Original Paper

Evaluating Large Language Models for Standardizing Clinical Data Using Health Level Seven Fast Healthcare Interoperability Resources: Development and Evaluation Study

Álvaro Riquelme-Tornel[1,2], MSc; Pedro Luis Costa del Amo[3], PhD; Catalina Martínez-Costa[1,2], PhD

[1]Department of Informatics and Systems, Faculty of Computer Science, Universidad de Murcia, Murcia, Spain
[2]Instituto Murciano de Investigación Biosanitaria, Murcia, Spain
[3]Latency Data, Madrid, Spain

Corresponding Author:

Catalina Martínez-Costa, PhD
Department of Informatics and Systems
Faculty of Computer Science, Universidad de Murcia
Campus de Espinardo
Murcia 30100
Spain
Email: cmartinezcosta@um.es

Abstract

Background: One of the barriers to the wider adoption of clinical information modeling standards is the high cost associated with the time and effort required to model and map data from legacy systems, as well as the need for specialized modeling experts. Meanwhile, large language models (LLMs) are becoming powerful tools for data transformation and semantic processing.

Objective: The aim of this study was to evaluate the capability of LLMs to support the mapping of structured clinical data to Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) resources and attributes.

Methods: Using data from the MIMIC-IV (Medical Information Mart for Intensive Care IV) database, a data pipeline integrating clustering techniques, retrieval-augmented resource grounding, and schema-guided prompting was implemented to support the FHIR mapping of structured clinical data. The pipeline was evaluated under 2 scenarios with different levels of complexity. In the first scenario, used as a baseline, the input data were well described and contextually rich. In the second scenario, designed as a controlled test under substantially reduced context, the original table structure and table-level descriptions were removed, while data element names and LLM-enriched descriptions were retained.

Results: Using 65 data elements, top 2 resource identification accuracy reached 100% in the baseline scenario. The highest mean attribute-level mapping accuracy was 80.61% (SD 2.06) for GPT-4o and 69.54% (SD 5.15) for Llama 3.3 70B. In the low-context scenario, clustering-based context recovery increased top 5 resource identification accuracy to 94%. Across temperature settings, GPT-4o achieved mean attribute-level mapping accuracies ranging from 73.23% (SD 2.92) to 76.77% (SD 1.14), whereas Llama 3.3 70B ranged from 54.31% (SD 4.17) to 58.31% (SD 2.23). Error analysis showed that the main remaining failures were associated with resource-level confusion and terminal-attribute semantic ambiguity, whereas other error categories were comparatively infrequent.

Conclusions: LLMs can support structured clinical data mapping to HL7 FHIR, but fine-grained semantic disambiguation remains the principal limitation. In context-rich settings, the main errors occurred when selecting the appropriate terminal FHIR attribute. Under reduced context, accurate candidate-resource retrieval became increasingly important, and clustering-based context recovery substantially improved resource identification. These findings support the use of LLMs as assistive components within human-supervised interoperability workflows, with particular attention to candidate-resource selection and ambiguity-sensitive attribute mappings.

JMIR Med Inform 2026;14:e100368; doi: 10.2196/100368

Keywords: clinical data interoperability; large language models; retrieval-augmented generation; clustering; HL7 FHIR; Health Level Seven; Fast Healthcare Interoperability Resources

Introduction

Background

The potential of health data to advance health care delivery and biomedical research is widely recognized [1]. However, despite the availability of mature technologies, interoperability standards, and implementation frameworks, their adoption across health systems remains limited [2]. In the context of the European Health Data Space (EHDS), the implementation of common digital health standards and interoperability frameworks [3] is one of the challenges to achieving a fully integrated ecosystem. A recent EHDS-related survey provided a good estimation of current health data scenarios in Europe, including the use of standardization [4]. The survey revealed that most European Union (EU) countries have not yet fully implemented open infrastructures for health data sharing, a situation attributed in part to the low adoption of health data standards in these countries.

For decades, semantic interoperability standards have been proposed to enable accurate, meaningful, and consistent exchange of health care data across different systems and organizations. Terminologies such as the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [5], International Classification of Diseases (ICD) [6], and Logical Observation Identifiers Names and Codes (LOINC) [7] provide standardized codes for representing medical concepts. Clinical information modeling standards and specifications such as Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) [8], Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) [9], and openEHR [10] reuse these terminologies and define how clinical data should be structured, represented, and exchanged.

