Normalizing Health Concepts with Biomedical Embedding and LLMs

Iram Azam[1], Keyuan Jiang[1], Gordon B. Bernard[2]
[1]Purdue University Northwest, Hammond, IN, USA
[2]Vanderbilt University, Nashville, TN, USA
{iazam, kjiang}@pnw.edu, gordon.bernard@vumc.org

Abstract

Accurate normalization of health-related expressions to standardized biomedical concepts is crucial for both healthcare and biomedical research. However, traditional string-based matching methods are limited by lexical variations. In this study, we propose a neural embedding-based normalization framework that utilizes an embedding model trained on biomedical terminology, generating over 3.59 million embeddings corresponding to UMLS terms and Concept Unique Identifiers (CUIs). For clinical data, CUIs were retrieved via semantic matching, while Twitter phrases were first processed using a large language model (LLM) to generate preferred terms prior to embedding-based CUI retrieval. Our approach substantially outperforms exact string matching and MetaMap Lite. For clinical data (3,144 phrases), normalization accuracy improved from 0.679 (string match) and 0.574 (MetaMap Lite) to 0.858. For Twitter data (102 phrases), accuracy increased from 0.235 (string match) and 0.118 (MetaMap Lite) to a range of 0.882 (Gemini 2.5 Flash) to 0.980 (GPT-4o mini). These findings highlight both the effectiveness of embedding-based semantic retrieval and the ability of LLMs to generate preferred terms, enhancing robustness in health concept normalization across diverse text sources.

1 Introduction

Health concepts can be found in free text from various sources, such as electronic health records (EHRs) and social media networks, and are often expressed in many different ways. Normalizing health concepts is a process of mapping various health terms or phrases to normalized, unique concept identifiers within an ontology (French and McInnes, 2023; Chen et al., 2025). This normalization is of unique importance in healthcare and biomedical research by enabling interoperability and integration of diverse medical/health data with a standard language.

Biomedical concept normalization plays a central role in clinical and public-health applications. It supports tasks such as tracking of emerging symptoms in outbreaks like the Coronavirus (Wu et al., 2023), detecting adverse drug reactions (Karimi et al., 2015), enabling clinical decision making (García-Barragán et al., 2025) and conducting large-scale epidemiological analysis (Wang et al., 2021).

In EHRs, medical concepts are typically expressed using medical terminology, as they are documented by healthcare workers. However, on social media, health concepts are largely expressed in *layman's terms*, posing more challenges for normalization. Although social media texts may contain valuable details, they often have repetitive phrases, inconsistent terminology, and vague wording, making automated data analysis and interoperability challenging (Scepanovic et al., 2020). This kind of variation in language causes information to be fragmented–identical concepts appearing in different phrases in datasets, limiting the effectiveness of natural language processing (NLP) systems in healthcare applications (Newman-Griffis et al., 2021).

The Unified Medical Language System (UMLS) Metathesaurus, developed by the U.S. National Library of Medicine, serves as the primary reference framework for biomedical concept normalization by grouping synonymous expressions from nearly 190 vocabularies under a single Concept Unique Identifier (CUI). Each CUI represents a distinct concept and is associated with defined semantic types and relationships, allowing diverse natural language expressions to be mapped to a single standardized concept (U.S. National Library of Medicine, 2016b).

The UMLS Metathesaurus integrates professional vocabularies like the International Classification of Diseases (ICD) and the Systematized Nomenclature of Medicine–Clinical Terms

Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026), pages 180–190

(SNOMED CT) but also includes layman's terms through the Consumer Health Vocabulary (CHV) (Zeng and Tse, 2006). However, since CHV's last update in 2011, it lacks coverage of evolving consumer language. This gap underscores the need for dynamic methods to map informal expressions to standardized biomedical concepts (He et al., 2017).

A recent survey of UMLS users highlights its broad adoption in academic and research settings, with the Metathesaurus among the most frequently accessed components and text processing for concept mapping and linking cited as its primary application (Amos et al., 2020).

UMLS has also been applied beyond formal clinical and research settings to interpret consumer-generated health content. Prior work has shown that informal social media expressions can be mapped to SNOMED IDs and ICD codes using neural models (Tutubalina et al., 2018), and that UMLS can support the extraction of medical concepts and semantic types from layman's vocabularies in online forums (Anik et al., 2024).

The importance of UMLS-based concept normalization is further highlighted in social media mining, where layman symptom expressions often differ substantially from clinical terminology (Manousogiannis et al., 2020) and include spelling errors, creative short texts, slang, figurative language, and paraphrasing (Tamine and Goeuriot, 2021). Clinical notes in EHRs contain shorthand, acronyms, abbreviations, and institution-specific jargon (Luo et al., 2020). Both settings highlight the difficulty of achieving consistent and accurate normalization in both professional and consumer languages.

