Utilizing Large Language Models for Zero-Shot Medical Ontology Extension from Clinical Notes

1st Guanchen Wu
Department of Computer Science
Emory University
guanchen.wu@emory.edu

2nd Yuzhang Xie
Department of Computer Science
Emory University
yuzhang.xie@emory.edu

3rd Huanwei Wu
College of Public Health
Temple University
huanmei.wu@temple.edu

4th Zhe He
School of Information
Florida State University
zhe@fsu.edu

5th Hui Shao
Hubert Department of Global Health
Emory University
hui.shao@emory.edu

6th Xiao Hu
Nell Hodgson Woodruff School of Nursing
Emory University
xiao.hu@emory.edu

7th Carl Yang
Department of Computer Science
Emory University
j.carlyang@emory.edu

***Abstract*—Integrating novel medical concepts and relationships into existing ontologies can significantly enhance their coverage and utility for both biomedical research and clinical applications. Clinical notes, as unstructured documents rich with detailed patient observations, offer valuable context-specific insights and represent a promising yet underutilized source for ontology extension. Despite this potential, directly leveraging clinical notes for ontology extension remains largely unexplored. To address this gap, we propose CLOZE, a novel framework that uses large language models (LLMs) to automatically extract medical entities from clinical notes and integrate them into hierarchical medical ontologies. By capitalizing on the strong language understanding and extensive biomedical knowledge of pre-trained LLMs, CLOZE effectively identifies disease-related concepts and captures complex hierarchical relationships. The zero-shot framework requires no additional training or labeled data, making it a cost-efficient solution. Furthermore, CLOZE ensures patient privacy through automated removal of protected health information (PHI). Experimental results demonstrate that CLOZE provides an accurate, scalable, and privacy-preserving ontology extension framework, with strong potential to support a wide range of downstream applications in biomedical research and clinical informatics.**[1]

***Index Terms*—ontology extension, clinical notes, large language models, entity extraction.**

I. Introduction

Ontologies are formal, structured representations of knowledge that define a set of concepts, entities, and the relationships between them within a specific domain [1] [2]. In the biomedical field, for instance, an ontology may represent diseases and symptoms along with their relationships (e.g., *pneumonia* is a type of *respiratory infection*). As medical knowledge evolves, timely ontology extension—systematically identifying, defining, and integrating new entities and their interrelations into the existing structure—is essential for maintaining accurate and up-to-date representations [3]. For example, during the early stages of the COVID-19 pandemic, terms such as "COVID-19" and related symptoms and complications had to be rapidly incorporated into ontologies like SNOMED CT to support reliable clinical interpretation. Without such updates, downstream systems risk relying on incomplete knowledge, leading to degraded performance and potential clinical harm. Scalable and accurate ontology extension is therefore critical for ensuring the effectiveness of biomedical applications.

Clinical notes, which are unstructured textual documents produced during patient encounters—including discharge summaries, nursing records, radiology reports, and ECG interpretations—represent a rich source of medical knowledge [4]. These narratives capture detailed clinical observations, treatment rationales, patient interactions, and social determinants of health—information frequently absent or poorly represented in other clinical documents such as structured Electronic Health Records (EHRs) [5] [6]. As a result, clinical notes offer valuable, context-specific insights that are highly relevant for ontology extension. By integrating novel medical concepts derived from these narratives, ontologies can achieve greater coverage, granularity, and practical applicability. Despite their potential, directly leveraging clinical notes for ontology extension remains largely unexplored, resulting in the loss of valuable clinical information embedded within these unstructured texts.

To leverage clinical notes for ontology extension, two key steps are essential: (i) Medical Entity Extraction, which identifies medical entities from unstructured clinical text, and (ii) Ontology Extension, which determines the appropriate insertion point for each extracted entity within an existing ontology. Existing approaches to ontology extension typically fall into two categories: (i) Rule-based and statistical methods, which rely on predefined patterns, co-occurrence analysis, and heuristic rules to extract new entities and relationships [7], [8]; and (ii) Learning-based methods, which apply machine learning techniques to infer and integrate new concepts from structured or unstructured data. However, these methods face significant limitations when applied to clinical notes. Rule-based and statistical approaches require substantial manual effort and lack scalability for large, evolving ontologies. They

[1]The implementation details, prompt designs, and codes are available in the anonymous repository at https://github.com/guanchenwu1015/CLOZE.

also perform poorly in capturing complex hierarchies—for instance, correctly placing "triple-negative breast cancer" demands nuanced domain understanding beyond simple lexical patterns. Learning-based methods, on the other hand, depend heavily on large annotated datasets and often lack interpretability, making it difficult to validate updates. Moreover, applying such models to clinical texts raises privacy concerns, as sensitive patient information may be inadvertently exposed.

In this work, we propose CLOZE (CLinical Notes Ontology Zero-shot Extension), a novel framework for hierarchical ontology extension from clinical notes, as illustrated in Figure 1. To the best of our knowledge, this is the first framework designed specifically for hierarchical ontology extension using clinical notes. The design of CLOZE is motivated by three core needs: (i) applying robust de-identification methods to protect patient privacy throughout the pipeline, (ii) leveraging the reasoning capabilities of LLMs to accurately capture complex hierarchical relationships that traditional approaches often overlook, and (iii) harnessing the rich biomedical knowledge embedded in pretrained language models to support a fully end-to-end, zero-shot pipeline that eliminates the need for manual annotation or training, thereby overcoming data scarcity in biomedical domains.