Although several standards coexist, and their selection may depend on specific needs or use cases [11], they share a common implementation challenge: modeling data according to these standards remains a labor-intensive and expert-dependent task. New data sources often require custom data pipelines, extensive manual mapping, and continuous maintenance. Focusing on HL7 FHIR, Marfoglia et al [12] proposed a modular conversion framework based on the use of templates to formalize mappings. However, the framework still required manual mapping and refinement. Similarly, Carbonaro et al [13] implemented a pipeline that involved extensive manual extraction, cleaning, and restructuring of local data. Their findings emphasized that extract, transform, and load (ETL) logic often remains tightly coupled to specific institutional databases, limiting scalability and requiring close collaboration between clinical experts, data engineers, and ontologists.

Implementation-oriented tools and FHIR harmonization pipelines have been proposed to support ETL workflows and facilitate the transformation of heterogeneous clinical data into standardized representations. Ignifyr [14], for example, provides a FHIR-first ETL environment for converting data from diverse source systems into HL7 FHIR resources. Its workflow supports connectivity with multiple data sources, the specification of FHIR resource templates, the use of FHIRPath expressions, and the testing, validation, execution, and persistence of generated FHIR resources. However, despite facilitating the operational implementation of FHIR transformation pipelines, such tools still require users to explicitly define, configure, test, and validate the underlying mapping logic. Similar limitations have been reported in previous FHIR harmonization initiatives. Williams et al [15] highlighted that the success of electronic health record (EHR)–to-FHIR harmonization pipelines largely depends on the availability and quality of local database documentation, while Zong et al [16] and Bennett et al [17] relied on manually created mappings to construct FHIR-based data profiles. Although these initiatives have successfully produced standardized data repositories, such as the MIMIC-IV on FHIR (Medical Information Mart for Intensive Care IV on FHIR) dataset, they also reveal a persistent bottleneck in the adoption of FHIR and interoperability standards more broadly: the heavy reliance on manual, expert-driven mapping between local data structures and standardized representations. This dependence increases implementation time and cost, constrains scalability and maintainability, and limits the transformation of heterogeneous legacy data into reusable standardized formats.

The rapid advancement of large language models (LLMs), such as GPT-4o (OpenAI), Gemini (Google LLC), and Llama 3 (Meta Platforms, Inc.), has expanded their applicability across numerous domains. In health care, LLMs offer a promising solution for facilitating interoperability and data reuse [18]. Their ability to interpret natural language, extract relevant information, classify concepts, and generate structured outputs has motivated their use in clinical data standardization workflows.

Building on these capabilities, several studies have investigated the integration of LLMs with interoperability frameworks such as HL7 FHIR. One line of work has examined the LLM-based conversion of unstructured clinical narratives into FHIR bundles, systematically comparing prompting strategies such as basic prompting, chain-of-thought, and 2-step decomposition approaches, alongside few-shot examples, temperature settings, and model selection [19]. Complementary efforts have addressed schema alignment between medical intake forms and FHIR resources or elements, comparing embedding-based similarity methods with LLM-based classification within constrained mapping spaces [20]. Other studies have focused on the direct generation of specific FHIR resources from free text. For example, FHIR-GPT [21] proposed a template-driven, few-shot approach for converting discharge summary medication snippets into FHIR MedicationStatement resources. Related work has also evaluated the effectiveness of LLMs in converting clinical data to FHIR representations

[22]. LLMs have also been explored as upstream normalization layers for clinical notes, including abbreviation expansion, spelling and grammar correction, and terminology normalization [23]. In parallel, patient-facing systems have leveraged existing FHIR APIs, combining prompt engineering with function-calling mechanisms to retrieve resources and generate lay explanations, while also revealing challenges related to response variability, omissions, filtering sensitivity, and reproducibility [24].

Collectively, these studies demonstrate the potential of LLMs for supporting clinical data standardization and interoperability. However, most existing approaches focus on unstructured clinical narratives, constrained form-field mappings, single-resource generation, note normalization, or patient-facing interaction with preexisting FHIR data. Less attention has been paid to structured clinical datasets that require systematic identification, validation, and transformation across a broader range of FHIR resources. This gap is particularly relevant in real-world interoperability scenarios, where clinical data are often stored in heterogeneous structured formats and must be mapped to appropriate FHIR resources through robust, traceable, and scalable procedures.

To address this gap, this study proposes a multistage prompt-chaining pipeline for mapping structured clinical inputs to HL7 FHIR. Unlike single-step generation approaches or studies focused on narrow mapping spaces, the proposed workflow separates resource identification, validation, and mapping refinement into distinct stages. This design aims to improve robustness, consistency, and interpretability in the transformation process, while expanding coverage to a broader set of FHIR resources, more representative of real-world interoperability requirements. In contrast to prior work centered mainly on unstructured narratives, intake-form schema matching, or the generation of isolated FHIR resources, our approach focuses on structured clinical data and explores how LLMs can support expert-driven interoperability workflows without fully replacing the need for human validation.