Traditional UMLS-based tools such as MetaMap (Aronson, 2001), MetaMap Lite (Demner-Fushman et al., 2017), cTAKES (Savova et al., 2010), and QuickUMLS (Soldaini and Goharian, 2016) rely primarily on lexical matching and dictionary lookup. Although effective for structured clinical text, their dependence on surface similarity reduces the accuracy when expressions deviate from standard terminology, motivating the need for more semantically robust approaches (Limsopatham and Collier, 2016).

Neural embedding methods can address this limitation by mapping biomedical terms into a continuous semantic vector space, where the directional similarity between embeddings reflects conceptual similarity. This allows normalization to be performed through semantic similarity matching rather than exact string comparison or lexicon lookup (Sung et al., 2020). While early static embeddings such as Word2Vec (Mikolov et al., 2013) captured broad semantic relationships, their fixed word representations limited their usefulness in biomedical settings, where meaning depends strongly on context. Contextual encoders such as BioBERT (Lee et al., 2020) improve biomedical understanding but do not explicitly align synonymous expressions. SapBERT (Liu et al., 2021) bridges this gap by training directly on UMLS synonym pairs using a metric-learning objective, producing concept-aware embeddings.

Since its release, SapBERT has been adopted across multiple biomedical NLP tasks, including clinical concept normalization, multilingual terminology alignment, and biomedical entity linking. Prior work demonstrates its effectiveness in mapping clinical notes to standardized vocabularies (Abdulnazar et al., 2023), aligning cross-lingual concepts (Lin et al., 2022), and achieving state-of-the-art performance in entity linking when combined with re-ranking methods (Gnecco et al., 2025).

Pretrained large language models (LLMs) such as OpenAI's GPT, Google's Gemini, and Meta's Llama models have emerged as powerful tools for biomedical text processing due to their strong capabilities in language understanding (Havlík, 2024). In the context of concept normalization, recent studies have examined their use for generating standardized terminology from informal biomedical text (Berkowitz et al., 2025; Dobbins, 2024).

We propose a unified normalization pipeline to standardize diverse health expressions by improving biomedical concept mapping. Our approach combines embedding-based CUI retrieval from the UMLS Metathesaurus and LLM-driven generation of preferred terms. The system is tested on both clinical dataset and social media content: clinical phrases are directly mapped via semantic retrieval, while informal expressions are first converted to preferred terms before CUI assignment.

2 Method

To improve the accuracy of retrieving CUIs from given phrases, a two-stage pipeline was proposed to perform health concept normalization. Figure 1 shows the workflow. First, UMLS terms and CUIs are encoded using the Self-Alignment Pretraining for Biomedical Entity Representations (SapBERT)

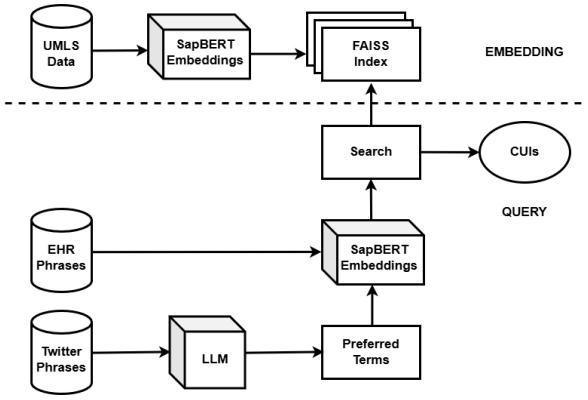

Figure 1: Pipeline for data processing and analysis.

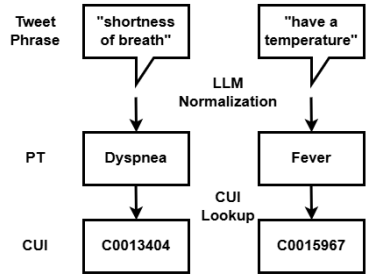

Figure 2: Examples of CUI retrieval from Twitter phrases by obtaining their preferred terms first and then finding the corresponding CUIs.

embedding model (Abdulnazar et al., 2023) and indexed with the Facebook AI Similarity Search (FAISS) (Facebook Engineering, 2017) to construct the concept search space. Each set of UMLS term and CUI is represented in a vector of real numbers. During inference, clinical phrases are retrieved directly via embedding search, while informal Twitter phrases are first normalized into preferred terms by an LLM before CUI retrieval through embedding search, as illustrated in Figure 2.

2.1 Data

Three sources of data were used in this work: (1) the UMLS Metathesaurus (U.S. National Library of Medicine, 2025b), which provides the controlled vocabulary and concept space for normalization, (2) a group of 137 annotated COVID-19 signs and symptoms and their associated synonyms, made up of a total of 3,144 expressions from EHRs of 3 different hospitals (Wang et al., 2021), and (3) a set of 102 symptom expressions from COVID-19 related tweets (Jiang and Bernard, 2024). The first dataset was for creation of neural embeddings of UMLS terms and CUIs, and the other two datasets were for validation.