CLOZE consists of two main steps: (i) Medical Entity Extraction and (ii) Hierarchical Ontology Extension. In the Medical Entity Extraction step, CLOZE first uses an LLM as the De-identification Agent to remove PHI (Protected Health Information), aligning with standard privacy regulations (e.g., HIPAA) to ensure safe handling of private data. Then another LLM acts as the Entity-extraction Agent, identifying candidate medical entities from de-identified notes. In the Hierarchical Ontology Extension step, these entities are further integrated into an existing ontology. Specifically, we employ a pretrained biomedical language model to compute semantic embeddings of new entities and existing ontology nodes, enabling the identification of the most semantically relevant anchor points in the ontology. Another LLM serves as the Relation-determination Agent, inferring the relationships between new and existing concepts. Based on this classification, CLOZE recursively inserts new entities at the appropriate hierarchical level, preserving both structural coherence and semantic fidelity.

In our experiments, we evaluate CLOZE using clinical notes from a major U.S. hospital, focusing on disease-related entities for ontology extension. We compare performance against state-of-the-art (SOTA) baseline models to validate CLOZE. Results show that CLOZE effectively and efficiently enables privacy-preserving processing and scalable, zero-shot ontology extension with high accuracy in complex biomedical settings.

II. Related Work

A. Rule-based and Statistical Methods for Ontology Extension

Traditional ontology extension techniques often rely on rule-based systems and statistical analysis to extract new entities, relationships, and hierarchical structures from domain-specific corpora. For example, in the medical treatment domain, Cruanes et al. applied semantic tagging and text mining to extract

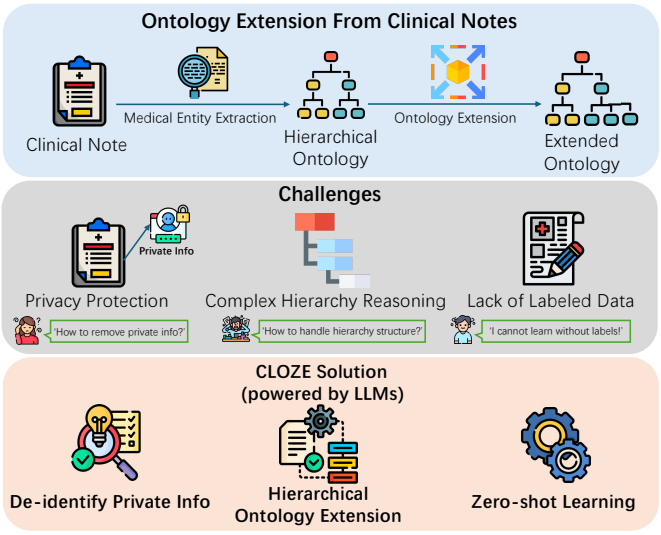

Fig. 1. Illustration of existing ontology extension methods and our proposed framework.

and integrate specialized terms, enabling structured updates to medical ontologies [7]. Phrase2Onto demonstrates the use of phrase-based topic modeling to detect novel concepts in free text and integrate them into ontologies, highlighting the effectiveness of unsupervised statistical techniques in real-world applications [9]. Co-occurrence analysis and word filtering have also been applied to identify candidate entities, followed by subsumption analysis to embed them within shallow ontology hierarchies [10].

While rule-based and statistical methods are often interpretable and lightweight, they typically require significant manual tuning, domain-specific knowledge, and validation by experts. These factors limit their scalability and adaptability.

B. Learning-based Methods for Ontology Extension

Learning-based approaches, including traditional machine learning and deep learning models, offer data-driven alternatives to rule-based systems for ontology extension. Early efforts used supervised classifiers (such as Bayesian models) and clustering techniques (such as k-means) to discover and integrate new concepts into domain ontologies [11]. Hybrid pipelines combining search engines, multilayer perceptrons (MLPs), Naive Bayes, and neural networks were also proposed to infer relationships between candidate entities [10].

With the rise of deep learning, more advanced models have been introduced to capture semantic and structural relationships. Word embeddings, such as Word2Vec, combined with hierarchical clustering, have enabled the automatic grouping of similar biomedical terms [8]. Attention-based models and visual explanation mechanisms have also been employed to improve interpretability and concept classification in chemistry and biomedicine [12]. In the biomedical domain, Pesquita and Couto proposed a supervised method that predicts which regions of an ontology are likely to require extension by analyzing historical versions and external annotations [13].

Despite their promise, learning-based methods often depend on large, annotated datasets and tend to be difficult to interpret,

which complicates the validation of their outputs. Furthermore, when applied to clinical text, these models may inadvertently compromise patient privacy by exposing sensitive information.

C. Pre-trained Language Models (PLMs)

Considering the prohibitive cost of training from scratch, researchers have turned to pretrained language models (PLMs)—such as BERT and GPT—which leverage massive corpora to produce rich, context-sensitive embeddings. In the biomedical domain, further specialization has led to models such as BioBERT [14] and SapBERT [15]. SapBERT, in particular, incorporates a self-alignment objective to preserve semantic similarity within its embeddings, making it one of the SOTA models for generating semantic representations used in biomedical ontology alignment.

More recently, LLMs have demonstrated strong zero- and few-shot performance on entity extraction and ontology extension tasks [16] [17]. For example, Wu et al. [18] proposed a zero-shot extension framework that combines LLM agents with online hierarchical clustering to integrate medical symptom concepts from patient forums. However, LLM-based approaches often struggle with deeper, multi-layer ontologies [19]. Their flat textual inputs and limited context windows impede the explicit modeling of hierarchical structure, leading to hallucinations or misclassifications when scaling beyond a few ontology levels.

