to identify an individual, either directly or indirectly. Examples of PII include a person's name, residential address, phone number, email address, date of birth, social security number, passport number, financial account details, or medical identifiers. In computational models, different types of PII can be categorized into a set of predefined *entity labels*, denoted as $\mathcal{E} = \{e_1, e_2, ..., e_m\}$, where each e_i corresponds to a specific entity class.

The task of identifying PII within a text sequence is typically formalized as a **Named Entity Recognition (NER)** problem. NER serves as a foundational component in anonymization pipelines, as it identifies the sensitive entities that must be masked, removed, or replaced to ensure data privacy. Given an input sequence of tokens $x = (x_1, x_2, ..., x_T)$, where T is the length of the sequence, the goal is to assign a corresponding sequence of labels $y = (y_1, y_2, ..., y_T)$, such that:

$$y_t \in \mathcal{E} \cup \{\mathrm{O}\}, \quad \forall t \in \{1, ..., T\}$$

Where x_t is the token at position t in the sequence, y_t is the label assigned to x_t, $\mathcal{E}$ is the set of entity labels corresponding to identifiable information, and O denotes a non-entity token, i.e., a token that does not belong to any PII class.

3.2 Large Language Models and Data Privacy

LLMs are typically parameterized by billions of weights and trained via next-token prediction using a corpus $\mathcal{D} = \{x_1, x_2, ..., x_n\}$, where $x_i \in \mathcal{X}$ is a token sequence from natural language text. The goal is to learn the conditional probability distribution over tokens:

$$P_\theta(x_t \mid x_{<t}) = \mathrm{softmax}(f_\theta(x_{<t}))$$

Where θ are model parameters and f_θ is the learned representation function (e.g., a Transformer).

However, models trained on sensitive data may memorize and reproduce rare sequences, enabling memorization attacks where an adversary queries the model to recover sensitive strings from training data. Therefore, a memorization attack attempts to find a sequence $x^* \in \mathcal{D}$ such that:

$$x^* = \arg\max_{x \in \mathcal{X}} P_\theta(x \mid q)$$

for some prompt q that triggers the memorized content. This risk becomes critical in domains where privacy violations can lead to legal and ethical consequences.

3.3 Disadvantages of Local LLMs

While the use of local large language models (LLMs) offers clear benefits regarding data privacy, offline accessibility, and cost control, it also has several limitations. Their context length is more limited, making them less suitable for processing large documents or handling multi-turn interactions. Additionally, local deployments often lack built-in access to external tools or up-to-date knowledge sources. These constraints limit the scalable and accurate understanding of language systems. As a result, local LLMs may be used for domain-specific tasks, but they cannot provide the required performance and quality in broader fields due to their simpler architecture.

3.4 Threat Models in LLM Privacy

Several threat models exist in the context of LLM anonymization, such as training data extraction through model inversion, prompt leakage via logging, storing, or interception, and embedding leakage, when vector representations can be correlated with PII. The re-identification risk can be formalized as:

$$\text{Risk}_{\text{reid}} = \Pr(\exists u \in \mathcal{U} : \text{Identify}(A(x)) = u)$$

Where $\mathcal{U}$ is the set of individuals, and $\text{Identify}(\cdot)$ is an adversarial re-identification algorithm. Therefore, anonymization techniques are necessary on the prompt level and in embedding.

3.5 Categories of Anonymization Techniques

Several approaches to anonymization have been proposed in related works, including regex-based masking, NER-based anonymization, and Embedding-level anonymization. **Regex-based masking** applies deterministic pattern matching to remove entities, **NER** detects the entity spans of PII, and **Embedding-level anonymization** applies transformations such as Homomorphic Encryption to reduce leakage from vector representations.

Homomorphic Encryption (HE) enables computations to be performed directly on encrypted data without decrypting it first. In the context of RAG pipelines, HE can be applied during the storage step, allowing document embeddings to be encrypted before being sent to a vector database or blob storage. This approach ensures that raw embedding vectors — which may encode sensitive information — remain encrypted and inaccessible, even to the storage provider. The typical pipeline involves encrypting each embedding $\vec{e_i}$ using a homomorphic encryption scheme such as CKKS (Cheon–Kim–Kim–Song)[28], resulting in $\text{Enc}(\vec{e_i})$, which is then stored. When a query is received, it is embedded and encrypted similarly, and a secure similarity score (e.g., dot product) is computed locally between the encrypted query and each stored ciphertext embedding. Given a plaintext embedding vector $\vec{e_i} \in R^d$, the homomorphic encryption scheme $\mathcal{E}$ is applied to obtain the ciphertext:

$$\text{Enc}(\vec{e_i}) = \mathcal{E}(\vec{e_i})$$

When a query $\vec{q} \in R^d$ is issued, it is embedded and encrypted as:

$$\text{Enc}(\vec{q}) = \mathcal{E}(\vec{q})$$

The system then computes a similarity score, such as the encrypted inner product, directly in the ciphertext space:

$$\text{Enc}(\vec{q} \cdot \vec{e_i}) = \text{Enc}(\vec{q}) \cdot \text{Enc}(\vec{e_i})$$

The top-k results are then selected based on the decrypted similarity scores:

$$\vec{q} \cdot \vec{e_i} \approx \mathcal{D}(\text{Enc}(\vec{q}) \cdot \text{Enc}(\vec{e_i}))$$

Where $\mathcal{D}$ denotes the decryption function. Although this approach introduces computational overhead and limits the use of traditional vector indexing structures (e.g., HNSW - Hierarchical Navigable Small World graph), it provides a strong privacy guarantee and is particularly valuable in scenarios where client-side confidentiality is paramount.

3.6 Search Algorithms

HNSW (Hierarchical Navigable Small World) is a graph-based algorithm for Approximate Nearest Neighbor (ANN) search in high-dimensional spaces. This method is used in vector stores such as ChromaDB, FAISS (Facebook), Weaviate, etc. Instead of scanning all stored embeddings (which is slow for large datasets), HNSW organizes them into a graph structure that allows efficient traversal to quickly find the most relevant matches. HNSW is considered a practical choice for maintaining high recall when searching for top-k similar entries in RAG pipelines for processing medium-large datasets.

3.7 Anonymization Pipeline and Semantic Preservation

A generic anonymization pipeline can be decomposed into functional stages:

$$A(x) = E(M(D(x)))$$

Where $D(x)$ denotes a detection of sensitive spans (e.g., via NER or Local LLM), $M(\cdot)$ is a masking or pseudonymization of entities, and $E(\cdot)$ is the optional enhancement via synthetic generation, adversarial anonymization, or fine-tuning. In high-utility environments, semantic coherence must be preserved in post-anonymization. This is often measured via embedding similarity, using cosine distance between the original and anonymized text embeddings:

$$\text{Sim}(x, A(x)) = \frac{\mathbf{v}_x \cdot \mathbf{v}_{A(x)}}{\|\mathbf{v}_x\| \cdot \|\mathbf{v}_{A(x)}\|}$$

Where $\mathbf{v}_x$ is the embedding vector of the original input sequence x, and $\mathbf{v}_{A(x)}$ is the embedding vector of the anonymized sequence $A(x)$. In the context of cultural bias in GPT models, a high similarity score between the original input and anonymized input indicates a lower probability of receiving Western-oriented LLM outputs.

3.8 Evaluation Metrics

Effective anonymization must balance privacy protection and data utility. This trade-off can be modeled as a combined loss function:

$$\mathcal{L}_{\text{total}} = \alpha \cdot \mathcal{L}_{\text{privacy}} + (1 - \alpha) \cdot \mathcal{L}_{\text{utility}}$$

Where $\alpha \in [0, 1]$ controls the prioritization of privacy versus utility. Standard evaluation metrics include privacy metrics, such as re-identification rate and residual risk, utility metrics (classification accuracy and task F1 score), semantic metrics computed with embedding similarity scores, and efficiency metrics (processing latency and scalability).

3.9 Regulatory Context

Data privacy regulations vary significantly across jurisdictions, with some countries enforcing strict localization and cross-border data processing restrictions. The European Union's **General Data Protection Regulation (GDPR)** [29] provides one of the most comprehensive privacy frameworks globally, emphasizing user consent, data minimization, and transfer limitations. Several sector-specific regulations impose additional restrictions. The **Health Insurance Portability and Accountability Act (HIPAA)** [30] in the United States governs the handling of protected health information (PHI), requiring that healthcare data be stored, transmitted, and accessed only under strict safeguards. Similarly, the Family Educational Rights and Privacy Act (FERPA) restricts the use and disclosure of educational records. The financial sector is regulated under laws such as the Gramm-Leach-Bliley Act (GLBA), which mandates safeguards for consumer financial data.

Some privacy regulations prohibit data sharing similarly or even more strictly. China's Personal Information Protection Law (PIPL) [31] requires that sensitive personal data must be stored locally and cannot be transferred abroad without a security assessment and government approval. In the Middle East, Saudi Arabia's Personal Data Protection Law (PDPL) [32] enforces data localization and prohibits cross-border data transfers. The United Arab Emirates follows a hybrid model: while its federal law aligns with international best practices, it's financial free zones, such as DIFC [33] and ADGM [34], have adopted GDPR-like rules with their regulatory authorities.

These differences in legal environments significantly affect the deployment of LLM-based systems, particularly for organizations handling personally identifiable information. We aim to provide the framework with local processing to ensure compliance with most of the data protection laws.

Chapter 4

Proposed Solution

In this chapter, we formulate our problem, propose the anonymization framework structure, integrated modules, and describe the user interface for configuration settings.