2.1.1 UMLS Metathesaurus (2025AA Release)

The 2025AA release of the Unified Medical Language System (UMLS) Metathesaurus (U.S. National Library of Medicine, 2025a) was obtained under the required license agreement and used as the reference terminology resource for CUI retrieval. The Metathesaurus contains 17,144,356 records in the Concept Names and Sources file `MRCONSO.RRF`, the main concept names and synonyms table, of which 10,492,355 are English-language entries. All English terms were extracted and paired with semantic type information from the Semantic Types file (`MRSTY.RRF`), which stores the semantic category assignments for each concept. Four fields were retained for downstream processing: CUI (Concept Unique Identifier), STR (surface string), TTY (term type), and STY (semantic type).

After removing duplicates to ensure each (CUI, STR, TTY, STY) mapped uniquely, preprocessing resulted in 3,594,105 unique English UMLS terms. These entries formed both the concept search space for FAISS-based nearest neighbor retrieval and the embedding index used across all experiments.

Within the UMLS Metathesaurus, each concept is associated with one or more preferred terms, designated by the term-type code PT, which serve as representative strings for the corresponding CUI. These preferred terms are computed from a list of ranked source vocabularies (U.S. National Library of Medicine, 2016a). Mapping informal phrases to preferred terms increases the likelihood of retrieving the correct CUI, especially when input text is noisy, abbreviated, or nonstandard.

2.1.2 COVID-19 Signs and Symptoms from EHR Data

To compare baseline methods with semantic retrieval using SapBERT embeddings, we adopted a corpus of 153 COVID-19 signs and symptoms and their associated CUIs[1] (Wang et al., 2021). This dataset was preprocessed to only include terms whose CUIs were present in the UMLS, resulting in 137 rows of signs and symptoms. Table 1 presents examples from the dataset, each listing a sign or symptom, example synonyms, and corresponding UMLS CUI.

We recompiled the dataset into a list of (CUI, phrase) pairs, where each pair represents either a preferred term or a synonym associated with

[1] https://github.com/Medical-NLP/COVID-19-Sign-Symptom

Table 1: Examples of COVID-19 signs and symptoms, synonyms, and UMLS CUIs.

Sign & Symptom	Synonyms	CUI
Anosmia	Sense of smell absent, loss of sense of smell	C0003126
Dyspnea	breathlessness, Shortness of Breath	C0013404
Fever	Increased body temperature, hyperthermia	C0015967
Sore Throat	Throat pain, throat discomfort	C0242429

the concept. After removing duplicates, the final dataset contained 3,144 distinct (CUI, phrase) pairs. These phrases served as the input for evaluating lexical matching against UMLS surface strings and for semantic retrieval using SapBERT embeddings.

2.1.3 COVID-19 Symptom Expressions from Tweets

To validate our approach for normalization under informal language conditions, a corpus of COVID-19 symptom phrases found in Twitter data was obtained from (Jiang and Bernard, 2024). The original dataset contains 3 major symptoms: breathing difficulty (48), fever (54), and loss of taste and/or smell (54). The *uniqueness* of this dataset is that there are dozens of mentions for each symptoms, representing a variety of ways to express each symptom concept. Given that each UMLS term only maps to a single CUI, finding 2 CUIs from a phrase such as "loss of taste and/or smell" is challenging. Therefore, we intentionally excluded all phrases of "loss of taste and/or smell", yielding a set of 102 phrases[2].

These expressions exhibit substantial lexical variability, including metaphoric phrasing (e.g., *"breathing has become a full-time job"*), misspellings (e.g., *"breathleness"*), and short text (e.g., *"high temp"*). To reduce such variations, an LLM was asked to generate the corresponding preferred term for each symptom phrase. Afterwards, a CUI was retrieved for the generated preferred term through its neural embedding.

2.2 Embedding with SapBERT

The embeddings of health concepts were generated using SapBERT[3], a BERT-based encoder initialized from PubMedBERT and fine-tuned on millions of UMLS synonym pairs using a metric-learning objective (Liu et al., 2021). The model produces 768-dimensional embeddings and is designed to cluster synonymous terms while separating expressions associated with different CUIs.

In our pipeline, each UMLS term was tokenized to a maximum length of 32 tokens and encoded in the evaluation mode. We extracted the `[CLS]` token representation and applied ℓ_2 normalization, so that FAISS inner-product search corresponds to cosine similarity (more discussions below). It rescales embeddings so that their direction (semantic meaning) matters more than their magnitude, allowing concepts such as *"struggling to breathe"* to consistently retrieve the correct dyspnea-related cluster during nearest-neighbor search.

For a total of 3,594,105 UMLS terms and 768 dimensions of each embedding vector, the resulting embedding matrix,

$$\mathbf{E} \in \mathbb{R}^{3,594,105 \times 768},$$

contains one vector per UMLS surface string (STR).

When retrieving corresponding CUIs, clinical phrases were embedded directly using SapBERT because they already contained medical terminology, but for Twitter phrases, they were first normalized to preferred terms before embedding. Since SapBERT was trained on standard biomedical phrases, this step can help improve alignment between informal language and the UMLS concept space.