III. Method

A. Problem Formulation

We address the task of extracting novel medical entities from clinical notes and integrating them into an existing hierarchical ontology in a privacy-preserving and zero-shot manner. Let $D = \{D_1, D_2, \ldots, D_N\}$ denote a collection of clinical notes, where D_i is the i-th document and N is the total number of notes.

We are also given a medical ontology $O = (E, R, \prec)$, where E is the set of existing medical entities, $R \subseteq E \times E$ is the set of ontological relationships (e.g., "is-a", "part-of"), and $\prec$ denotes a partial order capturing the hierarchical structure with depth k.

The goal is to extract a set of novel medical entities $E^{\text{new}} = \{e_1, e_2, \ldots, e_M\}$ from D, where $e_j \notin E$, and to generate candidate mappings $A = \{(e_j, e_p) \mid e_p \in E\}$ such that each new entity e_j is assigned a parent e_p in the ontology hierarchy. The final output is an extended ontology $\tilde{O} = (\tilde{E}, \tilde{R}, \prec)$, where:

- $\tilde{E} = E \cup E^{\text{new}}$ is the updated set of entities.
- $\tilde{R} = R \cup R^{\text{new}}$, where $R^{\text{new}} \subseteq E^{\text{new}} \times E$ represents the set of newly established relationships based on predicted parent-child mappings.
- $\prec$ is preserved or updated to maintain a valid hierarchy of depth $\tilde{k} \geq k$.

The entire process is conducted without access to any task-specific labeled data or manually annotated ontological links, thereby enabling scalable and privacy-preserving ontology extension in a zero-shot setting.

B. Framework Overview

To support the integration of novel medical entities into hierarchical ontologies, we propose CLOZE, a two-step framework: (i) Medical Entity Extraction and (ii) Hierarchical Ontology Extension, as illustrated in Figure 2. The extraction step uses two LLM-powered agents: a *De-identification Agent* to remove PHI and ensure privacy, and an *Entity-extraction Agent* to identify medical terms from the de-identified text. In the extension step, a *PLM (SapBERT)* generates embeddings for new entities and candidate ontology nodes, while a *Relation-determination Agent* classifies each pairwise relation to recursively insert new entities at the appropriate hierarchical level, preserving structural and semantic consistency. Further details are provided in Sections III-C and III-D.

C. Medical Entity Extraction

The Medical Entity Extraction step in CLOZE consists of two LLM-powered agents designed to ensure both privacy preservation and comprehensive entity identification. Let $D = \{D_1, D_2, \ldots, D_N\}$ denote the set of raw clinical notes, each potentially containing protected health information (PHI).

a) De-identification Agent: To comply with privacy regulations, we first apply a De-identification Agent to each note D_i, producing a de-identified version D_i^{deid}. Following recent advances in LLM-based PHI de-identification [20], [21], this step removes personal identifiers before downstream processing and is defined as:

$$D_i^{\text{deid}} = f_{\text{deid}}(D_i),$$

where f_{deid} is an LLM prompted to extract a predefined set of PHI types:

$$\mathcal{P} = \{\texttt{person}, \texttt{location}, \texttt{organization}, \texttt{age}, \texttt{phone number}, \texttt{email}, \texttt{date and time}, \texttt{zip}, \texttt{profession}, \texttt{username}, \texttt{id}, \texttt{url}\}$$

from the clinical notes. The LLM agent outputs a JSON dictionary:

$$\texttt{PHI\_dict} = \{p : [e_1^p, e_2^p, \ldots] \mid p \in \mathcal{P}\},$$

where each e_j^p is an extracted entity of PHI type p. This structured output improves interpretability, enables traceability, and promotes high recall, including for borderline PHI instances. Finally, the de-identified note D_i^{deid} is constructed by "masking" these detected PHIs in the raw clinical notes.

b) Entity-extraction Agent: After de-identification, the Entity-extraction Agent processes the de-identified clinical note D_i^{deid} to extract disease-related medical concepts. This step is defined as:

$$E^{\text{new}} = f_{\text{extract}}(D_i^{\text{deid}}),$$

where $E^{\text{new}} = \{e_1^{\text{new}}, e_2^{\text{new}}, \ldots, e_m^{\text{new}}\}$ denotes the set of disease entities extracted from the note. The extraction function f_{extract} is implemented as a large language model (LLM) guided by a structured prompt that emulates the behavior of a clinical