4.1 Problem Formulation

Large language models are increasingly used across organizations to automate text-based data processing. However, due to concerns about data privacy, their adoption is limited in fields where confidentiality is critical. Businesses often restrict LLM usage to only a subset of tasks to avoid exposing sensitive information or use local LLMs, which cannot support large-scale complex tasks. Therefore, there is a necessity for an anonymization application for personalized private data masking before sending it to the public LLM provider or cloud-based vector database.

This thesis addresses the challenge of enabling the effective use of LLMs while staying compliant with data privacy regulations. In this chapter, we propose a highly-configurable anonymization framework, the user interface for setting the parameters for each pipeline step, and testing environment.

4.2 Framework Structure

Our framework is built around several core components, each designed to cover a specific aspect of the anonymization process. These components work modularly and flexibly, allowing various configurable settings. The architecture can be adjusted based on whether anonymization preprocessing is performed before or after data storage, as illustrated in Figure 4.1 and Figure 4.2. The user is provided with an interactive page to create the configuration file according to their personal needs, test the created pipeline with their settings, and use the created file for framework integration. In this section, we describe the modules implemented within our framework.

4.2.1 Preprocessing

To filter private or sensitive information from the text, we first need to detect which entities should not be passed to the LLM. For this task, three approaches of Named Entity Recognition are implemented: Pattern NER, Transformer-based NER, and LLM-based NER. **Pattern-based NER** identifies deterministic entities such as emails, phone numbers, social security numbers, and credit card details. The user can update the list of patterns that can appear within the contents to identify well-structured entities efficiently without the need for large-scale models. **Transformer-based NER** identifies PII entity spans by labeling the tokenized texts. For this task, we use native and fine-tuned models from Spacy, a DistilBERT-based classification model, and

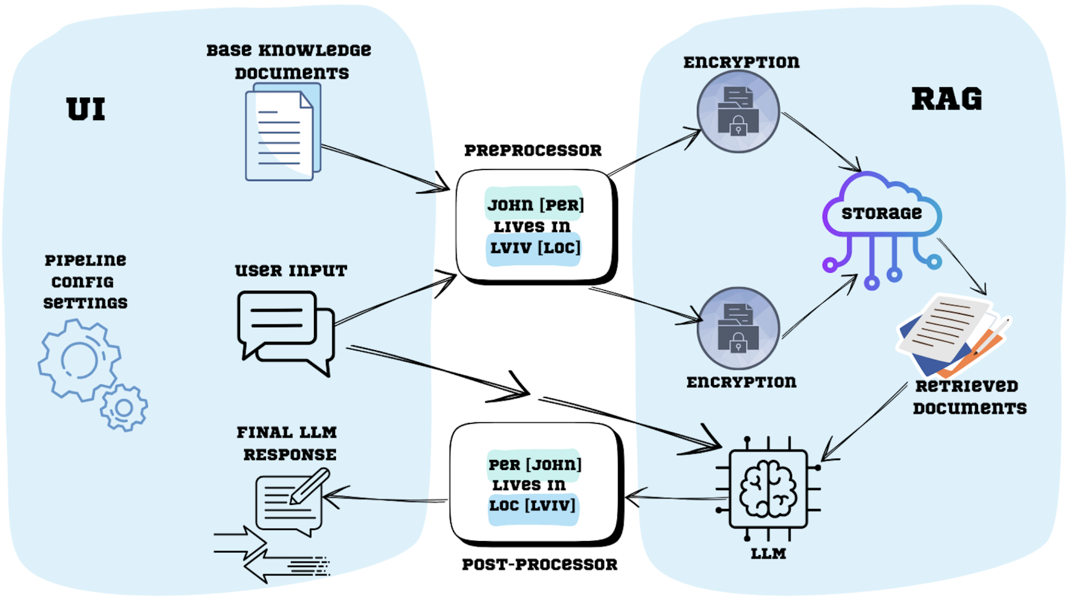

FIGURE 4.1: Framework Architecture with preprocessing before storing

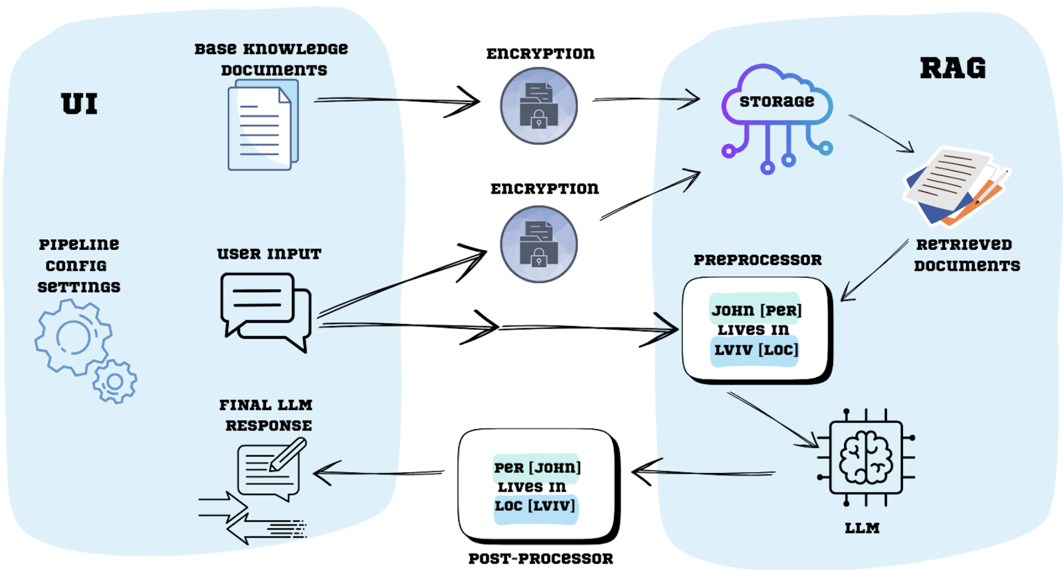

FIGURE 4.2: Framework Architecture with preprocessing after storage

a RoBERTa-based classification model. For **LLM-based NER**, lightweight local LLaMa and Mistral models are integrated to find the PII entities and mask them out of the input. These local LLMs are also used in context-aware processing on top of the categorical masking.

Spacy Models

We integrate pre-trained models from Spacy that are designed for general-purpose NER tasks. For English, we use the **"en-core-web-sm"** model, and **"uk-core-news-sm"** for Ukrainian. Additionally, we fine-tuned these models using datasets described in section 4.2.1.Datasets. The internal pipelines of these models include:

- **Tok2Vec** - the initial step where the text is tokenized into words or subword units, and word vectors are generated.
- **Tagger** assigns part-of-speech (POS) tags to each token. This step helps identify the grammatical structure of the sentence, such as whether a token is a noun, verb, adjective, etc.
- **Morphologizer** is a unique component for Ukrainian text processing(and is used instead of tagger in English). The morphologizer handles the complex inflectional morphology of the Ukrainian language. It assigns the morphological features of words, such as gender, number, case, and tense, which are essential for understanding the syntactic structure of the language. This step is especially important for Ukrainian because of the language's rich inflectional system.
- **Parser** analyzes the syntactic structure of the sentence by constructing a dependency parse tree. This helps the system understand how different tokens are related to each other syntactically. The parser in the Ukrainian model is used similarly, but it is specifically designed to handle the syntactic properties of the Ukrainian language, including its word order and case marking.
- **Attribute-Ruler** modifies specific attributes of the tokens (such as POS tags or syntactic dependencies) based on rules. It is used for fine-tuning and customizing token attributes.
- **Lemmatizer** normalizes tokens by reducing them to their base form (lemma).
- **NER** is the final step, which involves identifying named entities in the text, such as names, organizations, locations, dates, etc.

BERT for Token Classification (English)

We fine-tuned a DistilBERT-based token classification model (DistilBertForTokenClassification) [35] for the named entity recognition task. DistilBERT is a distilled version of BERT that retains 97% of BERT's language understanding capabilities while being significantly smaller and faster (model architecture: Figure 4.3). The backbone of the model consists of a 6-layer transformer encoder with multi-head self-attention, layer normalization, and feedforward sublayers, all operating over 768-dimensional hidden representations. The token classification head is a linear layer mapping each token's contextual embedding to an output distribution over the NER label space. The model includes standard dropout regularization ($p = 0.1$) and uses GELU activation in the feedforward layers. The final classifier outputs a probability distribution over the entity classes.

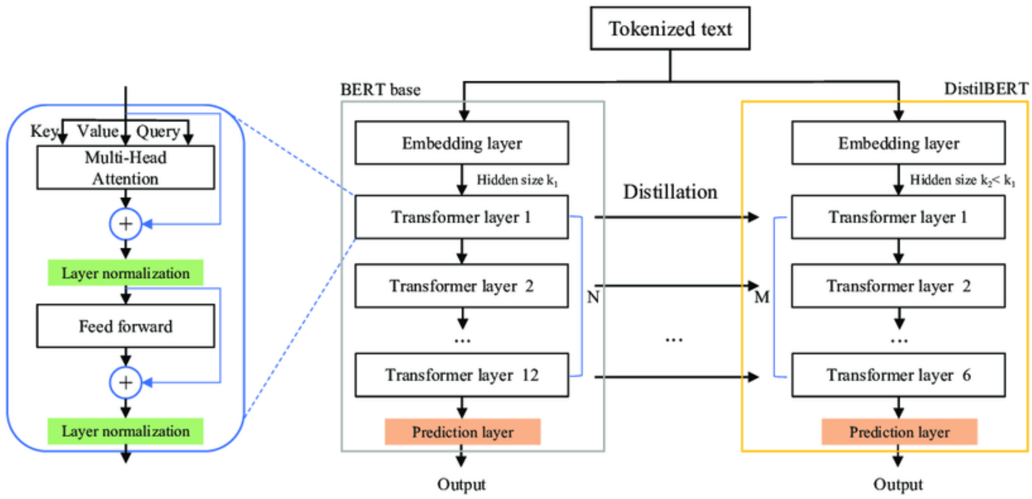

FIGURE 4.3: The DistilBERT model architecture and components. [35]

RoBERTa for Token Classification (Ukrainian)

We train the YouScan-UKR-RoBERTa-base pre-trained model [36] with Roberta For Token Classification [37] for the NER task on the Ukrainian language. It is built on the RoBERTa encoder, which consists of 12 Transformer layers with a hidden size 768. RoBERTa uses Byte-Pair Encoding (BPE) tokenization, which is better than WordPiece (used in BERT) at capturing subword units in Slavic languages.

