### Evaluating Local LLM-Based De-identification in a Hybrid Clinical RAG Framework

Arun Joshi, Aayush Adhikari  
Department of Computer Science  
University of Texas at Austin

**Abstract**

Large Language Models (LLMs) offer unprecedented capabilities for clinical decision support, yet their application is often hindered by stringent privacy regulations like the Health Insurance Portability and Accountability Act (HIPAA). We present a novel hybrid architecture that attempts to bridge this gap by employing local de-identification models to sanitize patient data before sending it to remote, high-performance LLMs for clinical analysis. Our approach compares two local de-identification strategies: a fine-tuned Google Gemma 2B model trained with LoRA on synthetic clinical data, and a baseline quantized Gemma 3B model. Through qualitative evaluation of this privacy-preserving RAG system, we demonstrate the technical feasibility of the hybrid approach while uncovering critical limitations. LLM-based de-identification, despite being effective, failed to achieve 100% removal of sensitive data. This finding highlights the need for more robust, multi-layered privacy solutions.

### 1. Introduction

The advent of powerful Large Language Models (LLMs) has created significant opportunities for advancing healthcare through intelligent clinical assistants. These models can synthesize vast amounts of patient data to answer complex clinical questions, assist with differential diagnoses, and recommend evidence-based treatments. Proprietary models like OpenAI's GPT series have demonstrated exceptional clinical reasoning, but their use is severely restricted by privacy laws like HIPAA.

Conversely, while local open-source models mitigate privacy risks, they often lack the sophisticated reasoning and tool-calling capabilities of their proprietary counterparts. This presents a critical challenge: **how can healthcare systems harness the power of state-of-the-art, remote LLMs for clinical analysis while maintaining strict, verifiable adherence to patient privacy regulations?**

To address this, we propose and evaluate a hybrid, privacy-preserving RAG system. The architecture's core principle is a division of labor: a small, local LLM is tasked with removing PHI from clinical records. This anonymized context is then sent to a more powerful remote LLM (OpenAI's GPT-4.1) for higher-order reasoning. We compare two approaches to local de-identification: a specialized fine-tuned model versus a general-purpose baseline.

### 2. Related Work

Our research is situated at the confluence of clinical data de-identification, privacy-preserving machine learning, and applied LLM architectures.

#### 2.1. Clinical Data De-identification

#### Clinical text de-identification has evolved significantly over the past decade. Early approaches relied heavily on rule-based systems like Philter, which employed regular expressions and carefully curated dictionaries to identify and redact PHI [2]. While these methods proved effective for structured data with predictable patterns, they struggled with the nuanced, varied language of clinical narratives.

#### The field has since shifted toward machine learning approaches, particularly Named Entity Recognition (NER) models. Tools like medspaCy and domain-specific models such as ClinicalBERT have become the gold standard, demonstrating superior performance in identifying PHI within complex clinical contexts [3,4]. However, these specialized models are typically designed for batch processing scenarios and lack the interactive capabilities required for real-time clinical assistants.

#### Our work explores whether general-purpose LLMs, with their broad language understanding and generative capabilities, can match or exceed the performance of specialized NER models while providing the flexibility needed for interactive clinical workflows.

#### 2.2. Privacy-Preserving Machine Learning (PPML)

#### The broader machine learning community has developed sophisticated techniques for privacy preservation, including Federated Learning and Differential Privacy [5]. These approaches have shown considerable success in training scenarios where the goal is to learn from distributed, sensitive data without centralizing it.

#### However, our use case presents a different challenge. We're working with pre-trained, proprietary models where the privacy concern centers on protecting input data during inference rather than during training. This shifts the focus from learning privacy-preserving representations to ensuring complete sanitization of individual data instances before they leave the secure environment.

#### 2.3. LLMs for Orchestration and Reasoning

### Recent advances in LLM capabilities, particularly the development of tool-calling and ReAct (Reasoning and Acting) frameworks, have opened new possibilities for clinical AI systems [6]. These capabilities allow models like GPT-4.1 to serve as intelligent orchestrators, dynamically deciding which tools to invoke and how to synthesize information from multiple sources.