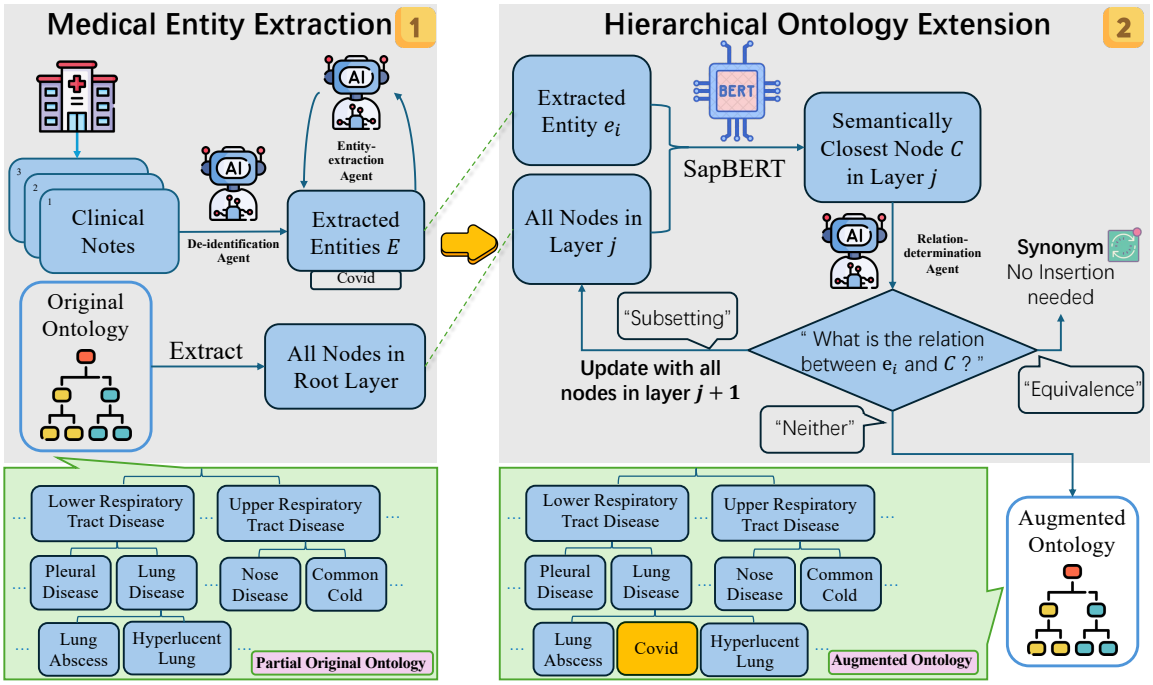

Fig. 2. Framework of CLOZE. This framework has two components: Medical Entity Extraction and Hierarchical Ontology Extension. The first module de-identifies PHI in clinical notes and extracts medical entities using two LLM agents. The second module integrates these entities into the ontology, using SapBERT for semantic matching and an LLM to determine relationships. The process iterates layer by layer until integration is complete.

expert. The prompt is carefully designed to: (i) frame the task specifically as disease-only entity recognition, (ii) instruct the model to return a flat, deduplicated list of disease mentions. This step provides the foundation for downstream ontology integration by producing a curated list of candidate entities for insertion into the ontology.

D. Hierarchical Ontology Extension

After extracting de-identified medical entities, the next step in our framework is to integrate them into an existing hierarchical ontology while preserving semantic meaning and structural consistency. We define the existing ontology as $O = (E, R, \prec)$, where E is the set of ontology nodes (entities), $R \subseteq E \times E$ represents the hierarchical relationships (e.g., "is-a"), and $\prec$ defines the partial ordering over nodes that encodes the ontology's hierarchical structure.

Let $e^{\text{new}} \in E^{\text{new}}$ denote a newly extracted entity from the clinical notes. To determine its appropriate insertion point within the ontology, we adopt a hybrid strategy that combines (i) PLM-based semantic matching, (ii) a Relation-determination Agent for relational classification, and (iii) recursive hierarchical extension. The goal is to identify an anchor node $e^* \in E$ that is most semantically similar to e^{new}, and to assess whether the new entity is either semantically equivalent to, or a more specific subclass of, e^* or one of its descendants. If no suitable child node is identified through recursive extension, which means e^{new} is a new found "child" we extracted from the clinical notes, then e^{new} is inserted as a new "child" node under candidate e^*, thereby maintaining both semantic alignment and structural consistency in the extended ontology.

a) PLM-based Matching: We leverage SapBERT, a SOTA PLM for generating semantic representations in biomedical ontology alignment, to compute embeddings for both the new entity and candidate ontology nodes. Specifically, we obtain an embedding $\mathbf{v}_{e^{\text{new}}} \in \mathbb{R}^d$ for the name of the new entity, and an embedding $\mathbf{v}_e \in \mathbb{R}^d$ for the name of each candidate node $e \in E_l$, where $E_l \subseteq E$ denotes the set of nodes in the current ontology layer l. Cosine similarity is then used to identify the most semantically similar candidate:

$$e^* = \arg\max_{e \in E} \cos\left(\mathbf{v}_{e^{\text{new}}}, \mathbf{v}_e\right).$$

b) Relation-determination Agent: Given the candidate node e^*, the Relation-determination Agent—an LLM prompted with contextual descriptions of e^{new} and e^*—classifies their relationship as one of the following:

- `Equivalence`: $e^{\text{new}} \equiv e^*$, indicating they refer to the same medical concept. The entity is treated as a synonym and is not added.
- `Subsetting`: $e^{\text{new}} \prec e^*$, meaning the new entity is more specific and should be inserted as a child of e^*.
- `Neither`: The entities are unrelated in a hierarchical sense, prompting the search to continue elsewhere in the ontology.

c) Recursive Hierarchical Extension: If the relationship between the new entity and the matched candidate node e^* is classified as `subsetting`, the process recurses into the next layer of the ontology, using the children of e^* as the new

candidate set E_{l+1}. This recursive search continues layer by layer until one of the following occurs:

- The relation is classified as `equivalence`, indicating that e^{new} already exists in the ontology and no insertion is needed.
- The relation is classified as `subsetting`, indicating that e^{new} should be put to the next layer, the search moves to another layer recursively.
- No child nodes of e^* are classified as `subsetting` or `equivalence`, suggesting that e^{new} is more specific than e^* but not semantically related to any of its children. In this case, e^{new} is treated as a new "child" node under e^*.

In either case where insertion is required, the new entity is added as a child of the most recent parent node:

$$\tilde{E} = E \cup \{e^{\text{new}}\}, \quad \tilde{R} = R \cup \{(e^{\text{new}}, e_{\text{parent}})\}.$$

The algorithm of the Hierarchical Ontology Extension step is summarized in Algorithm 1. By combining SapBERT's high-quality biomedical embeddings with the contextual reasoning abilities of the LLM, our framework enables accurate and scalable ontology extension. This hybrid approach ensures that new entities are integrated into deep hierarchical structures with both semantic alignment and structural consistency.