2.3 FAISS Index Construction

Facebook AI Similarity Search (FAISS) (Facebook Engineering, 2017), a high-performance open-source library for large-scale vector similarity search, was adopted as (1) the storage for our embedding vectors, and (2) the retrieval engine based on the nearest-neighbor search. SapBERT embeddings of 3.59 million English UMLS terms were indexed using an exact inner-product index (`IndexFlatIP`). Because all embeddings were ℓ_2-normalized, maximizing the inner product is equivalent to maximizing cosine similarity. Retrieval is defined as:

$$\hat{c} = \arg \max_{j \in \{1,\ldots,N\}} \hat{h}_q^\top \hat{h}_j,$$

where $\hat{h}_q$ denotes the query embedding, $\hat{h}_j$ the embedding of the j-th UMLS term, N the total number of indexed concepts, and $\hat{c}$ the retrieved concept with the highest cosine similarity.

[2] https://github.com/medeffects/covid_symptoms_corpus

[3] https://huggingface.co/cambridgeltl/SapBERT-from-PubMedBERT-fulltext

2.4 Large Language Models

To evaluate the effect of LLM-generated preferred terms from layman's phrases, seven state-of-the-art pretrained large language models from multiple development families of OpenAI, Google, and Meta AI were investigated. Some of these models are closed-weight models and the rest are open-weight models.

2.4.1 Closed-Weight LLMs

Three OpenAI models were evaluated: GPT-4o, GPT-4o mini, and GPT-5. GPT-4o is a multimodal transformer supporting text, image, audio, and video input and demonstrates strong reasoning performance across modalities (OpenAI, 2024). GPT-4o mini provides a cost-efficient alternative with similar instruction-following behavior, while GPT-5 represents OpenAI's most recent (at this writing) high-capacity model trained on a broader range of data (OpenAI, 2025).

Two models from the Google Gemini 2.x family were also tested: Gemini 2.0 Flash and Gemini 2.5 Flash. Both were built on a mixture-of-experts (MoE) transformer architecture, optimized for high-throughput inference. Gemini 2.5 Flash incorporates updated training data and has improved multimodal capabilities (Google DeepMind, 2025).

All closed-weight models were accessed through their respective cloud APIs and invoked directly within our Python pipeline.

2.4.2 Open-Weight Models

Two open-weight models from Meta AI, Llama 3.1-70B and Llama 3.3-70B, were also evaluated. They are autoregressive transformer models trained on large-scale text and code corpora (Meta AI, 2024). The 70B-parameter configuration offers a strong balance between capacity and deployability. The newer Llama 3.3 release includes improvements in training quality and instruction tuning.

Since Meta AI does not offer Llama models as a cloud API service, both variants were locally deployed on an NVIDIA GeForce RTX 4090 GPU machine using the `Ollama` (Ollama, 2025) runtime.

2.5 Prompt Design

Well-engineered prompting is needed to instruct LLMs to generate quality preferred terms. To normalize informal symptom expressions in tweets into medically valid terminology, we designed a concise instruction prompt for more deterministic normalization. Because social media text is often informal or ambiguous, the prompt enforces (1) standard clinical vocabulary, (2) a single-term output, and (3) suppression of extraneous text.

> **Prompt:** Normalize the following medical symptom phrase into standard medical terminology. Provide the single most appropriate preferred term. Respond only with the preferred term.
>
> Phrase: "<symptom expression>"
>
> Preferred term:

Figure 3: Prompt for finding preferred terms.

The prompt consisted of three components: an instruction framing the input as a medical symptom, a constraint requiring a single normalized term as output, and the raw input phrase. The complete format is shown in Figure 3. This design ensures a targeted response rather than open-ended generation.

Under this prompting formulation, input expressions such as *"fought for breath"* and *"temp drifted up into the 99s"* were normalized to appropriate medical concepts (e.g., *"dyspnea"* and *"fever"*). Because UMLS Metathesaurus primarily reflects professional biomedical terminology, normalization toward preferred terms (PT) forms can lead to more accurate retrieval of CUIs.

Although state-of-the-art closed-weight LLMs can be prompted to generate UMLS CUIs directly, our prior work showed that such mappings are often unreliable and hallucinative, with the same concept mapped to incorrect, unrelated, or different identifiers, which may be attributed to the requirement of the UMLS end user license (Jiang and Bernard, 2024). To ensure verifiability against the UMLS Metathesaurus, we therefore restrict the role of the LLM to linguistic normalization only, with CUI assignment handled separately afterwards. This design choice also avoids practical challenges associated with directly grounding LLMs in licensed biomedical ontologies such as UMLS.

To ensure reproducibility in the normalization step and to account for the non-deterministic nature of LLMs—where identical prompts may yield varying outputs—all models were configured with a *temperature* of 0. This configuration ensures consistent model outputs by forcing the selection of the most likely completion.