### Our hybrid architecture leverages these orchestration capabilities while extending the delegation pattern to include privacy protection. By treating de-identification as another tool in the LLM's toolkit, we can maintain the flexibility and reasoning power that make these models attractive for clinical applications.

### 3. Methodology

Our system implements a modular pipeline that processes clinical queries while maintaining strict separation between sensitive data handling and external reasoning capabilities.

#### 3.1. System Architecture

The architecture comprises two primary LLM components, a suite of data retrieval tools, and a central orchestrator, with a strict boundary between local and remote environments.

**Remote Orchestrator (GPT-4.1):** Serves as the system's cognitive center, interpreting clinical queries, orchestrating tool calls, and synthesizing responses using only de-identified data. This model never sees raw PHI but leverages its superior clinical reasoning capabilities to provide meaningful insights.

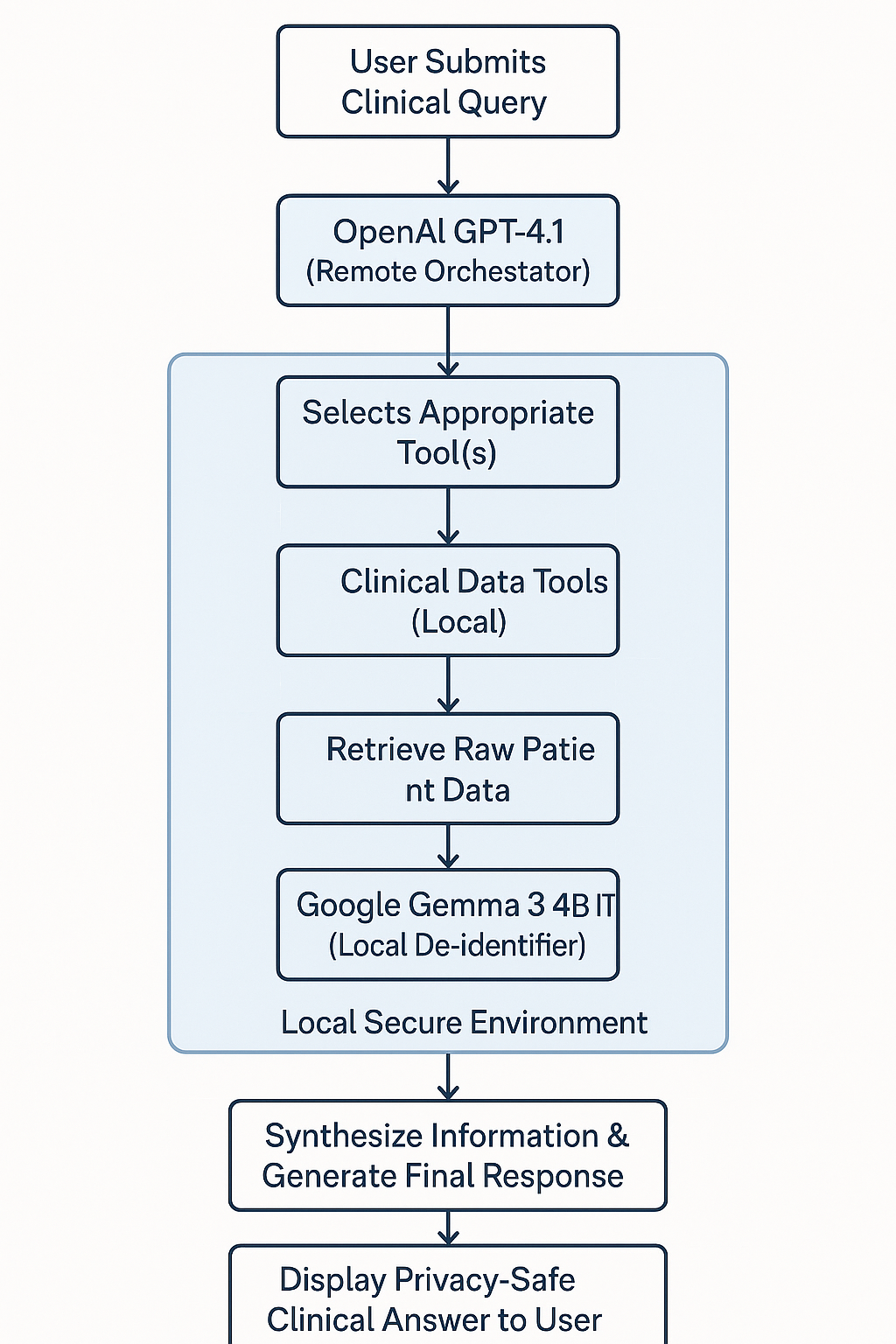
**Local De-identification Models:** Two distinct approaches operating entirely within the secure environment:

**Fine-tuned Gemma 2B:** A specialized model enhanced with LoRA adapters trained specifically on clinical de-identification tasks

**Baseline Gemma 3B:** A quantized general-purpose model serving as our comparison benchmark

**Clinical Data Tools:** Eight Python functions that simulate EHR access, each retrieving specific data types (observations, conditions, medications, procedures, imaging studies, immunizations, allergies, and care plans) from our synthetic dataset.

**Clinical Assistant Controller:** A Python orchestration class that manages the entire workflow, ensuring proper sequencing of data retrieval, de-identification, and remote reasoning.



**Figure 1:** System workflow, showing the local de-identifier processing data within the secure environment.

#### 3.2. Fine-Tuning for Specialized De-identification

To investigate whether specialized training could improve de-identification performance, we developed a fine-tuned variant of Google's Gemma 2B model using parameter-efficient techniques.

**3.2.1 Training Data Development**

We generated a comprehensive dataset of 343 clinical de-identification examples using a two-stage automated approach:

**Synthetic PHI Generation**: Python's Faker library provided realistic identifiers including names, Social Security numbers, addresses, phone numbers, email addresses, passport numbers, healthcare provider names, and organization details.

**Clinical Context Creation**: A local Ollama instance running GPT-OSS:20B (latest open ai open weight model) generated diverse clinical scenarios incorporating the synthetic PHI elements. This approach ensured realistic clinical language while maintaining complete control over ground truth labels.

**Label Generation**: Each training example contained paired inputs: the original note with embedded PHI (data\_with\_phi) and the corresponding sanitized version (data\_hipaa\_compliant) with appropriate redaction patterns.

**3.2.2 LoRA Configuration and Training**

We employed Low-Rank Adaptation to fine-tune the base Gemma 2B model with the following configuration:

Base Model: google/gemma-2b

LoRA Parameters: rank=8, alpha=16, targeting query and value projections

Training Setup: 3 epochs, learning rate 2e-4, batch size 2 with 8-step gradient accumulation

Hardware: Apple Silicon (MPS) with FP16 disabled for compatibility

The training process emphasized step-by-step reasoning about PHI identification and redaction decisions, encouraging the model to develop systematic approaches to privacy protection.

#### 3.3. System Prompts

**3.3.1 De-Identification model prompts**

For de-identification, both local models are guided by a streamlined prompt, as detailed below.

You are a local language model responsible for enforcing HIPAA compliance by identifying and removing all Protected Health Information (PHI) from clinical text and structured data before it is passed to an external system. Your task is to remove all 18 identifiers defined under HIPAA's Safe Harbor rule while preserving the clinical meaning of the data. Redact all identifiers like names, dates, addresses, SSNs, etc. with placeholder [REDACTED], without summarizing or altering clinical facts.

<data\_with\_phi>{user\_data}</data\_with\_phi>

<data\_hipaa\_compliant>

**3.3.2** **Main Orchestrator Prompt**

The orchestrator prompt defines the behavior of the remote GPT-4.1 model, positioning it as an expert clinical reasoner constrained to operate solely on the input data. The full prompt is provided below.

You are an AI Clinical Reasoning Assistant with expertise in internal medicine. Analyze clinical data and respond to patient-specific questions using structured reasoning.