Algorithm 1 Hierarchical Ontology Extension

Require: New extracted entity e^{new}, top-layer candidate set E_0

Ensure: Updated ontology $\tilde{O} = (\tilde{E}, \tilde{R})$

Set layer index $l \leftarrow 0$
while true **do**
  Compute embedding $\mathbf{v}_{e^{\text{new}}}$ and embeddings $\mathbf{v}_e$ for all $e \in E_l$
  Identify closest match:

$$e^* \leftarrow \arg\max_{e \in E_l} \cos(\mathbf{v}_{e^{\text{new}}}, \mathbf{v}_e)$$

  Query Relation-determination Agent for relation between e^{new} and e^*
  if relation is `equivalence` **then**
    return No insertion needed
  else if relation is `subsetting` **then**
    Set $E_{l+1} \leftarrow \text{Children}(e^*)$
    Increment layer: $l \leftarrow l + 1$
  else
    Insert e^{new} as child of e^*
    Update:

$$\tilde{E} \leftarrow E \cup \{e^{\text{new}}\}, \quad \tilde{R} \leftarrow R \cup \{(e^{\text{new}}, e^*)\}$$

    return Updated ontology $\tilde{O} = (\tilde{E}, \tilde{R})$
  end if
end while

IV. Experimental Settings

A. Dataset

Access to real-world clinical notes is highly restricted due to stringent privacy regulations, and fully annotated clinical datasets are particularly costly and difficult to obtain. For this study, we obtained a dataset consisting of 100 fully annotated clinical notes, in which all PHI entities were meticulously identified by multiple medical experts. These notes were generously provided by a large hospital in the U.S., ensuring real-world clinical authenticity and high annotation quality. These notes retain their original structure and include a diverse and detailed range of patient information, making them a rare and invaluable resource for advancing research in clinical informatics. On average, each clinical note contains approximately 1,000 tokens, reflecting a substantial level of detail and comprehensiveness in documenting patient care. Our goal is to extract the disease entities from the notes and integrate them into the hierarchical structure of a disease-specific ontology.

As a foundational resource, the Disease Ontology (DO) was utilized as the seed ontology [22]. DO is a standardized biomedical ontology offering a comprehensive and hierarchical classification of human diseases. It integrates seamlessly with major medical vocabularies and databases, including ICD and UMLS, thereby facilitating robust alignment and analysis of disease-related data. Widely regarded as a critical tool in biomedical research and healthcare, DO supports data integration, analysis, and interoperability across diverse systems, making it an essential resource for advancing ontology-based methodologies in the medical domain.

B. Language Models

For the CLOZE pipeline, we utilize one pre-trained language model (SapBERT) and three LLM-based agents (De-identification Agent, the Entity-extraction Agent, and the Relation-determination Agent). The SapBERT model is obtained from the Hugging Face platform[2] and is used to generate biomedical entity embeddings for semantic matching. All three LLM-based agents are built upon the LLaMA-3-70B-Instruct model[3], a SOTA model from Meta's LLaMA series. These language models are deployed locally, ensuring that sensitive clinical data remains entirely on-premises. This setup effectively addresses privacy concerns and supports compliance with data protection standards. The combination of local deployment and advanced language understanding enables efficient and confidential processing of clinical notes.

For the evaluation in Section V-B, considering the lack of ground truth, we also incorporated the GPT-4-0613 model[4] via Microsoft's Azure OpenAI Service[5]. Azure is a cloud-based platform provides robust safeguards, including end-to-

[2]https://huggingface.co/cambridgeltl/SapBERT-UMLS-2020AB-all-lang-from-XLMR

[3]https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

[4]https://platform.openai.com/docs/models/gpt-4

[5]https://azure.microsoft.com/en-us/products/ai-services/openai-service

end encryption, isolated execution environments, and HIPAA-compliant infrastructure. These features ensure that both prompt inputs and model outputs remain secure and inaccessible to external parties, including OpenAI. By leveraging Azure's secure environment, we were able to utilize the powerful reasoning capabilities of GPT models while maintaining the confidentiality of PHI. This configuration offers a reliable and compliant solution for deploying language model inference in real-world clinical contexts.

C. Baseline Methods

Since no existing work directly addresses ontology extension from clinical notes, we evaluate our framework by assessing each component separately.

In Section V-A, we evaluate the two agents (De-identification Agent and Entity-extraction Agent) in the Medical Entity Extraction step, each against relevant baseline methods. For the De-identification Agent, we exclude traditional learning-based approaches due to the limited availability of large annotated datasets required for training. Instead, we compare our design using LLaMA-3-70B-Instruct against the following zero-shot baselines for PHI annotation:

- PhysioNet: A widely used rule-based method, implemented via the PhysioNet de-identification software package[6], which uses regular expressions and dictionary-based lookups to identify and redact PHI in clinical text [23].
- LLaMA-3-8B-Instruct: A compact 8B-parameter version[7] of Meta's LLaMA-3 series, fine-tuned for instruction-following tasks and efficient inference with limited computational resources.
- GPT-3.5-turbo-0301: A cost-effective version[8] of the GPT-3.5 family, optimized for fast responses and general-purpose natural language understanding tasks, including few-shot and zero-shot reasoning.
- GPT-4-0613: A state-of-the-art LLM known for its superior reasoning, comprehension, and instruction-following capabilities. The 0613 version[9] is tailored for stable function-calling and prompt engineering in NLP tasks.

