Generalizable Approaches for Medical Term Normalization

Jacob Berkowitz, Yasaman Fatapour, Nicholas P. Tatonetti
Department of Computational Biomedicine, Cedars-Sinai Medical Center

Background

Approximately 80% of electronic health record (EHR) data consists of unstructured text, complicating the extraction of potentially life-saving medical insights (1). The complexity of medical language within EHRs present challenges for downstream analysis. Large language models (LLMs) offer promising solutions to these challenges by normalizing unstructured text to standardized medical terminologies. Here, we develop and evaluate generalizable approaches for medical text normalization using OpenAI's GPT-4. We selected GPT-4 for its wide availability and ease of use, as the computation is handled remotely, not requiring extensive local resources.

Methods

We developed four text normalization frameworks: Zero-Shot Recall, Prompt Recall, Semantic Search, and Retrieval-Augmented Generation (RAG) (Figure 1). Zero-Shot Recall (2) involves prompting GPT-4 to retrieve the normalized term from its training data and is computationally and cost efficient. Prompt Recall, similar to the needle in a haystack problem (3), passes a full dictionary of terms to GPT-4, ensuring the model has the correct match in its knowledge base. Semantic search (4) offers the most cost-effective solution, involving searching a semantic embedding space created by using an embedding model, in this case GIST-large-Embedding-v0. RAG (5) combines the strengths of Prompt Recall and Semantic Search, passing a dictionary of related candidate terms to GPT-4.

We evaluate these frameworks on their ability to map medical term synonyms to Systematized Nomenclature of Medicine—Clinical Terms (SNOMED CT) IDs (6) using two datasets: one oncology-specific and one covering a broad range of medical conditions. We generate these datasets using GPT-4 to produce ten synonyms for each term. For the oncology-specific dataset, we extracted terms related to "Malignant neoplastic disease" from the OMOP database, linked to ICD-10 billing codes. For the broader dataset, we randomly selected terms from institutional billing codes, ensuring a diverse representation of medical conditions.

To assess the performance of each framework, we used the proportion of correct responses, or accuracy. Additionally, to estimate cost and time complexity, we recorded tokens per call and number of total API calls.

Results

Zero-shot Recall returned the correct term 4.6% and 0.9% in the 106 and 750 term datasets, respectively. Prompt recall had accuracy of 86.4% and 40.8%. Semantic search achieved 86.0% and 76.2% accuracy. RAG had the highest observed accuracy at 89.8% and 80.0% of terms (Table 1).

Conclusions

While all approaches have their merit and may be optimal in specific use-cases, the RAG approach demonstrates the most promise in text normalization to SNOMED CT. Zero-Shot Recall’s poor performance may be attributed to lack of specific knowledge, however it correctly identified commonly used SNOMED CT such as “*primary malignant neoplasm of female breast*” and “*primary malignant neoplasm of prostate*.” Despite Prompt Recall ensuring the LLM has access to the correct term, the increase in irrelevant terminology overwhelms the model and reduces performance. Narrowing the candidate list down through semantic search saves on time and cost, while demonstrating greater performance. This study highlights the potential of LLMs in improving the accuracy and efficiency of EHR data management, which could lead to enhanced patient care and outcomes. Further research is needed to refine these techniques and their implementation in healthcare environments.