**Available Tools:**

get\_patient\_observations(): Laboratory test results and clinical observations

get\_patient\_conditions(): Diagnosis history, medical conditions, and medical history

get\_patient\_medications(): Current and past medications

get\_patient\_careplans(): Care plans and treatment plans

get\_patient\_procedures(): Medical procedures and interventions

get\_patient\_imaging\_studies(): Imaging studies and radiology reports

get\_patient\_immunizations(): Vaccination history and immunization records

get\_patient\_allergies(): Allergy information and adverse reactions

Clinical Reasoning Framework:

Analyze the question to determine needed information

Use appropriate tools to gather relevant patient data

Synthesize findings from multiple data sources

Provide clear, evidence-based responses with clinical reasoning

Rules:

Use ONLY data provided by tool outputs

Acknowledge limitations if data is insufficient

Always check for allergies before recommending treatments

#### 3.4. Grounding and Hallucination Mitigation

A primary risk in using LLMs for clinical applications is **hallucination**. Our architecture is explicitly designed to mitigate this risk through a strict **grounding** mechanism. The remote orchestrator is bound by its system prompt to answer questions *only* by using the provided tools and to "Use ONLY data provided by tool outputs." This forces the model to ground all of its analysis exclusively in the de-identified clinical information returned by the local tools. This RAG-like pattern ensures the assistant is always "patient-aware" and not a generic medical chatbot, dramatically reducing the risk of generating dangerous, non-factual clinical information.

### 3.5. Evaluation

Initial validation of the system's efficacy was performed subjectively. The methodology involved processing individual sample queries through the entire pipeline, one prompt at a time, and manually inspecting the output at each stage. This qualitative approach was designed to confirm the integrity of the tool-calling workflow, from query analysis by the orchestrator to the successful invocation of the de-identification component. While this method was effective for iterative development, it lacks the rigor of a formal, quantitative assessment. This initial evaluation was sufficient to reveal the shortcomings of the de-identification model, thereby confirming the need for better, more objective evaluation frameworks in future iterations

### 4. Results

The remote orchestrator consistently demonstrated appropriate clinical reasoning, successfully interpreting queries, selecting relevant data sources, and synthesizing meaningful responses from sanitized inputs. Tool-calling mechanisms operated reliably, and the strict grounding approach effectively prevented hallucination while maintaining clinical utility. However, a qualitative evaluation of the de-identification component revealed a critical limitation. We observed several instances where sensitive data was missed or overly redacted.

**4.1 Comparative De-identification Analysis**

Both local models achieved partial success in PHI removal but exhibited distinct failure patterns that illuminate the challenges inherent in LLM-based de-identification:

**Example Original Note:** Ms. Yaeko Ming Kshlerin, born on 06/07/1999 in Oakes, North Dakota (SSN: 999‑26‑7676), white, non‑Hispanic, female, presented to TOWNER COUNTY MEDICAL CENTER on 12/14/2015 for a medication review due (situation) at age 16; she resides at 523 O'Kon Orchard, Cando, North Dakota 58324 (Towner County, FIPS 38095, latitude 48.4746068, longitude ‑99.2336557) and holds driver’s license S99963354 and passport X23683209X. Dr. Jarvis Ankunding, a general practitioner, recorded the encounter as an outpatient check‑up (procedure) with a base cost of $96.45, a total claim of $1,257.75 fully covered by the payer, while her cumulative healthcare expenses amount to $127,546.31 against coverage of $673,780.87. The encounter was classified under SNOMED‑CT code 314529007 (Medication review due) and was documented as an outpatient encounter with a provider specialty of general practice. Follow‑up was scheduled for a future date to review medication adherence and any new concerns.

**4.1.1 Fine-tuned Gemma 2B**

**Example Redacted Note:**

Patient: [REDACTED], born on [REDACTED] in [REDACTED], white, non‑Hispanic, female, presented to [REDACTED] on [REDACTED] for a medication review due at age [REDACTED]; she resides at [REDACTED] and holds driver’s license [REDACTED] and passport [REDACTED]. Dr. [REDACTED], a general practitioner, recorded the encounter as an outpatient check‑up with a base cost of $96.45, a total claim of $1,257.75 fully covered by the payer, while her cumulative healthcare expenses amount to $127,546.31 against coverage of $673,780.87. The encounter was classified under SNOMED‑CT code 314529007 (Medication review due) and was documented as an outpatient encounter with a provider specialty of general practice. Follow‑up was scheduled for a future date to review medication adherence and any new concerns.