For the Entity-extraction Agent, we identify disease-related entities from de-identified clinical notes. For both agents, we evaluate the extracted disease entities using a standardized vocabulary from the Disease Ontology, ensuring consistent and ontology-aligned comparisons across methods. We compare our LLM-based approach (LLaMa-3-70B-Instruct) against:

- Stanza [24]: A widely adopted biomedical NLP toolkit[10] that performs named entity recognition on clinical text.
- BioEN [25]: A recent model[11] based on fine-tuned DistilBERT, trained on the MACCROBAT 2020 dataset. It is designed to extract a wide range of biomedical and epidemiological entities without relying on external ontologies for concept normalization.

In Sections V-B and V-C, we evaluate the performance of our Hierarchical Ontology Extension step. We compare our full method (SapBERT + LLM-Hierarchical) against the following three baseline strategies:

- LLM-Onetime: This method prompts a single large language model to perform ontology extension in one step. Due to context window limitations, the ontology is divided into segments. For each segment, the LLM identifies the node most similar to the new entity. The node with the highest overall similarity is selected, and its parent-child structure is used to determine the insertion point.
- LLM-Hierarchical: This baseline simulates hierarchical extension by prompting the LLM with one ontology layer at a time. At each step, the model selects the most semantically similar node and classifies the relationship as `subsetting`, `equivalence`, or `neither`. If classified as `subsetting`, the search continues recursively into the children of the selected node.
- SapBERT + LLM-Onetime: This variant combines SapBERT and a single-step LLM prompt. SapBERT is used to compute embeddings for both the new entity and all ontology nodes. The node with the highest cosine similarity is selected, and an LLM is prompted once to classify the relationship between the new entity and the selected node.

D. Evaluation Metrics

In Section V-A, we report precision, recall, and F1 score to evaluate the effectiveness of our design in the Medical Entity Extraction step. For the experiment evaluating the De-identification Agent, the ground truth consists of manually annotated PHI entities in the dataset. For the Entity-extraction Agent, an approximate ground truth is constructed by scanning each clinical note and identifying Disease Ontology (DO) terms that appear in the text using fuzzy string matching with a high similarity threshold (e.g., 90). These matched DO terms serve as the reference set. Predicted entities from each method are then compared against this reference using a range of relaxed fuzzy thresholds (e.g., from 60 to 80) to assess robustness under varying levels of string similarity.

In Section V-B, considering the lack of ground truth, we evaluate the Hierarchical Ontology Extension step using GPT-4-0613 for automatic assessment against baseline methods. For each updated ontology, the language model classifies every newly extended entity relationship as "Correct," "Incorrect," or "Not Sure." Precision is computed as the number of correct predictions divided by the total number of labeled predictions, excluding those marked as uncertain. This precision metric serves as the primary indicator of how accurately each method integrates new entities into the ontology structure.

In Section V-C, we further evaluate the effectiveness of the Hierarchical Ontology Extension step through human assessment. A random sample of newly inserted entity relation-

[6]https://www.physionet.org/content/deid/1.1/
[7]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct
[8]https://platform.openai.com/docs/models/gpt-3.5-turbo
[9]https://platform.openai.com/docs/models/gpt-4
[10]https://stanfordnlp.github.io/stanza/
[11]https://huggingface.co/MilosKosRad/BioNER

ships is manually reviewed based on four criteria: relevance (whether the entity is semantically appropriate within the disease ontology), accuracy (whether the hierarchical relationship is correctly specified), importance (whether the new entity adds meaningful value to the ontology), and overall satisfaction (the overall quality and consistency of the integrated triplet).

V. Experimental Results

A. Medical Entity Extraction

We begin by evaluating the performance of our two agents (the De-identification Agent and the Entity-extraction Agent) in the Medical Entity Extraction step. Results are presented in Table I and Table II, and compared against the baseline methods introduced in Section IV-C using the evaluation metrics defined in Section IV-D.

TABLE I
De-identification Agent evaluation results. Precision, Recall, and F1 score across different models. Best results are bolded.

Method	Precision	Recall	F1
PhysioNet	0.39	0.24	0.28
GPT-3.5-turbo	0.43	0.60	0.48
GPT-4	0.53	**0.69**	0.58
LLaMA-3-8B-Instruct	0.46	0.59	0.50
LLaMA-3-70B-Instruct(Ours)	**0.60**	0.68	**0.62**

TABLE II
Entity-Extraction Agent evaluation results. Precision (Pr), Recall (Re), and F1 score under different similarity thresholds. Best results are bolded.

Method	Threshold = 60			Threshold = 70			Threshold = 80		
	Pr	Re	F1	Pr	Re	F1	Pr	Re	F1
Stanza	0.1689	**0.3904**	0.2358	0.1373	0.3173	0.1917	0.1138	0.2630	0.1589
BioEN	0.2082	0.1065	0.1409	0.1469	0.0752	0.0994	0.1102	0.0564	0.0746
LLM(Ours)	**0.2120**	0.3528	**0.2649**	**0.1932**	**0.3215**	**0.2414**	**0.1606**	**0.2672**	**0.2006**

Table I reports the results of the De-identification Agent evaluated against SOTA zero-shot baselines. While the rule-based PhysioNet baseline exhibits limited ability to identify sensitive entities, all LLM-based methods achieve significantly better performance. Among them, our design based on LLaMA-3-70B-Instruct outperforms all other models, supporting its use as the De-identification Agent.