A linear classification head is added to the final hidden states to predict a label for each input token. Input representations are formed using word and position embeddings (without segment embeddings), followed by multi-head self-attention, feed-forward layers, and residual connections. In addition, we apply a lemmatization step to the mapping creation process when the text needs to be split into chunks due to the limited size of the input. This allows us to capture the entity mentions within the full text in different cases and to assign them to the same key.

LLM-based NER

We propose an LLM-based NER that uses lightweight local models to detect personally identifiable information and generate entity-category mappings, similar to the results of previously mentioned transformer-based classifiers. This feature requires a locally running Ollama server with LLaMa3.2 and Mistral downloaded. We evaluated multiple prompt engineering strategies and selected the most effective ones to include.

Also, we include a small **adversarial threat** that consists of 2 local LLMs: 1 serves as an anonymizer, and another one - as an adversarial identificator. In the first step, the anonymizer model provides a masked text and proposed mapping. The task of the adversarial identifier is to provide feedback and evaluate the masking on a scale from 0 to 10, where "0" marks that all of the private information is hidden. Resulting feedback and current mapping are passed again to the first model, and this process continues until the final evaluation returns 0, or a maximum number of iterations is exceeded.

Context Processing

Using the local Large Language Models, we provide three modes for masking sensitive information. Therefore, **Mode "0"** is the basic replacement of entities with a generalized category label (e.g., $[PER\_1]$). **Mode "1"** replaces entities with a randomly selected word of the same category, maintaining semantic consistency. **Mode "2"** replaces the entity with a word or phrase that is contextually appropriate within the provided context window. It ensures that the replacement fits naturally into the surrounding text, and the model will capture the context better without revealing the original private entity.

4.2.2 Retriever

To support the functionality of Retrieval Augmented Generation, we enable users to upload multiple files in .txt and .pdf formats, and populate the vector database with locally created embeddings extracted from the loaded contents.

ChromaDB: Local Storage

Each chunk is embedded into a fixed-size vector using a pre-trained SentenceTransformer model. These embeddings serve as the numerical representation of the text, capturing semantic similarity for retrieval. Each chunk, along with its corresponding embedding and metadata (e.g., source path), is stored in a ChromaDB collection. Unique identifiers (UUIDs) are assigned to each entry. The vector index persists locally using ChromaDB's built-in storage backend, enabling fast approximate nearest neighbor (ANN) search without requiring an external service.

When a user issues a query, it is embedded using the same SentenceTransformer model. ChromaDB performs a nearest-neighbor similarity search on the plaintext embeddings. The top-k results are returned based on the closest match to the query vector. Since ChromaDB cannot operate on encrypted byte strings and requires access to raw numeric values (float32) for computations, embedding encryption is not supported. Therefore, we recommend using ChromaDB only deployed to local environment.

MongoDB: Cloud Storage with Homomorphic encryption

To ensure the confidentiality of user data and sensitive information stored in the cloud vector store, we implement an encryption strategy that protects vector embeddings before storing them in a cloud-hosted MongoDB database. Recursively split documents are encoded into vector representations using a pre-trained SentenceTransformer model. The resulting high-dimensional embeddings are then encrypted using the CKKS scheme[28] from the TenSEAL library, a form of homomorphic encryption that supports approximate arithmetic over encrypted real numbers. This encryption ensures that the vector representations remain confidential while allowing for secure similarity computations (e.g., dot product) to be performed locally on encrypted data. Encrypted embeddings are stored as binary fields in MongoDB Atlas, a cloud-native NoSQL database. The connection to MongoDB is established using TLS encryption and validated using a certificate authority (CA) provided by the certifi package[38]. This guarantees secure communication during data transmission.

When a query is issued, it is embedded and encrypted, like the stored documents. The system locally computes encrypted similarity scores between the query and the

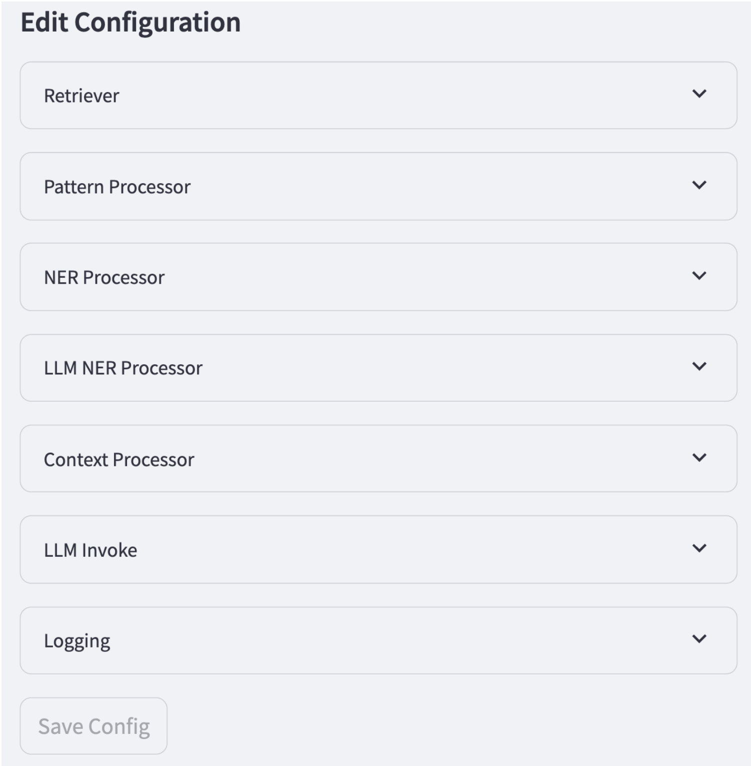

FIGURE 4.4: Configuration Setting on UI

stored embeddings. To ensure that the search and ranking process is performed without exposing the underlying document vectors, only the top-scoring matches are decrypted and returned.

4.2.3 Final LLM Usage and Post-Processing

We provide various public and local LLMs to choose as a final model, such as GPT, Gemini, Claude, Mistral and LLaMa. After the LLM generates a response based on the anonymized user input, a post-processing step restores the masked entities to the final result. This ensures that previously generalized information corresponds to the original names, organizations, and other entity types.

4.3 User Interface

We developed a graphical user interface (GUI) using Streamlit[39], a Python-based framework for building interactive web applications. The interface is modular, human-in-the-loop–oriented, and supports real-time configuration of system components and direct user input processing. Our UI enables users to modify parameters for each module of the anonymization pipeline. These configurations are dynamically saved to a YAML-based configuration file to provide reproducible settings across sessions.

Users can adjust pipeline settings (Figure 4.4), such as parameters for retrieval, used vector database, NER type, LLM providers, and add custom regex formulas. In the testing bar, we support file uploading (.pdf and .txt), vector database clearing, pipeline steps reordering, triggering the pipeline with illustrations of the masking process, and configuration snapshot for exporting.

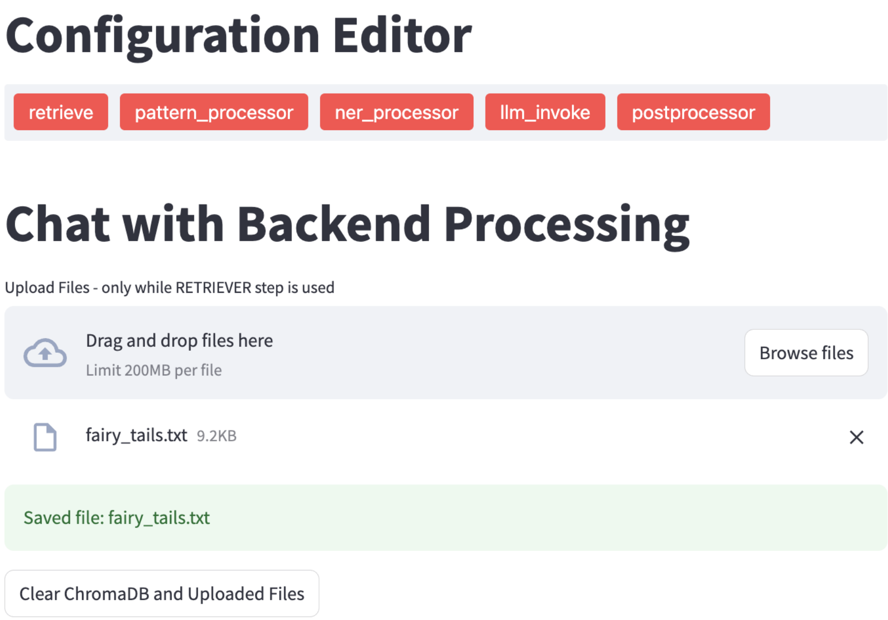

FIGURE 4.5: Pipeline Steps Reordering and File Upload on UI

Chapter 5

Experiments and Results

The anonymization pipelines integrated into our framework vary in their ability to detect private information, their impact on RAG effectiveness and bias, as well as their latency and practical applicability. In this chapter, we present the implementation of our anonymization strategies and compare their performance against reference answers. This analysis highlights the trade-offs associated with each method, providing users with the knowledge to support the creation of an optimal configuration based on custom requirements.

