**Research Proposal Template**

**Table of Contents**

[Abstract 2](#_Toc204372912)

[1. Background 2](#_Toc204372913)

[2. Literature Review 3](#_Toc204372914)

[3. Research Questions (If any) 3](#_Toc204372915)

[4. Aim and Objectives 4](#_Toc204372916)

[5. Significance of the Study 4](#_Toc204372917)

[6. Scope of the Study 4](#_Toc204372918)

[7. Research Methodology 4](#_Toc204372919)

[8. Requirements Resources 5](#_Toc204372920)

[9. Research Plan 5](#_Toc204372921)

[References 6](#_Toc204372922)

# Abstract

**This paper aims to explore at preserving privacy of client data** while using Large Language Models (LLMs). The study will explore current “Split-N-Denoise (SnD)” (Mai et al., 2024), and other privacy-preserving methods, such as cryptographic techniques (e.g., homomorphic encryption)(Chen et al., 2022),and other perturbation approaches to find out how they balance utility and privacy, particularly in local differential privacy (LDP) for text data.

This paper proposes **CleanSplit Model (CSM)**, a novel privacy-preserving inference framework by deploying only the lightweight token representation layer on the client side. Its main objective is to safeguard user privacy in LLM inference without modifying model parameters, optimizing the crucial trade-off between privacy and utility.

**Expected outcome is to find out** if new approach provides a **significantly reduced computational overhead for removing PII data at client side where secret data is involved** without compromising on LLM result efficiency.

# 1. Background

The **growing concern over data privacy** in LLMs is reflected in recent governmental and regulatory interventions, such as the EU's AI Act, the US Executive Order on Safe, Secure, and Trustworthy AI, and regulations like GDPR and HIPAA. These frameworks emphasize the necessity for AI systems to incorporate **robust privacy-preserving mechanisms**. Therefore, developing a practical and efficient solution for local privacy protection in LLM inference is paramount. This research aims to fill this critical gap in homomorphic encryption (HE) and secure multi-party computation (SMPC), incurring **substantial computational overhead** and Conversely, **perturbation approaches that** often **struggles to balance utility and privacy**, leading to reduced model accuracy process by proposing a novel framework for privacy-preserving LLM inference, Also Split-N-Denoise (SnD) framework tackles some of these limitations but it uses Differential Privacy algorithm which is good in adding mathematical noise can still reveal PII’s data to LLM Server. E.g. in Financial/ healthcare applications it becomes ultra important to remove PII data where SnD architecture might fail. it **doesn’t use of Named Entity Recognition (NER) due to which sometimes PII data gets missed out, we would a specific research on educational data available**. By addressing current limitations the study seeks to enhance security measures in educational, Heathcare, finance applications in preserving PII data.

# 2. Literature Review

A **notable dearth of research addresses local privacy during the inference phase with a fully frozen LLM**. This scenario is crucial for proprietary **black-box API access contexts**, such as GPT-4, where model parameters cannot be altered. Intuitive approaches like anonymizing sensitive terms before LLM input are insufficient, as they fail to conceal other linguistic elements vital for semantic interpretation, compromising full privacy and hindering tasks requiring exact semantic interpretation. Furthermore, existing denoising techniques, if deployed on the server side, are limited because the server lacks knowledge of the injected noise levels, creating a conflict with privacy protection

(Mai et al., 2024) ***proposed a* novel privacy-preserving inference framework that** introduces a **client-side denoising model** that leverages knowledge of raw inputs and noise levels to enhance embedding utility after noise injection but is unable to fully protect the PII data as it uses DP technique to protect privacy but DP can be used to **protect statistics or summaries** derived from user data but individual data still doesn’t get protected. This study aims to improve the privacy protection by adding NER techniques to remove PII data and replace it with same Fake data which even helps in protecting the performance of LLM as LLM doesn’t looses the context of conversation. Hence my proposed architecture should be able to effectively protect the PII data in chatbots for healthcare/financials domains where customer privacy is utmost important.

# 3. Research Questions (If any)

* How to effectively extract NER objects (Client side LLM vs simple python framework) without impacting performance at client side chatbots?
* Find out any Impact on LLM accuracy with NER replacements.
* How to preserve context of conversation with LLM and still protect the privacy?

# 4. Aim and Objectives

**Aim**:

To develop an Architecture pattern to efficiently protect PII data.

**Objectives**:

* Study existing patterns to protect PII data and come up with limitations of each.
* To find a pattern to find and replace the PII data from client side with redaction data to protect client data
* Propose an architecture that can protect PII data without compromising on LLM efficiency and still light-weight at client end.

# 5. Significance of the Study

Propose Architecture that could enhance the protection of PII data from client side chatbots whereby providing a more reliable and efficient way of preserving PII data. This would enhance the security and trust of LLM chatbots particularly amongst financial/ healthcare and education industry where protecting your PII data is utmost important and is mandated by regulators. This contributes in potentially reducing frauds but not leaking PII data over internet.