2.6 Evaluation Framework

2.6.1 Metrics

We evaluate normalization performance using accuracy, which is widely adopted for this task (French and McInnes, 2023; Tutubalina et al., 2018), and F1-score at the phrase level. Each input phrase is associated with exactly one gold-standard UMLS CUI, and each method produces at most one predicted CUI per phrase. A prediction is counted as a true positive (TP) if it exactly matches the gold-standard CUI.

Because the evaluation set consists exclusively of annotated biomedical concepts, true negative (TN) cases are not defined. Both incorrect predictions and instances where the model produces no output are therefore treated as false negatives (FN), yielding a single-label recovery setting. Under this formulation, *accuracy* is defined as:

$$\text{Accuracy} = \frac{TP}{TP + FN}.$$

Since each phrase has at most one prediction and no true negatives exist, *precision* is trivially equal to 1 whenever a prediction is produced and is therefore not informative. As a result, the F1-score simplifies to:

$$\text{F1} = \frac{2 \cdot TP}{2 \cdot TP + FN}.$$

For embedding-based CUI retrieval, including LLM-assisted normalization, the system always returns a single nearest neighbor from the UMLS embedding index, eliminating no-output cases. For preferred-term evaluation, a generated term is counted as an exact match if it exactly matches any UMLS preferred term associated with the gold CUI; for semantic matching, both generated and gold preferred terms are embedded using SapBERT, and the maximum cosine similarity across gold preferred terms is used to assess alignment. Similarity thresholds (e.g., >50% and >80%) are applied only as diagnostic criteria to assess whether this maximum similarity exceeds a given level.

2.6.2 Baselines

Two baselines are used for comparison. Exact string matching maps an input phrase to a UMLS concept only if the entire phrase exactly matches a UMLS surface form, reflecting the default lookup behavior of the UMLS Metathesaurus and providing a transparent and conservative point of comparison.

Table 2: Performance of CUI retrieval with phrases from clinical data.

Method	Accuracy	F1
Exact Match	0.679	0.809
MetaMap Lite	0.579	0.733
Neural Embedding	**0.858**	**0.924**

Table 3: Performance of CUI retrieval with LLMs from Twitter phrases.

Method	Accuracy	F1
Exact Match	0.235	0.381
MetaMap Lite	0.118	0.211
GPT-4o	0.961	0.980
GPT-4o mini	**0.980**	**0.990**
GPT-5	0.941	0.970
Gemini 2.0 Flash	0.961	0.980
Gemini 2.5 Flash	0.882	0.937
Llama 3.1-70B	0.921	0.959
Llama 3.3-70B	0.941	0.970

MetaMap Lite software (Demner-Fushman et al., 2017), developed by the U.S. National Libray of Medicine, is included as a UMLS-based normalization baseline. Since MetaMap Lite performs mention detection and may return multiple candidate CUIs for a phrase, only the highest-scoring returned CUI (top-1) for each input phrase was selected.

3 Results

Tables 2 and 3 show the results of retrieving CUIs from phrases found in clinical data and Twitter data, respectively. The CUI retrieval from the clinical data was performed *only* through neural embedding (Table 2), and that from Twitter phrases was conducted by first obtaining preferred terms from each LLM and then via neural embedding (Table 3).

Quantitative results for LLM-based normalization are reported in Table 4. The Exact Match columns represent string-level comparisons between LLM-generated preferred terms and UMLS counterparts (PTs), while the >50% and >80% columns report semantic similarity between their SapBERT embeddings.

4 Discussions

Table 2 highlights a clear performance gap between the baselines (0.679 for exact string matching and 0.579 for MetaMap Lite) and neural embedding-based retrieval (0.858), indicating that surface-form matching and traditional MetaMap algorithm are insufficient for robust biomedical concept normalization. Approximately one-third of clinical ex-

Table 4: Accuracy and F1-score of mapping LLM-generated preferred terms to UMLS preferred terms. Exact Match reports results without neural embeddings, while >50% and >80% denote cosine similarity thresholds. Best results per column are shown in bold.

Model	Exact Match		>50% Similarity		>80% Similarity	
	Acc	F1	Acc	F1	Acc	F1
GPT-4o	0.961	0.980	1.000	1.000	0.971	0.985
GPT-4o mini	**0.980**	**0.990**	1.000	1.000	**0.990**	**0.995**
GPT-5	0.941	0.970	1.000	1.000	0.961	0.980
Gemini 2.0 Flash	0.961	0.980	1.000	1.000	0.961	0.980
Gemini 2.5 Flash	0.873	0.932	1.000	1.000	0.902	0.949
Llama 3.1-70B	0.853	0.921	1.000	1.000	0.941	0.970
Llama 3.3-70B	0.922	0.960	1.000	1.000	0.971	0.985

pressions lacked direct matches to UMLS surface forms, limiting the effectiveness of exact string matching, while MetaMap Lite returned valid CUIs for most phrases but failed to recover the correct gold-standard CUI for over two-fifths of cases.