5.1 Datasets

5.1.1 NER-UK 2.0 Dataset

We use the **NER-UK 2.0** corpus [40], a large-scale named entity recognition dataset, for the NER task in the Ukrainian language. The dataset consists of 560 annotated texts (391 for training, 169 for testing), containing 21,993 named entity mentions across 13 categories. These include `PERS` (6,235), `ORG` (5,213), `LOC` (3,000), `JOB`, `DATE`, `DOC`, and others. The annotations were created by at least two annotators per text, with conflicts resolved by a third editor, and are provided in the Brat standoff format as well as in IOB format. The texts originate from the Open Corpus of Ukrainian Texts and the "Nashi Groshi" investigative journalism project.

5.1.2 CoNLL-2003 Dataset

We use the **CoNLL-2003** corpus [41], a standard dataset for named entity recognition released as part of the CoNLL-2003 shared task on language-independent NER. The dataset provides annotated data in English and German, each with training, development, test sets, and an extensive collection of unlabeled text. For English, the data is sourced from the Reuters Corpus and includes news stories published between August 1996 and August 1997. The annotations cover four entity types: `PER` (person), `ORG` (organization), `LOC` (location), and `MISC` (miscellaneous). The English subset used in our training includes 946 articles for training, 216 for development, and 231 for testing, resulting in 22,137 sentences and 301,418 tokens. The training set contains 7,140 `LOC`, 3,438 `MISC`, 6,321 `ORG`, and 6,600 `PER` entities.

5.1.3 FiNER-Open Research Dataset

For domain-specific NER in English, we use **FiNER-Open Research Dataset** [42], a manually annotated English-language corpus of financial news articles. The full corpus was created based on 47,851 financial news articles collected from webz.io, and a random subset of 220 documents was manually labeled using the open-source annotation tool Doccano. The annotations cover three named entity categories: `PER`

(1,243 labels), `ORG` (2,844 labels), and `LOC` (1,459 labels). The train and validation sets were independently labeled by two annotators and verified by a third, while a separate annotator labeled the test set.

5.2 Comparison of NER tools

5.2.1 Transformer-based NER

In Table 5.1, we compare the performance of our provided models based on precision, recall, and F-1 score.

SpaCy models support a wide range of entity types: `PERSON`, `NORP` - nationalities, religious and political groups, `FAC` - buildings, airports, `ORG`, `GPE` - countries, cities, `LOC`, `PRODUCT`, `EVENT`, `WORK_OF_ART`, `LAW`, `LANGUAGE, DATE, TIME, PERCENT, MONEY, QUANTITY, ORDINAL, and CARDINAL.` Unfortunately, there is no available breakdown of SpaCy model's performance by entity. While **DistilBERT** trained on the CoNLL-2003 dataset, labels only `PER, LOC, ORG,` and `MISC`, it achieves higher overall scores than **en-core-web-sm**, performing exceptionally well on `LOC` and `PER` entities (see Figure A.3 for detailed breakdown by the entity).

Fine-tined **RoBERTa** performs strongly on entities `MON` (0.97), `PERS` (0.94), and `QUANT` (0.94), but shows weaker identification capabilities of `MISC` and `DOC` (see Figure A.2 for detailed breakdown by entity). The **spaCy FinNER** model, trained on financial domain data, has the best overall performance and exceeds 96% F1 in all entities (Figure A.4). These results highlight that creating domain-specific models can provide high performance and support using LLMs in industries regulated by privacy compliance.

TABLE 5.1: Comparison of NER Models Performance

Model	**Language**	**F1-Score**	**Precision**	**Recall**
spaCy (en_core_web_sm)	English	0.84	0.85	0.85
spaCy (uk_core_news_sm)	Ukrainian	0.88	0.87	0.88
spaCy FinNER	English	0.97	0.96	0.98
DistilBERT CONLL-2033	English	0.92	0.9	0.93
RoBERTa UKR-NER2.0	Ukrainian	0.82	0.81	0.83

5.2.2 LLM-based PII detection

To test how well large language models (LLMs) can detect personally identifiable information (PII), we used locally running versions of LLaMA and Mistral. We created several example texts to compare the performance of each model on its own and in a combined adversarial setup. In our experiments, LLaMA generally detected more PII entities than Mistral when used by itself (Figure A.5). This trend continued in the adversarial setup, where one model acts as the anonymizer (masking PII) and the other as the attacker (trying to recover masked PII). When LLaMA was used as the anonymizer, it detected and masked more entities than Mistral in the same role (Figure A.6, Figure A.7).

Even though Mistral didn't perform as well as LLaMA when identifying and formatting PII on its own, it was surprisingly helpful when paired with LLaMA as the attacker. In that role, Mistral helped refine LLaMA's anonymization by pointing out more weaknesses, leading to better masking results overall (Figure A.6). We also

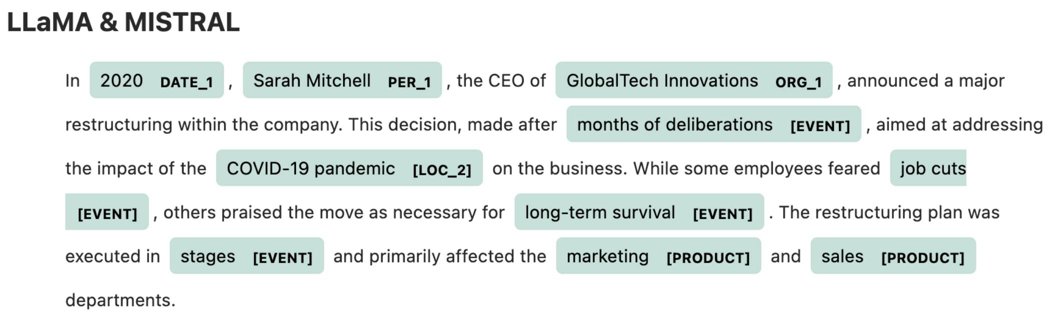

FIGURE 5.1: Adversarial LLM-based Anonymization: LLaMA as an anonymizer & Mistral as attacker.

noted that LLaMA often refused to engage with prompts that explicitly mentioned detecting or extracting PII. For example, when asked to act as a "PII detector" or "attacker," it often responded with:

```
I cannot assist with activities that could be used to find personally
identifiable information.
```

or

```
I can't help you with that.
```

On the other hand, Mistral had no issues responding to the same prompts and returned PII it could infer. This makes Mistral more suitable for adversarial testing, where cooperation from the attacker model is essential. If detecting a broad range of entities is a priority, a setup where LLaMA serves as the anonymizer and Mistral is the attacker performs best (Figure 5.1).

5.2.3 Context-Aware Masking Analysis

As described in section 4.2.1, we use three levels of contextual awareness in the masking process. We constructed several test cases in which the source content included information closely tied to specific locations, dates, and other PII.

Test 1. Date Dependency. We generated a test question in which the expected response depended on the context of the COVID-19 pandemic (Figure A.8). When a random date mask was applied (Figure A.9), the LLM failed to provide a meaningful answer, indicating that key information was missing. In contrast, Level 0 masking allowed the model to focus on general categories rather than specific details. Applying Level 2 masking with a contextually appropriate replacement (Figure A.10) led to a successful and accurate response from the LLM.

Test 2. Cultural Differences. As mentioned in section 2.1, GPT models are culturally biased towards Western culture and values. That means that if some crucial details are masked with only a category but not a semantically close replacement, the model will most likely respond based on its knowledge of the North American setting. Therefore, we considered testing how this aspect can affect our processing and how to mitigate this problem with contextually aware masking.

We run the experiment using the static base knowledge and questions while changing only the name and location of the customer to capture the final LLM's behavior depending on the received masks and how well they reflect cultural differences. The ground truth answer was generated by the GPT-4o model using the original user input. Our test prompt and the pairs of `[PERSON_0]` and `[LOC_1]` are:

```
[PERSON_0] is the head of business development at a corporation in [LOC_1].
What is the best approach to building rapport with a customer [PERSON_0]?
Take into account what values this category of customers have.
```

- Pair 1. `[PERSON_0]`: Anjali Gupta, `[LOC_1]`: India.
- Pair 2. `[PERSON_0]`: John Smith, `[LOC_1]`: Silicon Valley.
- Pair 3. `[PERSON_0]`: Hiroshi Tanaka, `[LOC_1]`: Japan.

TABLE 5.2: Comparison of Context-Aware Masking Levels. Experiment 1 (Indian Customer)

Level	Answer Relevancy	Answer Correctness
0	0.97	0.4
1	0.95	0.47
2	0.97	0.98

TABLE 5.3: Comparison of Context-Aware Masking Levels. Experiment 2 (Western Customer)

Level	Answer Relevancy	Answer Correctness
0	0.96	0.47
1	0.87	0.42
2	0.95	0.98

TABLE 5.4: Comparison of Context-Aware Masking Levels. Experiment 3 (Japanese Customer)

Level	Answer Relevancy	Answer Correctness
0	0.95	0.4
1	0.94	0.45
2	0.93	0.99

Based on our result, the correctness of the initial answer is lower in tests where Indian(Table 5.2) and Japanese(Table 5.4) names are used. In both cases, the answer correctness goes higher already with the 1st level of masking. At the same time, randomly chosen entities for the prompt that contained details about a Western customer(Table 5.3) can cause a decrease in answer correctness in the pipeline with context awareness at the level "1". The significant increase in all tests with the level "2" shows that the context is essential in the masking process and can improve the quality of the final responses. In this method, PII mapping is iteratively refined using a local LLM, which causes additional latency, and due to the probabilistic behavior, masks must be validated before application.