# 6. Scope of the Study

Propose Architecture only covers PII data protection and efficiency but doesn’t cover other type of secret data like

* PHI (Protected Health Information) - Medical conditions, prescriptions, diagnoses,
* PIB (Personally Identifiable Behaviour)- Browsing history, app usage, search queries
* Secrets or Credentials- API keys, passwords, access tokens, SSH keys
* Financial Information - Bank balances, investment data, transaction details
* Intellectual Property (IP) - Source code, internal documentation, trade secrets
* National Security or Classified Data - Defence-related terms, classified project names

# 7. Research Methodology

This study employs a **mixed-methods research approach** that combines both quantitative and qualitative methodologies to comprehensively investigate.

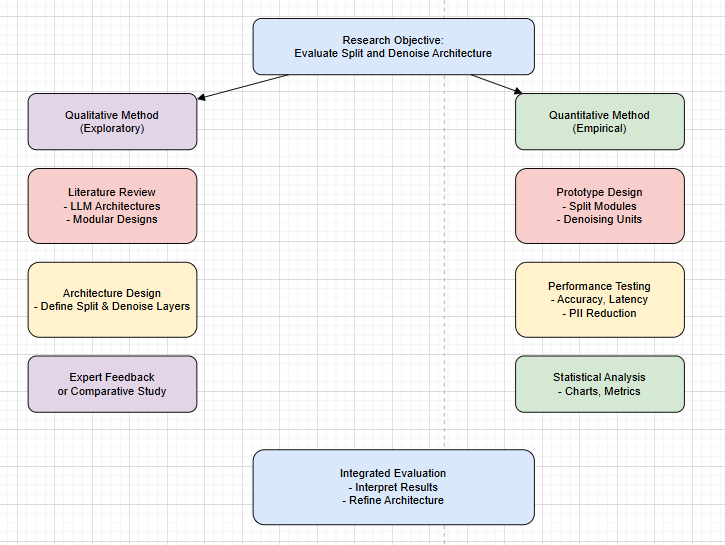
**Qualitative Component:**

Enables analysis of architectural design decisions and their impact on chatbot behaviour.

**Quantitative Component:**

Enables precise measurement of performance metrics such as response accuracy, latency, computational efficiency, and PII removal effectiveness

Benchmark existing PII protecting methods using leakage metrics.



# 8. Requirements Resources

#### ****Software/Tools****:

| **Category** | **Tools** |
| --- | --- |
| Programming | **Python, Jupyter, or Colab** |
| ML/LLM Frameworks | **OLLAMA, LangChain, OpenAI API, spaCy** |
| PII Detection | **Presidio (Microsoft) or Regex-based custom matchers** |
| Metrics & Privacy Testing | **PrivacyMeter** |
| Visualization | **Matplotlib, Seaborn, or Plotly** |
| Backend Services | **Docker, FastAPI** |

**Hardware**: Good spec machine with at least 16 GB of RAM and access to LLM models like Chat GPT, Sonet

**Data Sources**: [PII | External Dataset](https://www.kaggle.com/datasets/alejopaullier/pii-external-dataset) , The dataset includes extensive samples containing personally identifiable information (PII) such as names, contact details, and addresses.

# 9. Research Plan

* Phase 1: Literature review and hypothesis formulation (Month 1).
  + - Finalize research questions, scope, and hypotheses.
    - Define evaluation metrics: privacy protection, response quality, latency
  + **Risk**: Limited access to recent studies.
  + **Contingency**: Use interlibrary loans and online databases to access necessary resources.
* Phase 2: Design & Data collection and preprocessing (Month 2)
  + - Decide on NER model (e.g., fine-tune spaCy or transformer-based NER)
    - Define categories of PII to target (e.g., names, addresses, emails)
    - Plan architecture integration of NER & fake data injection
  + Risk:
    - Delays in data acquisition.
    - Understanding some of the python libraries.
  + Contingency:
    - Have backup data sources and extend the data collection period if needed.
    - Take some courses on python libraries.
* Phase 3: Data analysis and interpretation (Month 3)
  + - Train or fine-tune the NER model
    - Replace PII in user inputs and generate altered prompts
    - Log original vs modified queries for analysis
  + Risk: Software or hardware malfunctions.
  + Contingency: Regularly back up data and have alternative analysis tools available.
* Phase 4: Writing and dissemination of findings (Month 4-6)
  + Risk: Delays in writing or peer review.
  + Contingency: Allocate extra time for revisions and seek early feedback from peers.

# References

* (Mai et al., 2024)
* (Chen et al., 2023)
* (Zhang and Zhong, 2013)