MetaMap Lite leverages UMLS knowledge and linguistic analysis, but *surprisingly* underperforms exact string matching in this setting because it focuses on span-level concept identification rather than strict concept specificity. As a result, it often assigns valid but more general or semantically adjacent CUIs based on sub-phrase matches. For instance, the phrase *"ache in joint"* was mapped to the general concept *"ache"* (C0234238) instead of the gold-standard symptom concept *"arthralgia"* (C0003862).

In contrast, semantic retrieval using SapBERT embeddings effectively identifies concept variants expressed with differing wording. This improvement stems from SapBERT's synonym-aware training objective, which clusters semantically equivalent expressions even when phrasing deviates noticeably from canonical terminology.

While semantic retrieval improves concept matching within clinical phrasing, informal social media language introduces challenges that embeddings alone fail to resolve. Table 3 illustrates significant improvements in accuracy in finding CUIs from Twitter phrases through LLM-generated preferred terms, from 0.235 for exact string matching to a range of 0.882 to 0.980, corresponding to improvements of 275% to 317%; relative to MetaMap Lite (0.118), the same accuracy range represents larger gains of 647% to 732%.

Tweets frequently contain slang (*"temperature is still crazy at 39.4"*), metaphor (*"breathing has become a full-time job"*), short texts (*"temp still up"*), and misspellings (*"breathleness"*) that do not resemble clinical vocabulary closely enough for direct retrieval.

Incorporating LLM-based preferred-term generation provides a powerful tool for bridging this gap. By translating informal phrases into medically grounded terminology, LLMs function as a linguistic normalization layer that enables more accurate embedding-based CUI retrieval. This division of labor supports a two-stage architecture in which LLMs perform language standardization and embeddings provide biomedical concept alignment.

A survey by French and McInnes reports that accuracies of 23 systems performing biomedical concept normalization on various clinical datasets range from the low of 62% to the high of 96.9% (French and McInnes, 2023). Compared to previously published results, our accuracies of matching Twitter phrases to CUIs meet or exceed the upper level of reported accuracies on clinical datasets, indicating the effectiveness of our approach under the evaluated conditions.

To better understand where these gains originate within the two-stage pipeline, we briefly examined the quality of the intermediate preferred terms generated by the LLMs before CUI retrieval (please see Section 2.6.1 for details). As shown in Table 4, while these terms may not always exactly match UMLS preferred terms (as reflected in the Exact Match column), relaxing the similarity thresholds (to >50% or >80%) yields improved alignment with UMLS PTs.

Analysis of LLM-generated outputs indicate that these performance gains stem primarily from consistent mapping of informal expressions to medically grounded terminology. Across models, common symptom phrases were frequently normalized to standard clinical concepts (e.g., *"dyspnea"* for breathing-related expressions and *"fever"* for temperature-related mentions), enabling reliable downstream CUI retrieval.

Variability in preferred-term generation was observed mainly in cases involving differences in clin-

Table 5: Representative normalization patterns observed across multiple LLMs. For each input phrase, a single example preferred term is shown; cosine similarity values summarize semantic proximity to the gold concept, with ranges reported only when multiple models produced different scores.

Input Phrase	Representative PT (example)	Cosine Similarity (<80%)	Gold CUI	Observed Pattern
"affecting my breathing"	respiratory distress	0.735	C0013404	Granularity mismatch (related concept)
"breathing issues"	respiratory symptom	0.555	C0013404	Overly general concept selection
"temp that won't come down"	refractory fever	0.612	C0015967	Persistent-based severity inference
"temp between 37.3–37.8"	subfebrile	0.674–0.783	C0015967	Boundary ambiguity within fever spectrum

ical granularity or boundary conditions, such as selecting overly general descriptors (e.g., *"respiratory symptom"*) or adjacent severity categories (e.g., *"subfebrile"* versus *"fever"*). As illustrated in Table 5, these cases reflect semantic ambiguity rather than hallucinated or unrelated outputs, suggesting that remaining mismatches arise from fine-grained clinical distinctions rather than failures to interpret the input phrase.

GPT models demonstrated superior performance overall, with GPT-4o-mini notably achieving strong results comparable to larger models, suggesting that conceptual understanding and domain exposure are more critical than size. Open-weight models, such as Llama 3.1 and 3.3, performed competitively but inconsistently, likely due to variations in biomedical training data and instruction tuning. Gemini models showed robust performance collectively, though Gemini 2.5 Flash exhibited lower accuracy, possibly influenced by prompt design or domain familiarity.

The main contribution of this work is a unified approach that integrates generative and embedding-based methods into a cohesive normalization process. The proposed framework offers: (1) generalizability across clinical and consumer text, (2) interpretability through intermediate preferred-term generation, and (3) scalability via FAISS indexing over millions of UMLS entries. The improved accuracy and robustness demonstrated in this study suggest that this architecture can strengthen downstream applications such as syndromic surveillance, population-level analysis, and clinical NLP systems for concept extraction.