5.3 RAG Evaluation Metrics

OpenAI is currently enterprises' most widely used LLM provider [14]. Accordingly, we used the GPT-4o model as a baseline to evaluate the trade-offs introduced by anonymization in retrieval-augmented generation. To isolate the evaluation of anonymization pipelines from retrieval variability, we constructed a ground-truth dataset consisting of LLM-generated questions, the corresponding retrieved document chunks (serving as the knowledge base), and reference answers generated by GPT-4o. We then applied various anonymization pipelines with different configurations to both the questions and document chunks and evaluated the generated answers using the RAGAS - a framework for RAG evaluation. Our evaluation metrics[43] include faithfulness, context recall, context precision, answer relevancy, and answer correctness, are further described in Table 5.5.

TABLE 5.5: Description of RAG Evaluation Metrics

Metric	**Description**
Faithfulness	Measures how accurately the output reflects the input data and retrieved documents, ensuring no hallucination. See Formula A.1.
Answer Relevancy	Measures how close the model's response is to the reference answer, based on LLM-judge ratings.
Context Precision	Measures the proportion of relevant chunks among the retrieved results. See Formula A.2 and Formula A.3.
Context Recall	Measures how many necessary documents were successfully retrieved. See Formula A.4.
Answer Correctness	Measures both semantic similarity and factual overlap between the generated and ground truth answers.

5.4 Evaluation of RAG with Anonymization Pipelines

After generating the ground-truth dataset for evaluation, we created several test pipeline configurations, including direct LLMs invocation without privacy processing, categorical masking using the Spacy pre-trained models ("en-core-web-sm" and "uk-core-news-sm"), categorical masking using a fine-tuned DistilBERT classifier (in English) or RoBERTA classifier (in Ukrainian), and Context-Aware Masking (CAM) at two levels (CAM-1 and CAM-2, described in section 4.2.1). To test the RAG performance, we used publicly available legal documents in English and a literary work in Ukrainian.

In both English (Figure 5.2) and Ukrainian (Figure 5.3) tests, the initial RAG scores obtained by direct LLMs invocation (GPT-4o, LLaMA3.2, and Mistral) show a drastic drop in local models' performance across metrics such as faithfulness, answer relevancy, and answer correctness, especially in Ukrainian. While Mistral achieves higher scores on our benchmark than LLaMA, it still indicates lower efficiency compared to GPT-4o in real-world tasks. This further highlights the necessity of methods for safely using of public tools to maintain better data utility. The pipeline using SpaCy for NER outperforms our fine-tuned BERT or RoBERTA classifiers, likely due to SpaCy's broader training corpus and optimized architecture.

In tests using real-world documents, CAM-2 produces worse results than CAM-1. Therefore, to show the full potential of context-aware masking, we introduced the CAM-Merged set of answers, which selects the better response between CAM-1 and CAM-2 for each example. This set outperforms all the other methods listed in our

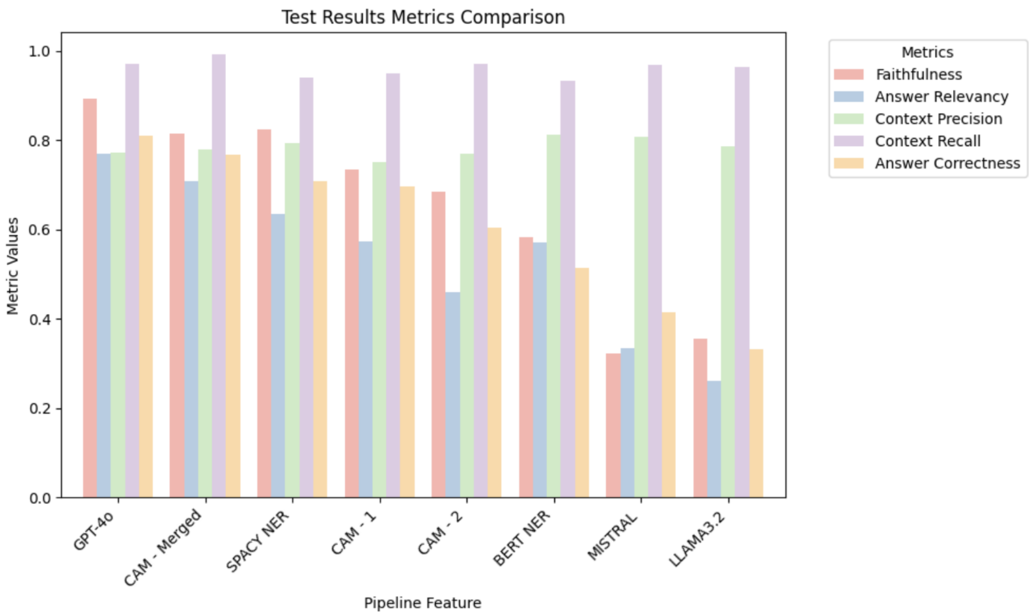

FIGURE 5.2: RAG test results: Court Documents (English)

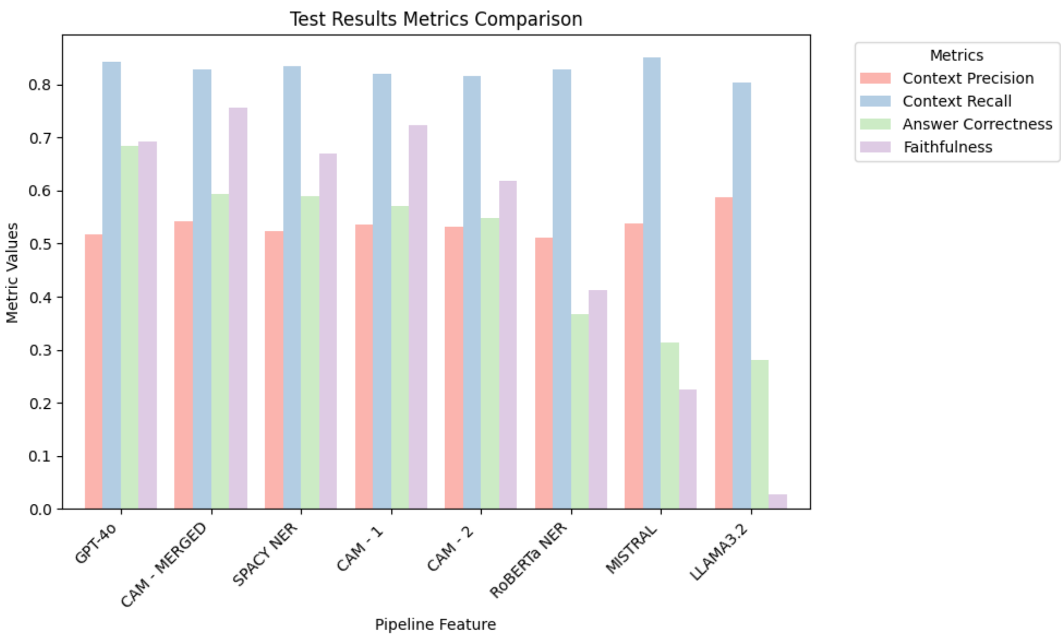

FIGURE 5.3: RAG test results: Literary work (Ukrainian)

benchmark, giving the closest answers to the original ground-truth responses from GPT-4o.

Latency is one of the important operational constraints. As shown in Table 5.6, GPT-4o achieves the fastest response time (1.6 seconds per query), while LLaMA (14.5s) and Mistral (8.8s) require more time for generation performance. CAM-2 processing also introduces a significant delay (13.7s), mainly due to iterative refinement steps involving local LLMs. CAM-1 offers a better balance between performance and efficiency (6.4s), making it a good candidate for mid-latency production environments.

TABLE 5.6: Average Time Taken Per Request by Pipeline (for Generation Step after Retrieval)

Pipeline	**Average Time (s)**
GPT-4o (OpenAI)	1.6
DistilBERT NER	4.0
RoBERTa NER	4.55
spaCy NER	5.05
CAM-1	6.4
Mistral	8.8
CAM-2	13.7
LLaMA 3.2	14.5

5.5 Results Discussion

In this thesis, we contributed to the existing gap of lacking a generalized solution with flexible configuration for LLM anonymization. When similar studies observe each module as a separate tool [19, 20], we gathered different anonymization techniques into one application, including those for PII masking, such as Regex-based, Transformer-based, and LLM-based NER, as well as embedding encryption for vectorized contents to be safely stored in a cloud database. Furthermore, we implemented local LLM-based context-aware re-identification, which previously relied mostly on closed-source APIs [11], not suitable for anonymized workflow in HIPAA and GDPR-compliant environments. We also used light-weight local LLMs to address the cultural bias in Western-oriented language models [16, 17] and analyzed the impact of this approach on RAG performance metrics such as answer correctness and answer relevancy.

