Automated OMOP Concept Mapping Using Multi-Agent Large Language Models and Graph-Enhanced Semantic Retrieval

Adil Ahmed[1], Selvin Soby[1], Boudewijn Aasman[1], Parsa Mirhaji[2]
[1] Montefiore Medicine, [2] Albert Einstein College of Medicine at Montefiore

Background

Mapping local clinical terminologies to standard OMOP concepts remains a foundational task in observational health data harmonization. However, current tools such as USAGI depend heavily on string matching and lexical similarity and offer limited support for semantic disambiguation[1]. These approaches perform poorly when mapping non-standard, institution-specific terms, particularly abbreviations, misspellings, and domain-specific variants. Manual curation is time-consuming, not as scalable, and subject to inter-mapper variability. Furthermore, existing tooling fail to fully leverage the hierarchical and semantic richness of OMOP vocabularies, including ancestor-descendant relationships, specimen types, and contextual metadata, limiting the quality of mappings for downstream cohort identification and analytics[2]. Additionally, the assumption that existing manual mappings represent ground truth is problematic, necessitating robust validation methodologies for automated systems[3]. To address these limitations, we developed a multi-agent pipeline using large language models (LLMs) and a graph-enhanced retrieval framework to support automated, scalable, and semantically informed concept mapping.

Methods

We developed a multi-stage, modular pipeline for automated OMOP concept mapping, with initial focus on the measurement domain, using LLMs, biomedical text embeddings, vector similarity retrieval, and ontology-based graph traversal. The system was designed to process proprietary or institution-specific clinical labels and generate a single standardized OMOP concept identifier along with explanatory metadata to support downstream validation and review.

The pipeline comprises four sequential stages: (1) concept expansion, (2) semantic retrieval with graph-based enrichment, (3) candidate filtering, and (4) final concept selection.

In the initial stage, an LLM performs concept expansion and normalization. Given a raw source term, typically an unstandardized label from an electronic health record system (e.g., “Hgb-HospA”), the model generates a semantically expanded version of the term in natural language. This step improves alignment with standardized clinical vocabulary by incorporating clinical context and resolving abbreviations, synonyms, and institution-specific variations. This stage used *Claude 4 Opus* accessed via secure API endpoints, with structured outputs validated using a schema enforcement layer built with Pydantic.

In the second stage, the normalized text is embedded using a pretrained domain-specific model, S-PubMedBert-MS-MARCO, a fine-tuned version of PubMedBERT optimized for biomedical sentence-level similarity. Embeddings were computed with L2 normalization to enable cosine similarity–based retrieval[4]. The vector store backend was implemented using ChromaDB, a persistent, metadata-supported storage layer enabling efficient top-k nearest neighbor queries. The retrieved concepts (top 20) are augmented with metadata extracted from a custom knowledge graph constructed from the OMOP vocabulary tables.

The OMOP knowledge graph was implemented using the NetworkX library and included nodes representing each OMOP concept with attributes such as domain identifier, vocabulary source, and class. Directed edges encoded semantic relationships defined in the CONCEPT_RELATIONSHIP and CONCEPT_ANCESTOR tables, including hierarchical links (e.g., “Is a”, “Subsumes”), compositional relationships (e.g., “Has component”), and structural associations (e.g., “Panel contains”). Synonym information was integrated via CONCEPT_SYNONYM entries. For each retrieved candidate concept, the system performed a traversal of the graph to extract relevant ancestors, descendants, associated panels, specimen types, and measurement units, thereby creating a contextual profile for each candidate.

In the third stage, the list of enriched candidate concepts is evaluated by a second LLM serving as a filtering agent. The agent was prompted with the original input label and metadata for each candidate. It applied a structured decision framework to assign a likelihood rating (“High,” “Medium,” or “Low”) to each concept along with a rationale. Candidates rated as “Low” were removed from the list. This filtering step significantly reduces the search space and improves the signal-to-noise ratio in downstream selection, while also minimizing token usage for subsequent LLM inference.