5 Limitations

Although the UMLS Metathesaurus covers a wide range of biomedical semantic types, this study focused only on sign and symptom concepts from two relatively small datasets, limiting empirical validation of generalizability to other domains such as diseases, medications, or procedures. This reflects the limited availability of publicly accessible annotated datasets, especially for informal social media text. In addition, the analyzed corpora consist of pre-identified phrases rather than full running text and considers only single-concept expressions, whereas real-world clinical notes and social media posts may contain multiple or overlapping concepts. Finally, the datasets include only phrases with gold-standard annotations, and true negative cases are therefore not defined; extending the framework to support abstention when no clinical concept is present remains a direction for future work.

6 Conclusion

We propose a neural embedding-based approach for normalizing health concepts from clinical and social media data using SapBERT, trained on UMLS data. To enhance accuracy with informal social media (Twitter in this case) phrases, we first use an LLM to generate preferred terms, which are then input into our embedding system for CUI retrieval. Our results demonstrate significant improvements over traditional string matching and UMLS-based rule-driven normalization methods such as MetaMap Lite, particularly for diverse clinical expressions and noisy social media text. This method effectively maps user-generated language to standardized biomedical concepts, showing that integrating generative normalization with semantic retrieval offers a robust solution for biomedical concept mapping.

Ethical Considerations

This study analyzes health-related text derived from de-identified clinical datasets and publicly available social media posts, without using any personal

or identifiable information. All use of the UMLS Metathesaurus complies with the U.S. National Library of Medicine license agreement, and large language models were restricted to linguistic normalization only, with all CUI assignments performed through deterministic embedding-based retrieval to improve the output accuracy while remaining compliant with licensing requirements. The proposed framework is intended for research and decision-support purposes rather than clinical diagnosis, and its performance may be affected by biases present in social media data and pretrained language models. Any practical use of this approach requires the involvement of domain experts and appropriate validation.

Acknowledgment

The authors wish to thank three reviewers for their thoughtful critiques and constructive feedback, which helped improve this manuscript.

References

Akhila Abdulnazar, Markus Kreuzthaler, Roland Roller, and Stefan Schulz. 2023. Sapbert-based medical concept normalization using snomed ct. In *Caring Is Sharing–Exploiting the Value in Data for Health and Innovation*, pages 825–826. IOS Press.

Liz Amos, David Anderson, Stacy Brody, Anna Ripple, and Betsy L Humphreys. 2020. Umls users and uses: a current overview. *Journal of the American Medical Informatics Association*, 27(10):1606–1611.

Adib Ahmed Anik, Paramita Basak Upama, Masud Rabbani, Shiyu Tian, Min Sook Park, Sheikh Iqbal Ahamed, Jake Luo, and Hyunkyoung Oh. 2024. Identifying medical concepts and semantic types in lay vocabularies of health consumers who are concerned with diabetes on social media using the umls and nlp. In *2024 IEEE 48th Annual Computers, Software, and Applications Conference (COMPSAC)*, pages 862–869. IEEE.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17.

Jacob S Berkowitz, Apoorva Srinivasan, Jose Miguel Acitores Cortina, Yasaman Fatapour, and Nicholas P Tatonetti. 2025. Biomedical text normalization through generative modeling. *Journal of Biomedical Informatics*, page 104850.

Haihua Chen, Yuhan Zhou, Ruochi Li, Aryan Murthy Illa, Ana Cleveland, and Junhua Ding. 2025. A comprehensive survey on medical concept normalization: Datasets, techniques, applications, and future directions. *Techniques, Applications, and Future Directions (November 10, 2025)*.

Dina Demner-Fushman, Willie J Rogers, and Alan R Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.

Nicholas J Dobbins. 2024. Generalizable and scalable multistage biomedical concept normalization leveraging large language models. *Research Synthesis Methods*, pages 1–12.

Facebook Engineering. 2017. Faiss: A library for efficient similarity search. https://engineering.fb.com/2017/03/29/data-infrastructure/faiss-a-library-for-efficient-similarity-search/. Accessed: 2025-11-18.

Evan French and Bridget T McInnes. 2023. An overview of biomedical entity linking throughout the years. *Journal of biomedical informatics*, 137:104252.

Álvaro García-Barragán, Ahmad Sakor, Maria-Esther Vidal, Ernestina Menasalvas, Juan Cristobal Sanchez Gonzalez, Mariano Provencio, and Víctor Robles. 2025. Nssc: a neuro-symbolic ai system for enhancing accuracy of named entity recognition and linking from oncologic clinical notes. *Medical & Biological Engineering & Computing*, 63(3):749–772.

Daniel Peña Gnecco, Jairo Serrano, Edwin Puertas, and Juan Carlos Martinez-Santos. 2025. Hybrid reranking for biomedical entity linking using sapbert embeddings: a high-performance system for bionne-l 2025-1. In *CLEF*.

Google DeepMind. 2025. Gemini 2.5 Flash: High-speed multimodal model. https://deepmind.google/models/gemini/flash/. Accessed: 2025-11-18.

Vladimír Havlík. 2024. Meaning and understanding in large language models. *Synthese*, 205(1):9.