**Evaluation:**

Across all cases, the fine-tuned model consistently achieved high recall, successfully removing explicit identifiers including patient names, SSNs, dates of birth, facility names, provider names, addresses, geographic coordinates, and identification numbers (driver’s license, passport). In Example 1, it also removed embedded identifiers within parentheses and non-standard formats, demonstrating robust pattern recognition. However, this aggressive approach occasionally over-redacted non-PHI content, such as allergy types and patient age at vaccination, thereby reducing the clinical richness of the text. Overall, the fine-tuned variant favored maximal privacy protection at the cost of occasional information loss, and exhibited incomplete coverage for certain numeric identifiers.

**4.1.2** **Base Gemma 3B**

**Example Redacted Note:**

Patient: Ms. [REDACTED], born on [REDACTED] in [REDACTED], [REDACTED], white, non‑Hispanic, female, presented to [REDACTED] on [REDACTED] for a medication review due (situation) at age [REDACTED]; she resides at [REDACTED], [REDACTED] ([REDACTED], FIPS [REDACTED], latitude [REDACTED], longitude [REDACTED]) and holds driver’s license [REDACTED] and passport [REDACTED]. Dr. [REDACTED] recorded the encounter as an outpatient check‑up (procedure) with a base cost of $[REDACTED], a total claim of $[REDACTED] fully covered by the payer, while her cumulative healthcare expenses amount to $[REDACTED] against coverage of $[REDACTED]. The encounter was classified under SNOMED‑CT code [REDACTED] and was documented as an outpatient encounter with a provider specialty of general practice. Follow‑up was scheduled for a future date to review medication adherence and any new concerns.

**Evaluation:**

The base model adopted a more conservative redaction strategy, generally preserving non-PHI clinical details such as allergy descriptions, encounter classifications, and procedure codes. This yielded outputs with higher clinical fidelity but at the cost of reduced recall for PHI. We also noticed that this model was not consistent and frequently failed to redact obvious identifiers across different instances, even when presented in similar formats. For example, in one case it correctly removed a patient’s full name and SSN but left in financial amounts, coverage totals, and certain medical codes; in another, it entirely missed geographic coordinates, driver’s license numbers, passport numbers, and provider names. This inconsistency poses a significant compliance risk, as even a single missed identifier can undermine the privacy guarantees of the entire de-identification process.

**4.2 Clinical Analysis**

**Sample query: The patient wants to try and get pregnant, is there any relevant medical history or medications that should be considered?**

**Final Response:**

Executing tool call: get\_patient\_conditions

Executing tool call: get\_patient\_medications

executing tool call: get\_patient\_allergies

Summary of findings:

- Medical conditions: The available data does not specify any chronic medical conditions or diagnoses relevant to pregnancy (such as diabetes, hypertension, thyroid disease, or reproductive disorders).

- Medications: The patient has previously received Acetaminophen 300 mg / Hydrocodone Bitartrate 5 mg oral tablets. This is a combination opioid analgesic, which is not recommended during pregnancy due to potential risks to the fetus (including neonatal opioid withdrawal syndrome and possible teratogenicity).

- Allergies: The patient has a history of moderate rhinoconjunctivitis and mild skin eruptions, suggesting possible allergic tendencies, but no specific drug allergies are documented.

Clinical reasoning:

- If the patient is still taking Acetaminophen/Hydrocodone, this medication should be reviewed and ideally discontinued or substituted with a safer alternative prior to conception.

- Allergic tendencies may be relevant if considering medications or supplements during pregnancy.

- No other significant medical history is documented that would directly impact pregnancy planning.

Recommendations:

1. Review current medication use—if the patient is still on opioid analgesics, discuss tapering and alternative pain management strategies.

2. Consider preconception counseling, including review of immunization status, folic acid supplementation, and screening for other medical conditions not listed here.

3. Monitor for allergic reactions if new medications or prenatal vitamins are started.

Limitations:

- The data does not provide a comprehensive list of all medical conditions or a current medication list. Additional information about current health status, menstrual/reproductive history, and other medications would be helpful.

If you need a more detailed assessment, please provide or request additional information regarding current medications, medical conditions, or recent laboratory results.

**Evaluation:**

In this test, the remote orchestrator understood the pregnancy-related question and ran the right tools to check conditions, medications, and allergies. It built its answer entirely from that data, which shows the overall design is working. The orchestrator can identify what information it needs, retrieve it from local sources, and follow the rule of not making up answers.