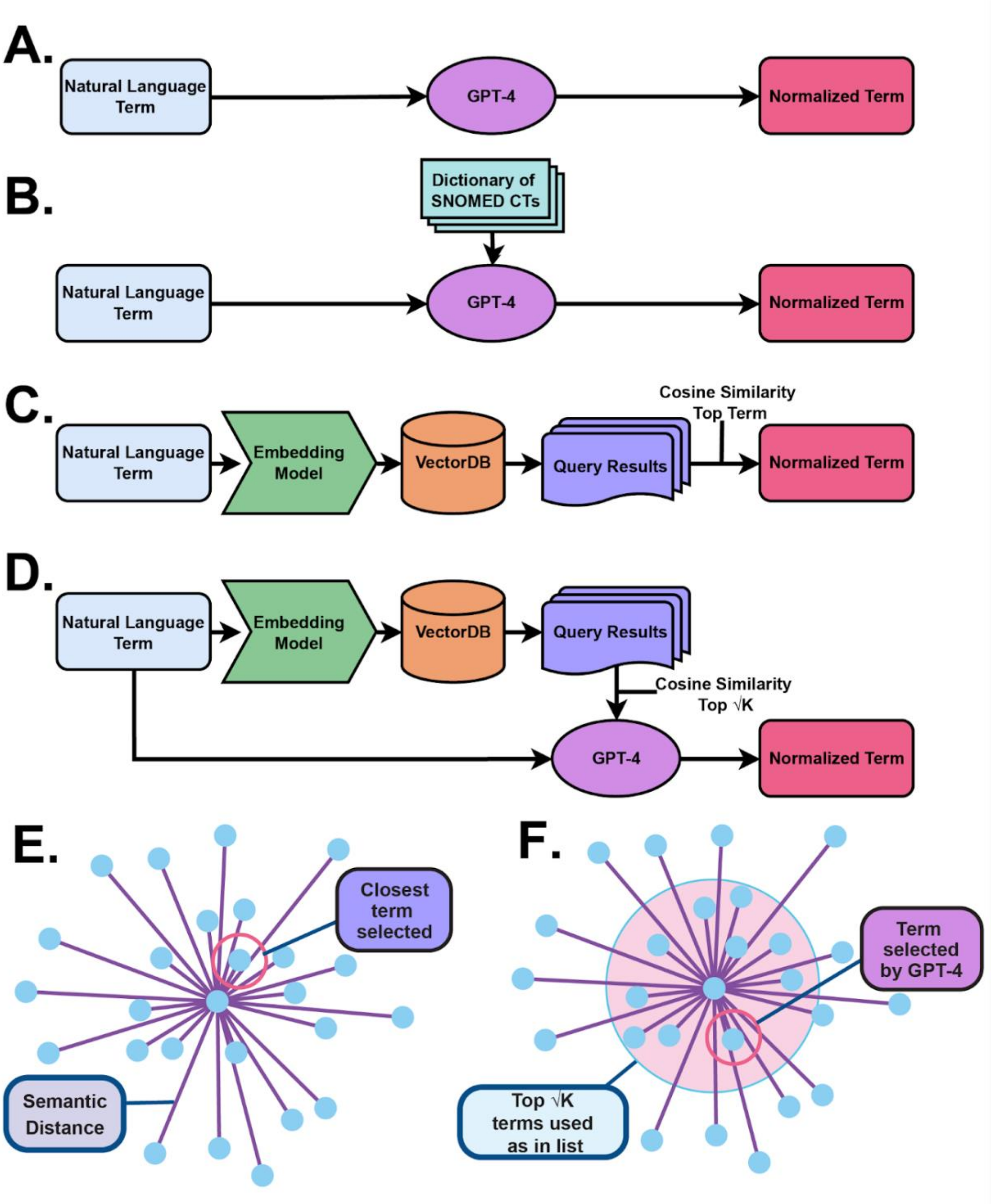

Figure 1: Methodology Flowchart Step-by-step approach for four different normalization methods— (A) Zero-Shot Recall: Utilizes a single prompt to elicit the correct term from the model without any prior examples or fine-tuning. (B) Prompt Recall: Feeds the model a comprehensive list of terms, prompting it to select the most appropriate one based on the input context. (C) Semantic Search: Matches input terms with the closest semantic equivalents from a precomputed vector space of embeddings. (D) RAG: First retrieves relevant documents or data snippets and then generates the normalized term by

synthesizing the retrieved information. An embedding space visualization illustrates the differences between Semantic Search (E) and RAG (F).

Table 1: Model Performance and Utility

	Malignant Neoplastic Disease Terms (N=106)			**Randomly sampled diagnosis codes (N=750)**		
Approach	**# API Calls**	**Average Tokens per call**	**Accuracy**	**# API Calls**	**Average tokens per call**	**Accuracy**
Zero-Shot Recall	106	36	4.6%	750	36	0.9%
Prompt Recall	106	1606	86.4%	750	11,529	40.8%
Semantic Search	0		86.0%	0		76.2%
RAG	106	189	**89.8%**	750	455	**80.0%**

References

1. Leaman R, Khare R, Lu Z. Challenges in clinical natural language processing for automated disorder normalization. J Biomed Inform. 2015 Oct 1;57:28–37.
2. Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. 2022 May 24; Available from: http://arxiv.org/abs/2205.11916
3. Kuratov Y, Bulatov A, Anokhin P, Sorokin D, Sorokin A, Burtsev M. In Search of Needles in a 11M Haystack: Recurrent Memory Finds What LLMs Miss. 2024 Feb 16; Available from: http://arxiv.org/abs/2402.10790
4. Wang S, Koopman R. Semantic embedding for information retrieval [Internet]. Available from: http://www.worldcat.org/
5. Gao Y, Xiong Y, Gao X, Jia K, Pan J, Bi Y, et al. Retrieval-Augmented Generation for Large Language Models: A Survey. 2023 Dec 18; Available from: http://arxiv.org/abs/2312.10997
6. Vuokko R, Vakkuri A, Palojoki S. Systematized Nomenclature of Medicine–Clinical Terminology (SNOMED CT) Clinical Use Cases in the Context of Electronic Health Record Systems: Systematic Literature Review. Vol. 11, JMIR Medical Informatics. JMIR Publications Inc.; 2023.