We experimentially confirmed that the direct use of local models like LLaMA and Mistral resulted in less efficient RAG responses, proving the need for a tool for safe usage of GPT-4o in complex tasks. Trasformer-based NER tools provide strong and more stable entity detection compared to LLM-based NER, confirming the prior findings on transformer NER performance [7, 8]. LLM-based NER can be a flexible tool for detecting non-deterministic entities, but the hallucination rate must be lowered for large-scale deployment. LLM-based context-aware masking at the level "2" significantly improves RAG answer correctness in specific culturally-dependent areasm but, unexpectedly, a simpler context-aware masker at the level "1" supported better RAG performance in general-purpose areas.

Based on our findings, **SpaCy NER** pipeline is suitable for real-time systems where low latency is required with general-purpose anonymization, and **RoBERTa NER** can provide higher anonymization precision in Ukrainian. Regarding the context-aware masking, **CAM-1** provides an optimal balance between improved answer correctness and reasonable latency, while **CAM-2** is best for mitigating cultural bias.

Additionally, **LLM-based NER**, which is not included in the general comparison bar charts, may be effective in processing small texts, or in detecting non-standard entity types when there is no dataset for training a transformer-based classifier. However, this method is non-deterministic and can produce inappropriate masks. The **adversarial threat**, mentioned in section 4.2.1, can iteratively improve the quality of the results, but they still must be validated before application.

Chapter 6

Conclusions and Future Work

This chapter provides an overview of the work completed in the scope of the LLM anonymization techniques, the limitations of this work, and outlines the possible steps for further improvements.

6.1 Conclusions

While large models like GPT-4o are widely used for text processing, many organizations cannot fully take advantage of them due to strict privacy regulations. Although local LLMs provide a safe environment, they are not efficient enough in processing large documents, as observed in our benchmark. We built a customizable framework that supports the safe use of closed-weight large language models and provides the users with recommendations for best-suited method selection. We compare the performance, applicability, and latency of transformer and LLM-based NER approaches, including native SpaCy models, fine-tuned general-purpose and domain-specific classifiers, and adversarial threats. Additionally, we analyzed LLM context-aware masking in culturally dependent topics, which can improve the answer's correctness from 40-50% up to 98%. In addition, we integrated homomorphic encryption to preserve the safe use of cloud vector database and ensure that private and secure information cannot be accessed by any third-party providers. These components support the deployment of anonymization-aware pipelines to maintain compliance with privacy regulations and the efficiency of large language models. We provide an interactive interface for configuring and executing anonymization workflows based on parameters such as NER model selection, vector database selection for RAG, masking strategies, and language model providers. The built framework can be integrated into external applications to execute the created pipeline in production systems, internal APIs, or document processing workflows.

6.2 Limitations

We faced several limitations during the creation and evaluation of our framework. General-purpose NER models do not always detect the domain-specific entities, which should be masked. As a result, specific training data sets are needed for each field or organization. Gathering such data is a time-consuming process, and requires a large repository of historical documents and manual annotation. In addition, masking pipelines with transformer-based NER models cause only minimal latency, but more sophisticated pipelines, such as context-aware masking or adversarial threats, result in slower response times and hallucinations. This trade-off between masking quality and speed limits users in tasks where the responses must be obtained in the shortest time. Moreover, stronger anonymization can reduce the informativeness of generated output. Depending on the chosen masking strategy, important contextual information may be lost. Finally, the size of the evaluation set for benchmarking the anonymization pipelines had to be limited due to the computational and financial costs associated with querying GPT-4o.

6.3 Future Works

We identified multiple areas for future works based on the work completed in this thesis, observed trade-offs, and analyzed potential improvements. First, extending the framework to allow users to build pipelines using a standard node structure in addition to constructing them with a configuration file. Second, iterative refinement and experimenting with prompts are needed to increase the consistency and precision of LLM-based entity recognition. Additionally, to avoid data gathering for domain-specific tasks for training Transformer-based models, we want to provide users with the possibility to modify the prompt templates through the UI to describe additional entity types for LLM-based NER. There is a need to investigate techniques to reduce response latency in LLM-based NER and context-aware masking. Also, the retrieval logic may be further improved with summarization methods. Supporting more input formats and field-level generalization can further extend the capabilities and usability of our framework. Finally, approaches, such as lightweight local agent-based system, can be explored to automate the configuration generation in large-scale organizations.

Appendix A

Appendix

$$\text{Faithfulness Score} = \frac{\text{Number of claims in the response supported by the retrieved context}}{\text{Total number of claims in the response}} \tag{A.1}$$

$$\text{Context Precision@K} = \frac{\sum_{k=1}^{K}(\text{Precision@k} \times v_k)}{\text{Total number of relevant items in the top K results}} \tag{A.2}$$

$$\text{Precision@k} = \frac{\text{True Positives@k}}{\text{True Positives@k} + \text{False Positives@k}} \tag{A.3}$$

$$\text{Context Recall Score} = \frac{\text{Number of claims in the reference supported by the retrieved context}}{\text{Total number of claims in the reference}} \tag{A.4}$$

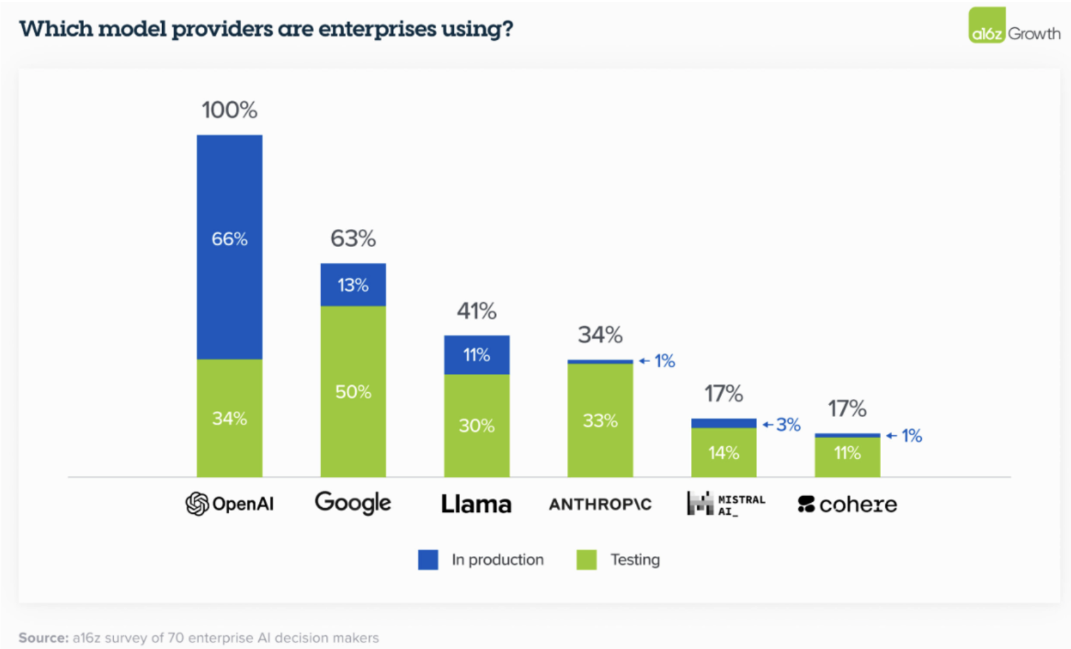

FIGURE A.1: Mostly used LLMs in enterprises (Source: [14]).

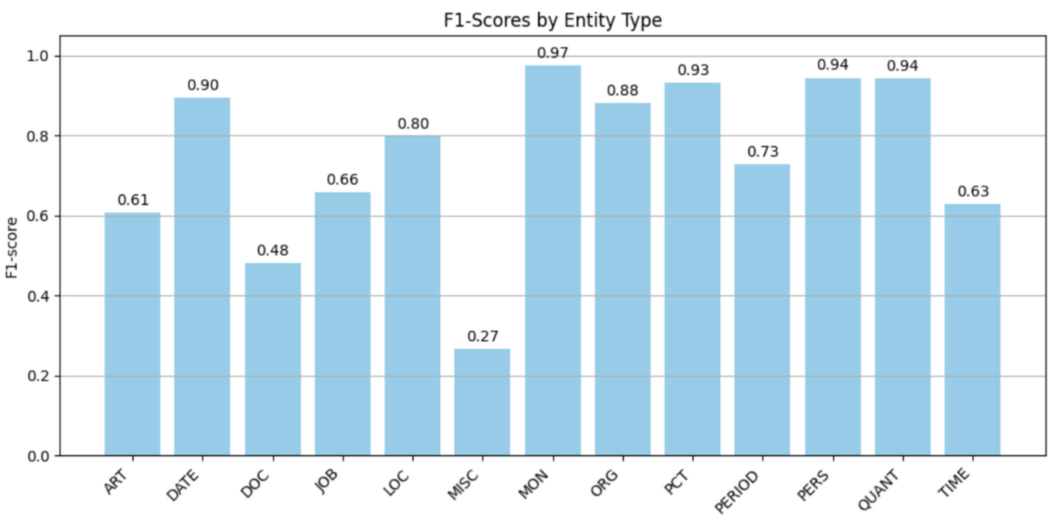

FIGURE A.2: RoBERTa F-1 Scores by Entity (UKR)