Table II shows the evaluation of the Entity-extraction Agent compared to SOTA baseline methods. Our model consistently outperforms both Stanza and BioEN in terms of F1 score across all similarity thresholds. While Stanza achieves the highest recall, especially at lower thresholds, its precision remains lower than that of the Entity-extraction Agent, reflecting a higher number of false positives. BioEN performs well in precision at the lowest threshold but suffers from very low recall, likely due to its broader but less disease-focused entity coverage. As the fuzzy matching threshold increases, precision decreases for all methods, reflecting stricter matching with ontology terms. The Entity-extraction Agent maintains a strong balance between precision and recall, leading to the highest F1 scores across all thresholds.

B. Hierarchical Ontology Extension: LLM Evaluation

We evaluate the effectiveness of our Ontology Extension design against the baseline methods described in Section IV-C, using the evaluation metrics detailed in Section IV-D.

TABLE III
Hierarchical Ontology Extension LLM Evaluation Results. Comparison of precision and the number of Correct, Incorrect, and Not Sure predictions across our method and three baselines. Best results are bolded.

Models	Correct	Incorrect	Not Sure	Precision
LLM-Onetime	293	461	99	38.859
LLM-Hierarchical	268	463	53	36.662
SapBERT+LLM-Onetime	241	318	322	43.113
SapBERT+LLM-Hierarchical(Ours)	617	158	106	**79.613**

Table III presents the performance of our method's design (SapBERT+LLM-Hierarchical) compared to the three baselines. Our method achieves the highest precision, approaching 80%, and produces the largest number of correct entity-node triplets. LLM-Onetime and LLM-Hierarchical suffer from incorrect placements, likely due to their lack of domain-specific embeddings and inability to reason over hierarchical structures. Although they produce fewer ambiguous cases labeled "Not Sure", they tend to misclassify the correct location, especially in deeper hierarchies. SapBERT+LLM-Onetime improves similarity matching via domain-specific embeddings but still lacks layer-wise reasoning, resulting in fewer correct placements compared to our proposed hierarchical ontology extension method.

These results highlight the importance of combining domain-specific semantic embedding (via SapBERT) and hierarchical reasoning (via LLM) to achieve accurate and scalable ontology extension.

C. Hierarchical Ontology Extension: Human Evaluation

In addition to automated evaluations in Section V-B, we conducted a comprehensive human evaluation to assess the semantic relevance and hierarchical accuracy of the extended ontology. Two annotators with expertise in biomedicine terminology, recruited from a nursing school affiliated with a major research university, participated in the evaluation process. From the ontology extension results produced by all four methods, we randomly sampled 100 new nodes that appeared in the outputs. For each sampled node, we extracted its hierarchical context and constructed evaluation triplets in the form: {*new_node, node_from_original_ontology, relationship*}.
Each triplet was rated by both annotators on a 3-point scale (0–2), where 0 indicates that the criterion was not met, 1 denotes partial alignment, and 2 represents full alignment. Ratings were assigned based on four evaluation criteria: *Relevance* (whether the new entity is semantically appropriate within the disease ontology), *Accuracy* (whether the hierarchical relationship is correctly specified), *Importance* (whether the new entity is a meaningful addition to the ontology), and *Overall Satisfaction* (the overall quality and consistency of the triplet).

TABLE IV
HIERARCHICAL ONTOLOGY EXTENSION HUMAN EVALUATION RESULTS. COMPARISON ACROSS RELEVANCE (REL.), ACCURACY (ACC.), IMPORTANCE (IMP.), AND OVERALL. BEST RESULTS ARE BOLDED.

Models	Rel.	Acc.	Imp.	Overall
LLM-Onetime	0.09	0.67	0.63	0.06
LLM-Hierarchical	0.09	0.96	0.63	0.09
SapBERT+LLM-Onetime	1.04	1.53	1.00	1.00
SapBERT+LLM-Hierarchical(Ours)	**1.05**	**2.00**	**1.87**	**1.87**

The results, summarized in Table IV, show that our proposed hierarchical ontology extension framework achieved the highest average ratings across all four evaluation metrics. Specifically, it outperformed the second-best model, SapBERT+LLM-Onetime, by an average of 51.2% across the metrics. Both LLM-Onetime and LLM-Hierarchical methods showed notably low performance on the Relevance metric, highlighting a key limitation of using LLMs without domain-specific pretraining. Without adequate grounding in biomedical knowledge, these models struggle to capture semantically appropriate relationships, regardless of whether they incorporate hierarchy. Furthermore, our framework achieved the highest score on the Accuracy metric, demonstrating its ability to identify precise hierarchical relations between medical entities.

VI. CONCLUSION

This study presents CLOZE, a novel zero-shot hierarchical ontology extension framework tailored for clinical notes, leveraging the power of LLMs to address key challenges including privacy-preserving preprocessing, hierarchical complexity, and data scarcity. Experimental results demonstrate that CLOZE consistently outperforms baseline methods across multiple evaluation settings, highlighting its effectiveness in aligning new medical concepts with existing ontological structures. Despite the promising results, the current evaluation is limited by a small dataset of 100 clinical notes due to the labor-intensive nature of privacy-compliant annotation. Ongoing efforts focus on expanding the dataset and improving methodological robustness. Future directions include enhancing scalability, interpretability, and alignment with clinical decision-making to support real-world applications in healthcare.

REFERENCES

[1] A. Luschi, C. Petraccone, G. Fico, L. Pecchia, and E. Iadanza, "Semantic ontologies for complex healthcare structures: a scoping review," *IEEE Access*, vol. 11, pp. 19 228–19 246, 2023.