In the final stage, a third LLM agent receives the filtered list and performs concept selection. The agent reviews the contextual metadata, relevance scores, and semantic content to identify the most appropriate OMOP concept. The output includes the selected concept identifier, a natural language justification, a confidence estimate (“High,” “Moderate,” or “Low”), and a list of alternative candidate identifiers. Outputs are serialized in structured JSON for downstream integration and human validation.

To enable human-in-the-loop review, we developed a REDCap-based interface for structured expert annotation. Clinical informatics reviewers assessed each mapping for correctness, provided alternate suggestions when necessary, and rated their confidence in the system-generated outputs. These annotations were used to construct a gold-standard dataset and assess inter-rater reliability, system accuracy, and calibration performance. To address potential bias in existing manual mappings, we implemented a blind validation approach where reviewers are presented with both the LLM-selected concept and the previously mapped concept without identification, along with the original EHR term and source context.

The entire pipeline was orchestrated using LangGraph for state machine management and LangChain for LLM tool interfacing. All data processing tasks, including vocabulary extraction and preprocessing, were performed using Python-based data science libraries including pandas and NumPy. The implementation was containerized and executed in a secured computational environment integrated with our institution’s research data warehouse.

Results

The pipeline was evaluated using 100 proprietary laboratory measurement terms from Montefiore Medicine's institutional EHR data warehouse, representing real-world challenges including legacy terminology, abbreviations, and department-specific labels previously mapped through manual processes. The LLM-based system achieved direct concordance with existing manual mappings in 48% of cases, with an additional 22 cases containing the previously mapped concept among the top 4-7 filtered candidates, yielding a broader agreement rate of 70%. Recognizing that manual mappings may not represent perfect ground truth, we implemented a blind validation approach using REDCap that presents reviewers with both the LLM-selected and previously mapped concepts without identification, addressing inherent uncertainty in clinical terminology mapping. Total inference cost was $5.05 for 100 mappings using Claude Opus 4, consuming 1,175,399 input tokens and 101,613 output tokens across all pipeline stages. The blind validation process will establish a more reliable gold standard, with direct matches assumed valid and disagreement cases undergoing expert review to determine optimal mappings.

Conclusion

This work demonstrates the feasibility of using multi-agent LLM frameworks to automate OMOP concept mapping at scale with high fidelity. By integrating vector-based semantic retrieval, graph-based enrichment, and LLM-based filtering and selection, the system captures both lexical and ontological dimensions of concept similarity. The resulting pipeline outperforms manual and string-based approaches in accuracy, interpretability, and scalability. Current limitations related to cost and ambiguous input resolution are being addressed through hierarchical clustering of retrieved concepts and fine-tuning of smaller, domain-specific models. Future work includes expansion to procedures and conditions, integration into production OMOP-CDM workflows, and deployment in multi-site research environments. This approach holds potential to accelerate data standardization efforts essential for distributed observational research and learning health systems.