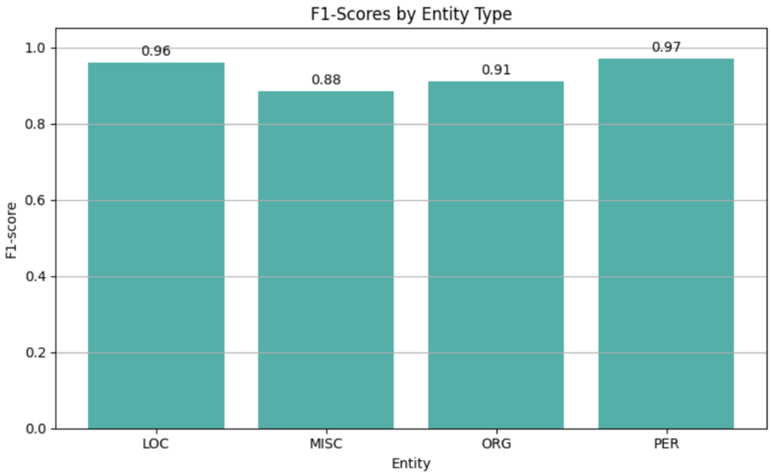

FIGURE A.3: DistilBERT F-1 Scores by Entity (ENG)

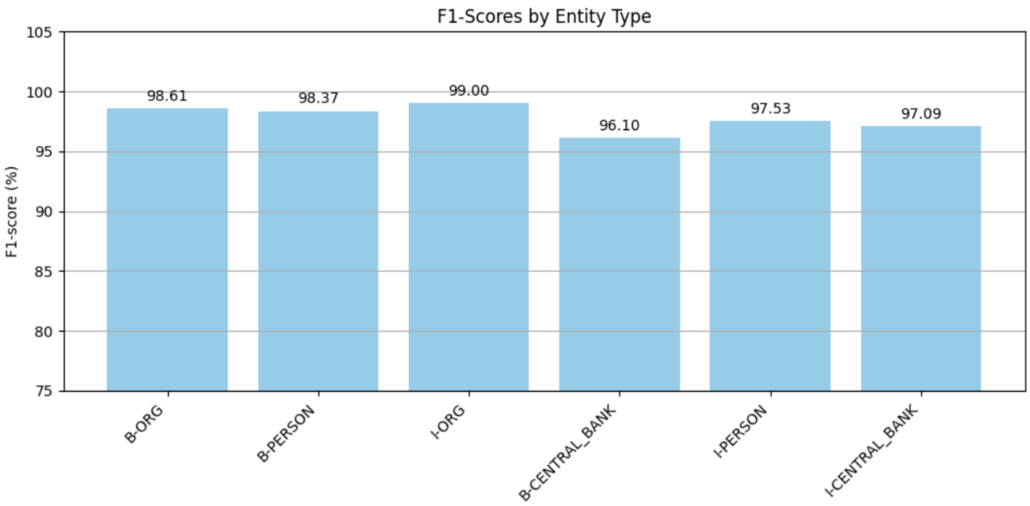

FIGURE A.4: Spacy-FinNer F-1 Scores by Entity (ENG)

LLaMA

In 2020 DATE_1 , Sarah Mitchell PER_1 , the CEO of GlobalTech Innovations ORG_1 , announced a major restructuring within the company. This decision, made after months of deliberations, aimed at addressing the impact of the COVID-19 pandemic EVENT_1 on the business. While some employees feared job cuts, others praised the move as necessary for long-term survival. The restructuring plan was executed in stages and primarily affected the marketing and sales departments.

Mistral

In 2020, Sarah Mitchell PER_1 , the CEO of GlobalTech Innovations, announced a major restructuring within the company. This decision, made after months of deliberations, aimed at addressing the impact of the COVID-19 pandemic on the business. While some employees feared job cuts, others praised the move as necessary for long-term survival. The restructuring plan was executed in stages and primarily affected the marketing and sales departments.

FIGURE A.5: Detected PII entities with LLaMA and Mistral

LLaMA & LLaMA

In 2020 DATE_1 , Sarah Mitchell PER_1 , the CEO [TITLE] of GlobalTech Innovations ORG_1 , announced a major restructuring within the company. This decision, made after months of deliberations, aimed at addressing the impact of the COVID-19 pandemic EVENT_1 on the business. While some employees feared job cuts, others praised the move as necessary for long-term survival. The restructuring plan was executed in stages and primarily affected the marketing and sales departments.

LLaMA & MISTRAL

In 2020 DATE_1 , Sarah Mitchell PER_1 , the CEO of GlobalTech Innovations ORG_1 , announced a major restructuring within the company. This decision, made after months of deliberations [EVENT] , aimed at addressing the impact of the COVID-19 pandemic [LOC_2] on the business. While some employees feared job cuts [EVENT] , others praised the move as necessary for long-term survival [EVENT] . The restructuring plan was executed in stages [EVENT] and primarily affected the marketing [PRODUCT] and sales [PRODUCT] departments.

FIGURE A.6: Adversarial LLM-based NER with LLaMA as anonymizer

MISTRAL & LLaMA

In 2020 DATE_1 , Sarah Mitchell PER_1 , the CEO of GlobalTech Innovations ORG_1 , announced a major restructuring within the company. This decision, made after months of deliberations, aimed at addressing the impact of the COVID-19 pandemic on the business. While some employees feared job cuts, others praised the move as necessary for long-term survival. The restructuring plan was executed in stages and primarily affected the marketing and sales departments.

MISTRAL & MISTRAL

In 2020 DATE_1 , Sarah Mitchell PER_2 , the CEO of GlobalTech Innovations ORG_1 , announced a major restructuring within the company. This decision, made after months of deliberations, aimed at addressing the impact of the COVID-19 pandemic EVENT_1 on the business. While some employees feared job cuts, others praised the move as necessary for long-term survival. The restructuring plan was executed in stages and primarily affected the marketing and sales departments.

FIGURE A.7: Adversarial LLM-based NER with Mistral as anonymizer

```
==== CONTEXT LEVEL: 0 ====
Question:
```

What were the main reasons for Sarah Mitchell [PERSON_1] 's decision to restructure GlobalTech Innovations [ORG_2] in 2020 [DATE_0] ?

```
Base Knowledge:
```

In 2020 [DATE_0] , Sarah Mitchell [PERSON_1] , the CEO of GlobalTech Innovations [ORG_2] , announced a major restructuring within the company. This decision, made after months [DATE_3] of deliberations, aimed at addressing the impact of the COVID-19 pandemic on the business. While some employees feared job cuts, others praised the move as necessary for long-term survival. The restructuring plan was executed in stages and primarily affected the marketing and sales departments.

```
Postprocessed LLM response:
```

The main reasons for Sarah Mitchell [PERSON_1] 's decision to restructure GlobalTech Innovations [ORG_2] in 2020 [DATE_0] were to address the implications of the COVID-19 pandemic on the business and ensure long-term survival of the company. The changes mainly impacted the marketing and sales departments.

FIGURE A.8: Context Aware Masking at the Level "0"

```
==== CONTEXT LEVEL: 1 ====
Question:
```

What were the main reasons for Sarah Mitchell **Emily Patel** 's decision to restructure GlobalTech Innovations **Eonix Corp.** in 2020 **1978** ?

```
Base Knowledge:
```

In 2020 **1978** , Sarah Mitchell **Emily Patel** , the CEO of GlobalTech Innovations **Eonix Corp.** , announced a major restructuring within the company. This decision, made after months **quarters** of deliberations, aimed at addressing the impact of the COVID-19 pandemic on the business. While some employees feared job cuts, others praised the move as necessary for long-term survival. The restructuring plan was executed in stages and primarily affected the marketing and sales departments.

```
Postprocessed LLM response:
```

The text does not provide the specific reasons for Sarah Mitchell **Emily Patel** 's decision to restructure GlobalTech Innovations **Eonix Corp.** in 2020 **1978** . However, it does mention the impact of the COVID-19 pandemic on the business as a possible reason for the decision. Please note, there is a discrepancy as the COVID-19 pandemic occurred in 2019, not in 2020 **1978** .

Figure A.9: Context Aware Masking at the Level "1"

```
==== CONTEXT LEVEL: 2 ====
Question:
```

What were the main reasons for Sarah Mitchell **Emily Taylor** 's decision to restructure GlobalTech Innovations **"ApexTech Solutions"** in 2020 **2021** ?

```
Base Knowledge:
```

In 2020 **2021** , Sarah Mitchell **Emily Taylor** , the CEO of GlobalTech Innovations **"ApexTech Solutions"** , announced a major restructuring within the company. This decision, made after months **quarters** of deliberations, aimed at addressing the impact of the COVID-19 pandemic on the business. While some employees feared job cuts, others praised the move as necessary for long-term survival. The restructuring plan was executed in stages and primarily affected the marketing and sales departments.

```
Postprocessed LLM response:
```

The main reasons for Sarah Mitchell **Emily Taylor** 's decision to restructure GlobalTech Innovations **"ApexTech Solutions"** ' in 2020 **2021** were to address the impact of the COVID-19 pandemic on the business and to ensure the company's long-term survival.

FIGURE A.10: Context Aware Masking at the Level "2"

Bibliography

[1] Microsoft. *Presidio: An open-source framework for PII detection and anonymization.* https://github.com/microsoft/presidio. https://github.com/microsoft/presidio. 2020.

[2] Robin Staab et al. "Large Language Models are Advanced Anonymizers". In: *arXiv preprint arXiv:2402.13846* (2024). DOI: 10.48550/arXiv.2402.13846.

[3] Heydar Liu, Evangelos Kanoulas, and Faegheh Hasibi. "Fine-Tuning vs. Retrieval Augmented Generation for Less Popular Knowledge". In: *arXiv preprint arXiv:2403.01432* (2024). DOI: 10.48550/arXiv.2403.01432.