Zhe He, Zhiwei Chen, Sanghee Oh, Jinghui Hou, and Jiang Bian. 2017. Enriching consumer health vocabulary through mining a social q&a site: A similarity-based approach. *Journal of biomedical informatics*, 69:75–85.

Keyuan Jiang and Gordon R Bernard. 2024. The ability of pretrained large language models in understanding health concepts in social media posts. In *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 5494–5498. IEEE.

Sarvnaz Karimi, Chen Wang, Alejandro Metke-Jimenez, Raj Gaire, and Cecile Paris. 2015. Text and data mining techniques in adverse drug reaction detection. *ACM Computing Surveys (CSUR)*, 47(4):1–39.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Nut Limsopatham and Nigel Collier. 2016. Normalising medical concepts in social media texts by learning semantic representation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (volume 1: long papers)*, pages 1014–1023.

Ying-Chi Lin, Phillip Hoffmann, and Erhard Rahm. 2022. Enhancing cross-lingual biomedical concept normalization using deep neural network pretrained language models: Yc. lin et al. *SN Computer Science*, 3(5):387.

Fangyu Liu, Ehsan Shareghi, Zaiqiao Meng, Marco Basaldella, and Nigel Collier. 2021. Self-alignment pretraining for biomedical entity representations. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 4228–4238.

Yen-Fu Luo, Sam Henry, Yanshan Wang, Feichen Shen, Ozlem Uzuner, and Anna Rumshisky. 2020. The 2019 n2c2/umass lowell shared task on clinical concept normalization. *Journal of the American Medical Informatics Association*, 27(10):1529–e1.

Emmanouil Manousogiannis, Sepideh Mesbah, Alessandro Bozzon, Robert-Jan Sips, Zoltan Szlanik, and Selene Báez Santamaría. 2020. Normalization of long-tail adverse drug reactions in social media. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 49–58.

Meta AI. 2024. Llama 3 model family. https://www.llama.com/models/llama-3/. Accessed: 2025-11-18.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Denis Newman-Griffis, Guy Divita, Bart Desmet, Ayah Zirikly, Carolyn P Rosé, and Eric Fosler-Lussier. 2021. Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. *Journal of the American Medical Informatics Association*, 28(3):516–532.

Ollama. 2025. Ollama: Run open models locally. https://ollama.com/. Accessed: 2025-11-18.

OpenAI. 2024. GPT-4o system card. https://openai.com/index/gpt-4o-system-card/.

OpenAI. 2025. GPT-5 system card. https://openai.com/index/gpt-5-system-card/.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Sanja Scepanovic, Enrique Martin-Lopez, Daniele Quercia, and Khan Baykaner. 2020. Extracting medical entities from social media. In *Proceedings of the ACM conference on health, inference, and learning*, pages 170–181.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*, pages 1–4.

Mujeen Sung, Hwisang Jeon, Jinhyuk Lee, and Jaewoo Kang. 2020. Biomedical entity representations with synonym marginalization. *arXiv preprint arXiv:2005.00239*.

Lynda Tamine and Lorraine Goeuriot. 2021. Semantic information retrieval on medical texts: Research challenges, survey, and open issues. *ACM Computing Surveys (CSUR)*, 54(7):1–38.

Elena Tutubalina, Zulfat Miftahutdinov, Sergey Nikolenko, and Valentin Malykh. 2018. Medical concept normalization in social media posts with recurrent neural networks. *Journal of biomedical informatics*, 84:93–102.

U.S. National Library of Medicine. 2016a. Preferred terms. https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_004.html. Last reviewed July 29, 2016. Accessed: 2026-01-26.

U.S. National Library of Medicine. 2016b. Unique identifiers in the metathesaurus. https://www.nlm.nih.gov/research/umls/new_users/online_learning/Meta_005.html. Last reviewed July 29, 2016. Accessed: 2025-11-18.

U.S. National Library of Medicine. 2025a. Umls knowledge sources. https://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html. Release 2025AA, National Library of Medicine (US), Bethesda, MD. Accessed: 2025-09-08.

U.S. National Library of Medicine. 2025b. Unified medical language system (umls). https://www.nlm.nih.gov/research/umls/index.html. Accessed: 2025-11-18.

Jingqi Wang, Noor Abu-el Rub, Josh Gray, Huy Anh Pham, Yujia Zhou, Frank J Manion, Mei Liu, Xing Song, Hua Xu, Masoud Rouhizadeh, and Yaoyun Zhang. 2021. Covid-19 signsym: a fast adaptation of a general clinical nlp tool to identify and normalize covid-19 signs and symptoms to omop common data model. *Journal of the American Medical Informatics Association*, 28(6):1275–1283.

Jiageng Wu, Lumin Wang, Yining Hua, Minghui Li, Li Zhou, David W Bates, and Jie Yang. 2023. Trend and co-occurrence network of covid-19 symptoms from large-scale social media data: infoveillance study. *Journal of Medical Internet Research*, 25:e45419.

Qing T Zeng and Tony Tse. 2006. Exploring and developing consumer health vocabularies. *Journal of the American Medical Informatics Association*, 13(1):24–29.