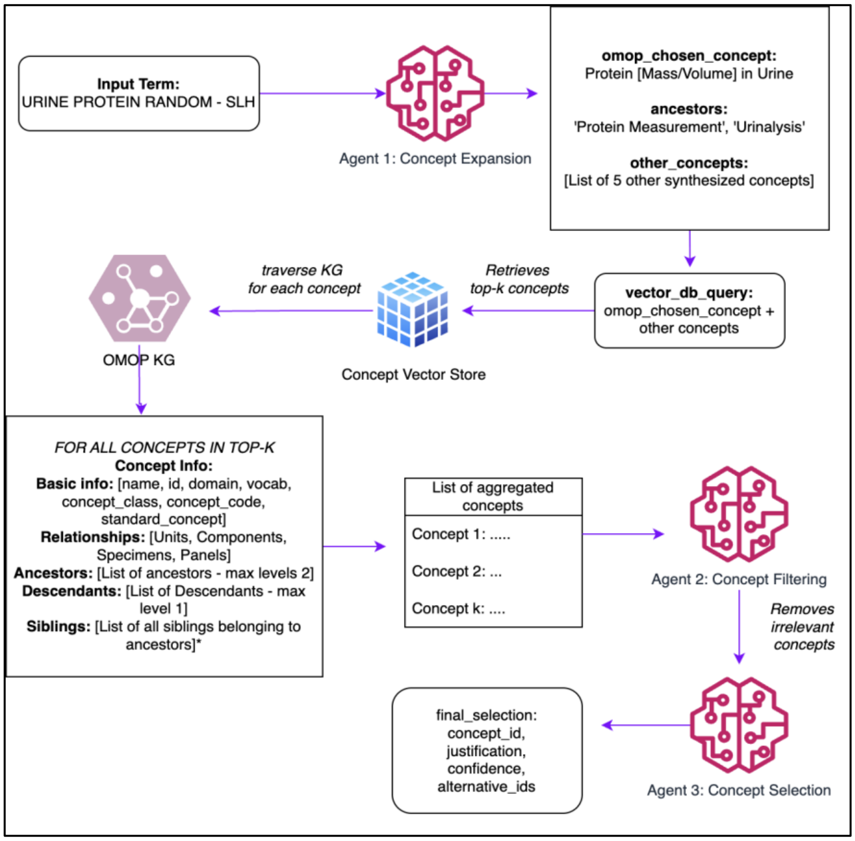
Input Term:
URINE PROTEIN RANDOM - SLH
Agent 1: Concept Expansion
omop_chosen_concept:
Protein [Mass/Volume] in Urine
ancestors:
'Protein Measurement', 'Urinalysis'
other_concepts:
[List of 5 other synthesized concepts]
vector_db_query:
omop_chosen_concept + other concepts
Retrieves top-k concepts
Concept Vector Store
traverse KG for each concept
OMOP KG
FOR ALL CONCEPTS IN TOP-K
Concept Info:
Basic info: [name, id, domain, vocab, concept_class, concept_code, standard_concept]
Relationships: [Units, Components, Specimens, Panels]
Ancestors: [List of ancestors - max levels 2]
Descendants: [List of Descendants - max level 1]
Siblings: [List of all siblings belonging to ancestors]*
List of aggregated concepts
Concept 1: .....
Concept 2: ...
Concept k: ....
Agent 2: Concept Filtering
Removes irrelevant concepts
Agent 3: Concept Selection
final_selection:
concept_id,
justification,
confidence,
alternative_ids

Figure 1: Workflow depicting the concept mapping process

References:

1. Xiao G, Pfaff E, Prud'hommeaux E, Booth D, Sharma DK, Huo N, Yu Y, Zong N, Ruddy KJ, Chute CG, Jiang G. FHIR-Ontop-OMOP: Building clinical knowledge graphs in FHIR RDF with the OMOP Common Data Model. *J Biomed Inform*. 2022 Sep 8;134:104201. doi:10.1016/j.jbi.2022.104201. PubMed PMID: 36089199; PMCID: PMC9561043
2. Wu J, Zhu J, Qi Y, Chen J, Xu M, Menolascina F, Grau V. Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation. arXiv [Preprint]. 2024 Aug 8. Available from: https://arxiv.org/abs/2408.04187
3. Cho S, Sin M, Tsapepas D, Husain SA, Natarajan K, Mohan S, et al. Content Coverage Evaluation of the OMOP Vocabulary on the Transplant Domain Focusing on Concepts Relevant for Kidney Transplant Outcomes Analysis. *Appl Clin Inform*. 2020 Oct 7;11(4):650–658. doi:10.1055/s-0040-1716528. PubMed PMID: 33027834; PubMed Central PMCID: PMC7557323.
4. Kang M, Alvarado-Guzman JA, Rasmussen LV, Starren JB. Evolution of a Graph Model for the OMOP Common Data Model. *Appl Clin Inform*. 2024 Oct;15(5):1056–1065. doi:10.1055/s-0044-1791487. PubMed PMID: 39631779; PMCID: PMC11617070