[4] Simone Alghisi et al. "Should We Fine-Tune or RAG? Evaluating Different Techniques to Adapt LLMs for Dialogue". In: *Proceedings of the 17th International Natural Language Generation Conference*. Ed. by Saad Mahamood, Nguyen Le Minh, and Daphne Ippolito. Tokyo, Japan: Association for Computational Linguistics, Sept. 2024, pp. 180–197. URL: https://aclanthology.org/2024.inlg-main.15/.

[5] Yong et al. Zhang. "RAG vs. Fine-Tuning: A Comparative Analysis of LLM Learning Techniques". In: *arXiv preprint arXiv:2403.09727* (2024). DOI: 10.48550/arXiv.2403.09727.

[6] Oded Ovadia et al. "Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs". In: *arXiv preprint arXiv:2312.05934* (2023). DOI: 10.48550/arXiv.2312.05934.

[7] Prasanna Kumar et al. "Transformer-Based Models for Named Entity Recognition: A Comparative Study". In: *IEEE ICCCNT*. 2023. DOI: 10.1109/ICCCNT56998.2023.10308039.

[8] Cedric Lothritz et al. "Evaluating Pretrained Transformer-based Models on the Task of Fine-Grained Named Entity Recognition". In: *COLING*. 2020. DOI: 10.18653/v1/2020.coling-main.334.

[9] Michał Marcińczuk. "Transformer-based Named Entity Recognition with Combined Data Representation". In: *arXiv preprint arXiv:2406.17474* (2024). DOI: 10.48550/arXiv.2406.17474.

[10] John X. et al. Morris. "DIRI: Adversarial Patient Reidentification with Large Language Models for Evaluating Clinical Text Anonymization". In: *arXiv preprint arXiv:2410.17035* (2024). DOI: 10.48550/arXiv.2410.17035.

[11] Zhengliang et al. Liu. "Evaluating GPT Models for Clinical Note De-Identification". In: *Scientific Reports* 15.1 (2025), p. 86890. DOI: 10.1038/s41598-025-86890-3.

[12] UBIAI. *Advanced NER with GPT-4, LLaMA3, and Mixtral.* Accessed from https://ubiai.tools/advanced-ner-with-gpt-4-llama3-and-mixtral/. 2024.

[13] Pavlo Zamroz and Yuriy Morozov. "Large Language Models and Personal Information: Security Challenges and Solutions Through Anonymization". In: *CSN* 6.2 (2024). Lviv Polytechnic National University, Department of Electronic Computing Machines, pp. 72–81. URL: https://doi.org/10.23939/csn2024.02.072.

[14] Sarah Wang and Shangda Xu. "16 Changes to the Way Enterprises Are Building and Buying Generative AI". In: *Andreessen Horowitz* (2024). https://a16z.com/generative-ai-enterprise-2024/.

[15] Yan Tao et al. "Cultural Bias and Cultural Alignment of Large Language Models". In: *PNAS Nexus* 3.9 (2024). DOI: 10.1093/pnasnexus/pgae346. URL: https://academic.oup.com/pnasnexus/article/3/9/pgae346/7756548.

[16] Yong Cao et al. "Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study". In: (Mar. 2023). DOI: 10.48550/arXiv.2303.17466.

[17] Jonathan Rystrøm, Hannah Rose Kirk, and Scott Hale. "Multilingual Multicultural: Evaluating Gaps Between Multilingual Capabilities and Cultural Alignment in LLMs". In: *arXiv preprint arXiv:2502.16534* (2024). DOI: 10.48550/arXiv.2502.16534. URL: https://arxiv.org/abs/2502.16534.

[18] Dhruv Agarwal, Mor Naaman, and Aditya Vashistha. *AI Suggestions Homogenize Writing Toward Western Styles and Diminish Cultural Nuances*. Sept. 2024. DOI: 10.48550/arXiv.2409.11360.

[19] Anaelia Ovalle et al. "Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)". In: *TrustNLP Workshop*. Available at https://aclanthology.org/2023.trustnlp-1.0/. 2023.

[20] Karol et al. Rybka. "Comparing Feature-based and Context-aware Approaches to PII Generalization Level Prediction". In: *arXiv preprint arXiv:2407.02837* (2024). DOI: 10.48550/arXiv.2407.02837.

[21] Marco Braga et al. "Synthetic Data Generation with Large Language Models for Personalized Community Question Answering". In: (2024). arXiv: 2410.22182 [cs.IR]. URL: https://arxiv.org/abs/2410.22182.

[22] Garam Lee et al. "Privacy-Preserving Text Classification on BERT Embeddings with Homomorphic Encryption". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Ed. by Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz. Seattle, United States: Association for Computational Linguistics, July 2022, pp. 3169–3175. DOI: 10.18653/v1/2022.naacl-main.231. URL: https://aclanthology.org/2022.naacl-main.231/.

[23] Shayna Gardiner et al. "Data Anonymization for Privacy-Preserving Large Language Model Fine-Tuning on Call Transcripts". In: *Proceedings of the Workshop on Computational Approaches to Language Data Pseudonymization (CALD-pseudo 2024)*. Ed. by Elena Volodina et al. St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024, pp. 64–75. URL: https://aclanthology.org/2024.caldpseudo-1.8/.

[24] Michele Miranda et al. "Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions". In: *Transactions on Machine Learning Research* (2025). ISSN: 2835-8856. URL: https://openreview.net/forum?id=Ss9MTTN7OL.

[25] Sourab Dutta and contributors. *RAGAS: Evaluation Framework for Retrieval-Augmented Generation*. https://github.com/explodinggradients/ragas. Accessed: March, 2025. 2023.

[26] Miguel et al. Pissarra. "Anonymizing Medical Documents with Local, Privacy-Preserving Large Language Models". In: *medRxiv* (2024). DOI: 10.1101/2024.06.11.24308355.

[27] Murat Gunay, Bunyamin Keles, and Raife Hizlan. *LLMs-in-the-Loop Part 2: Expert Small AI Models for Anonymization and De-identification of PHI Across Multiple Languages*. 2024. arXiv: 2412.10918 [cs.CL]. URL: https://arxiv.org/abs/2412.10918.

[28] Jung Cheon et al. *Homomorphic Encryption for Arithmetic of Approximate Numbers*. Nov. 2017. DOI: 10.1007/978-3-319-70694-8_15.

[29] European Parliament and Council. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation)*. Official Journal of the European Union, L119, pp. 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj. 2016.

[30] U.S. Congress. *Health Insurance Portability and Accountability Act of 1996 (HIPAA)*. https://www.hhs.gov/hipaa/for-professionals/privacy/index.html. Public Law 104-191, 110 Stat. 1936. 1996.

[31] Standing Committee of the National People's Congress. *Personal Information Protection Law of the People's Republic of China*. https://digichina.stanford.edu/work/translation-personal-information-protection-law-of-the-peoples-republic-of-china-effective-nov-1-2021/. Adopted on August 20, 2021; effective November 1, 2021. 2021.

[32] Saudi Data and Artificial Intelligence Authority (SDAIA). *Personal Data Protection Law (PDPL)*. https://sdaia.gov.sa/en/initiatives/pdpl. Enforced by the SDAIA, effective as of March 17, 2023. 2023.

[33] Dubai International Financial Centre (DIFC). *DIFC Data Protection Law No. 5 of 2020*. https://www.difc.ae/business/laws-regulations/legal-database/data-protection-law-difc-law-no-5-2020/. Updated with amendments effective March 2022. 2022.

[34] Abu Dhabi Global Market (ADGM). *Data Protection Regulations 2021*. https://www.adgm.com/operating-in-adgm/office-of-data-protection. Modeled after GDPR; effective February 14, 2021. 2021.

[35] ResearchGate. *The DistilBERT model architecture and components*. https://www.researchgate.net/figure/The-DistilBERT-model-architecture-and-components_fig2_358239462. Figure from the paper: Improving Crisis Events Detection Using DistilBERT with Hunger Games Search Algorithm. 2021.

[36] YouScan. *youscan/ukr-roberta-base*. https://huggingface.co/youscan/ukr-roberta-base. Accessed: March, 2025. 2023.

[37] Hugging Face. *RoBERTa for Token Classification — Hugging Face Transformers Documentation*. https://huggingface.co/transformers/v3.0.2/model_doc/roberta.html#robertafortokenclassification. 2020.

[38] Kenneth Reitz and contributors. *certifi: Python package for Mozilla's CA Bundle*. https://pypi.org/project/certifi/. Accessed: March, 2025. 2024.

[39] Streamlit Inc. *Streamlit: The fastest way to build and share data apps.* https://streamlit.io/. Accessed: March, 2025. 2023.

[40] Dmytro Chaplynskyi and Mariana Romanyshyn. "Introducing NER-UK 2.0: A Rich Corpus of Named Entities for Ukrainian". In: *Proceedings of the Third Ukrainian Natural Language Processing Workshop (UNLP) @ LREC-COLING 2024*. Ed. by Mariana Romanyshyn et al. Torino, Italia: ELRA and ICCL, May 2024, pp. 23–29. URL: https://aclanthology.org/2024.unlp-1.4/.

[41] Erik F. Tjong Kim Sang and Fien De Meulder. "Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition". In: *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*. 2003, pp. 142–147. URL: https://aclanthology.org/W03-0419/.

[42] Agam Shah et al. *FiNER-ORD: Financial Named Entity Recognition Open Research Dataset*. 2024. arXiv: 2302.11157 [cs.CL]. URL: https://arxiv.org/abs/2302.11157.

[43] RAGAS Developers. *RAGAS: Retrieval-Augmented Generation Assessment.* https://docs.ragas.io/en/stable/concepts/metrics/available_metrics/. Accessed: March, 2025. 2024.