[2] Y. Xie, X. Han, R. Xu, X. Hu, J. Lu, and C. Yang, "Hypkg: Hypergraph-based knowledge graph contextualization for precision healthcare," in *International Semantic Web Conference*. Springer, 2025, pp. 328–348.

[3] S. Althubaiti, Ş. Kafkas, M. Abdelhakim, and R. Hoehndorf, "Combining lexical and context features for automatic ontology extension," *Journal of biomedical semantics*, vol. 11, pp. 1–13, 2020.

[4] T. M. Seinen, J. A. Kors, E. M. van Mulligen, and P. R. Rijnbeek, "Using structured codes and free-text notes to measure information complementarity in electronic health records: Feasibility and validation study," *Journal of Medical Internet Research*, vol. 27, p. e66910, 2025.

[5] M. Tayefi, P. Ngo, T. Chomutare, H. Dalianis, E. Salvi, A. Budrionis, and F. Godtliebsen, "Challenges and opportunities beyond structured data in analysis of electronic health records," *Wiley Interdisciplinary Reviews: Computational Statistics*, vol. 13, no. 6, p. e1549, 2021.

[6] Y. Xie, H. Cui, Z. Zhang, J. Lu, K. Shu, F. Nahab, X. Hu, and C. Yang, "Kerap: A knowledge-enhanced reasoning approach for accurate zero-shot diagnosis prediction using multi-agent llms," *arXiv preprint arXiv:2507.02773*, 2025.

[7] J. Cruanes, "Ontology extension and population: an approach for the pharmacotherapeutic domain," in *Natural Language Processing and Information Systems: 16th International Conference on Applications of Natural Language to Information Systems, NLDB 2011, Alicante, Spain, 2011. Proceedings 16*. Springer, 2011.

[8] A. S. Behr, M. Völkenrath, and N. Kockmann, "Ontology extension with nlp-based concept extraction for domain experts in catalytic sciences," *Knowledge and Information Systems*, pp. 5503–5522, 2023.

[9] M. A. N. Pour, H. Li, R. Armiento, and P. Lambrix, "Phrase2onto: a tool to support ontology extension," *Procedia Computer Science*, vol. 225, pp. 1415–1424, 2023.

[10] W. He, S. Li, and X. Yang, "A hybrid approach for extending ontology from text," in *CCF International Conference on Natural Language Processing and Chinese Computing*. Springer, 2013, pp. 255–265.

[11] Q. Song, J. Liu, X. Wang, and J. Wang, "A novel automatic ontology construction method based on web data," in *2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, 2014.

[12] A. Memariani, M. Glauer, F. Neuhaus, T. Mossakowski, and J. Hastings, "Automated and explainable ontology extension based on deep learning: A case study in the chemical domain," *arXiv:2109.09202*, 2021.

[13] C. Pesquita and F. M. Couto, "Predicting the extension of biomedical ontologies," *PLOS Computational Biology*, 2012.

[14] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "Biobert: a pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020.

[15] F. Liu, E. Shareghi, Z. Meng, M. Basaldella, and N. Collier, "Self-alignment pretraining for biomedical entity representations," *arXiv preprint arXiv:2010.11784*, 2020.

[16] Y. Xie, J. Lu, J. Ho, F. Nahab, X. Hu, and C. Yang, "Promptlink: leveraging large language models for cross-source biomedical concept linking," in *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, pp. 2589–2593.

[17] B. Bhasuran, Q. Jin, Y. Xie, C. Yang, K. Hanna, J. Costa, C. Shavor, W. Han, Z. Lu, and Z. He, "Preliminary analysis of the impact of lab results on large language model generated differential diagnoses," *npj Digital Medicine*, vol. 8, no. 1, p. 166, 2025.

[18] G. Wu, C. Ling, I. Graetz, and L. Zhao, "Ontology extension by online clustering with large language model agents," *Frontiers in Big Data*, vol. 7, p. 1463543, 2024.

[19] S. Tsaneva, S. Vasic, and M. Sabou, "Llm-driven ontology evaluation: Verifying ontology restrictions with chatgpt," *The Semantic Web: ESWC Satellite Events*, vol. 2024, 2024.

[20] G. Wu, Z. Chen, Y. Xie, and C. Yang, "Towards automatic evaluation and selection of phi de-identification models via multi-agent collaboration," *arXiv preprint arXiv:2510.16194*, 2025.

[21] G. Wu, L. Zheng, H. Xie, Z. Xiang, J. Lu, D. Liu, D. Bold, B. Li, X. Hu, and C. Yang, "Large language model empowered privacy-protected framework for phi annotation in clinical notes," *arXiv preprint arXiv:2504.18569*, 2025.

[22] J. A. Baron, C. S.-B. Johnson, M. A. Schor, D. Olley, L. Nickel, V. Felix, J. B. Munro, S. M. Bello, C. Bearer *et al.*, "The do-kb knowledgebase: a 20-year journey developing the disease open science ecosystem," *Nucleic acids research*, vol. 52, no. D1, pp. D1305–D1314, 2024.

[23] I. Neamatullah, M. M. Douglass, L.-W. H. Lehman, A. Reisner, M. Villarroel, W. J. Long *et al.*, "Automated de-identification of free-text medical records," *BMC medical informatics and decision making*, vol. 8, no. 1, p. 32, 2008.

[24] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A python natural language processing toolkit for many human languages," *arXiv preprint arXiv:2003.07082*, 2020.

[25] S. Raza, D. J. Reji, F. Shajan, and S. R. Bashir, "Large-scale application of named entity recognition to biomedicine and epidemiology," *PLOS Digital Health*, vol. 1, no. 12, p. e0000152, 2022.