Despite these advances, the deployment of LLMs in clinical data standardization remains constrained by significant challenges. Beyond scalability, it is critical to ensure the reliability, consistency, and clinical safety of model outputs. A key barrier lies in the model's ability to accurately interpret complex, domain-specific medical contexts without introducing errors or unsupported mappings. To address these limitations, recent studies have emphasized approaches such as fine-tuning, on-premises deployment, retrieval-based methods, and human-in-the-loop validation to improve contextual grounding, data privacy, and trustworthiness in clinical environments [24].

Study Objective and Research Questions

The objective of this work was to evaluate the capability of LLMs to support the mapping of structured clinical attributes to HL7 FHIR resources and attributes. While most existing studies addressed narrow data domains such as diagnoses or medications, they often did not consider more complex and realistic mapping scenarios requiring multiple candidate FHIR resources and explicit attribute-level disambiguation.

Using data from MIMIC-IV, a pipeline integrating clustering-based context recovery, retrieval, and schema-guided prompting was evaluated under 2 scenarios with differing contextual richness. Because the goal was to assess LLM behavior in structured-data mapping, only structured data were used, and entity identification from free text was intentionally excluded.

Specifically, this study addressed the following research questions (RQs):

RQ1: How accurately can GPT-4o and Llama 3.3 70B map structured clinical data elements to HL7 FHIR resources and attributes when rich structural and descriptive metadata are available, and how does prompting strategy affect this performance?

RQ2: When the original table-level structure and descriptions are unavailable, to what extent can clustering-based context recovery improve FHIR resource retrieval and downstream attribute-level mapping?

RQ3: How stable are the resulting mappings across decoding temperatures, and which error types constitute the main remaining sources of mapping failure?

Methods

Experimental Design

A controlled experimental design was used to evaluate clinical data mapping to FHIR under variations in 3 factors: contextual metadata availability, resource-retrieval strategy, and prompting configuration. All experiments used the same MIMIC-IV–derived evaluation set, the same HL7 FHIR reference corpus, and the same evaluation criteria. Performance was assessed at 2 levels: resource identification and data element–to-attribute path mapping. A total of 2 scenarios were defined: a baseline scenario with richer metadata and preserved table structure, and a low-context scenario with reduced metadata and weaker contextual grounding.

Data Sources

The MIMIC-IV database [25] was used in this study, which contains deidentified intensive care unit (ICU) patient data from Beth Israel Deaconess Medical Center between 2008 and 2019. The database is organized into 3 source-based modules: hosp, icu, and note. We used the hosp and icu modules and excluded the note module, which contains free-text clinical narratives. The hosp module includes 22 tables derived from the hospital-wide EHR, covering demographics, admissions, laboratory and microbiology results, medication administrations, and billed diagnoses. The icu module comprises 9 tables with ICU clinical data, including intravenous medications, ventilator settings, and other routinely charted variables.

More specifically, the MIMIC-IV on FHIR representation was used as the gold standard for evaluating the

mappings [17]. MIMIC-IV on FHIR was originally implemented in HL7 FHIR Release 4 (R4), whereas this study targets Release 5 (R5). We therefore adapted the reference mappings to R5 resource structures and attributes wherever they differed from R4, and used these adapted mappings as the gold standard for scoring model predictions. Both the FHIR reference corpus given to the models and the evaluation mappings were thus represented in R5. The adaptations are detailed in Multimedia Appendix 1. Tables from the hosp and icu modules were selected based on two criteria: (1) inclusion of clinically relevant patient-care data and (2) availability of a direct HL7 FHIR mapping. Applying these criteria, the following 17 tables were included. From the hosp module: patients, admissions, procedures_icd, diagnoses_icd, hcpcsevents, microbiologyevents, prescriptions, transfers, pharmacy, services, and emar. From the icu module: d_items, datetimeevents, icustays, inputevents, outputevents, and procedureevents. Figure 1 depicts the correspondence between the selected MIMIC tables and the FHIR resources. The 17 tables contain 119 data elements, of which 65 were retained for evaluation. We excluded elements with no reference mapping in MIMIC-IV on FHIR, and those whose source-to-FHIR mapping pattern was already covered by a retained element, to avoid duplicate evaluation of equivalent patterns (not on the basis of clinical importance). The full list of the 65 included elements and their reference mappings are provided in Multimedia Appendix 2.

Figure 1. Correspondence between the selected tables from MIMIC (Medical Information Mart for Intensive Care IV) and Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR).