The usefulness of the answer was reduced by an upstream problem. Allergy information existed in the original note, but it was removed during local de-identification before reaching the orchestrator. Without that detail, the system could only report that no allergy information was available and recommend getting more data. While this is a safe fallback that avoids incorrect conclusions, it is a clear example of how small mistakes in de-identification can limit the completeness of the final clinical advice.

The medication review worked well. From the retrieved data, the orchestrator identified that acetaminophen/hydrocodone is not recommended during pregnancy and suggested considering safer alternatives. It also noted the absence of major chronic conditions and provided reasonable pre-pregnancy guidance such as checking immunizations and starting folic acid.

This example shows both strengths and weaknesses. On the positive side, the orchestrator makes correct tool calls, bases its answers on available data, and offers relevant, safe advice when the data is complete. On the negative side, the quality of its answers depends heavily on accurate local de-identification. If that process removes important clinical details, the reasoning will be limited.

### 5. Conclusion

Our work confirms the architectural viability of a hybrid LLM system for privacy-preserving clinical assistance. By delegating the mechanical task of PHI removal to a local model, our approach presents a promising paradigm for leveraging state-of-the-art AI in a HIPAA-compliant manner.

However, our primary finding is a cautionary one: **achieving complete and reliable de-identification using a single, general-purpose local LLM is a non-trivial and, with current technology, unsolved challenge.** The inconsistencies and observed false negatives are critical failures that render the system unsuitable for production deployment. Consequently, the contribution of this work pivots from demonstrating a turnkey solution to articulating a core problem for the field: purely generative, prompt-based de-identification currently lacks the reliability required for sensitive clinical data.

### 6. Future Work

The limitations identified in this project directly inform the most critical areas for future research.

**Hybrid De-identification Engine:** The failure of the LLM-only approach makes a hybrid model the most logical next step. This would involve a pipeline where data is first passed through a **rule-based filter** (using regex) and a **specialized NER model** (medspaCy). The output would then be passed to the local LLM, which would focus on its strengths: catching nuanced PHI and performing the final de-identification and output transformation.

**Evaluating Advanced Open-Source Models:** Newer, more powerful open-source models should be evaluated for the local de-identification role, as they may offer superior instruction-following and accuracy.

**EHR Integration via FHIR:** A crucial step toward validating the system's real-world utility is to move beyond the clean, structured synthetic dataset used in this project. Future work must focus on integrating the assistant with genuine clinical data sources, such as Electronic Health Records (EHRs), by leveraging interoperability standards like HL7 FHIR. This integration must intentionally include diverse data from various clinical settings, specialties, and patient populations. Using diverse data is essential to test the robustness and generalizability of both the de-identification and reasoning models, ensuring they can handle the wide variations in documentation styles, terminologies, and data formats found in practice. Successfully processing this heterogeneous data will be critical for demonstrating that the system is not merely a proof-of-concept but a viable tool for clinical deployment.

**Formal Clinical Validation:** A critical next step is to collaborate with clinicians to formally validate the accuracy, relevance, and clinical utility of the assistant's responses through structured evaluations.

**Enhanced Privacy Auditing and Human-in-the-Loop:** A production-grade system would require a comprehensive logging and auditing mechanism, potentially flagging low-confidence de-identification outputs for human review to provide a final layer of safety and ensure compliance.

**REFERENCES**

[1] Brown, T., et al. "Language Models are Few-Shot Learners." Advances in Neural Information Processing Systems, 2020.

[2] McMurry, A. J., et al. "SHUTTLE to privacy: Syntactic transformation of health information into less identifiable forms." Methods of Information in Medicine, 2013.

[3] Neumann, M., et al. "ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing." Proceedings of the 18th BioNLP Workshop, 2019.

[4] Alsentzer, E., et al. "Publicly Available Clinical BERT Embeddings." Proceedings of the 2nd Clinical Natural Language Processing Workshop, 2019.

[5] Li, T., et al. "Federated Learning: Challenges, Methods, and Future Directions." IEEE Signal Processing Magazine, 2020.

[6] Yao, S., et al. "ReAct: Synergizing Reasoning and Acting in Language Models." International Conference on Learning Representations, 2023.