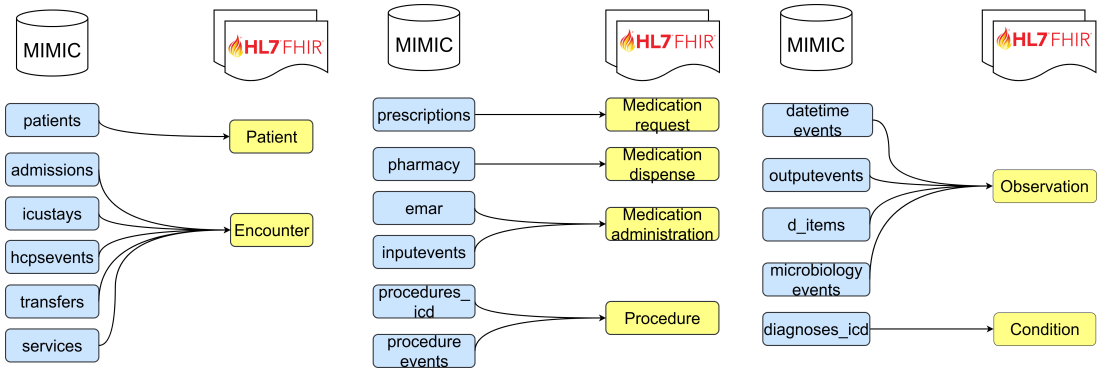
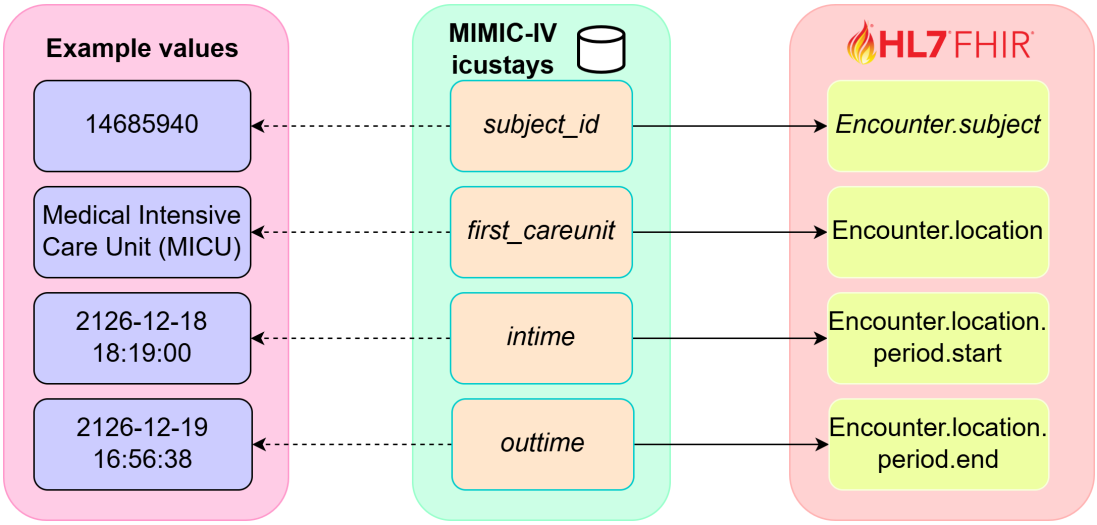


Figure 2 summarizes the data elements and sample values from the icustays table and their correspondence in FHIR. Example values (left) illustrate each icustays data element (center), and dotted arrows indicate the reference correspondence to the target FHIR fields (right). Specifically, subject_id maps to Encounter.subject (patient identifier/reference), first_careunit maps to Encounter.location (the first care-unit location where the patient was treated), intime maps to Encounter.location.period.start (admission time to that care unit), and outtime maps to Encounter.location.period.end (discharge time from that care unit). Detailed attribute definitions were available on the official MIMIC-IV website [25]. Each selected table is accompanied by a brief description of its intended use and detailed definitions of its data elements and permissible values. All tables were downloaded in CSV format and contained only data element values.

Figure 2. Ground-truth mapping used for evaluation between data elements in the MIMIC-IV (Medical Information Mart for Intensive Care IV) icustays table and their corresponding attributes in the Health Level Seven (HL7) Fast Healthcare Interoperability Resources (FHIR) Encounter resource.



In addition, a corpus describing the HL7 FHIR standard was developed. It will be referred to as the FHIR reference standard corpus hereafter, covering 45 HL7 FHIR resources. For each resource, the corpus included a brief description of its purpose and JSON-formatted definitions of all associated attributes. This corpus was used to provide contextual grounding for the LLMs.

Ethical Considerations

This study involved secondary methodological analysis of previously deidentified data from the MIMIC-IV database, accessed under PhysioNet credentialed access requirements. No participants were recruited, no new patient data or biological samples were collected, and no reidentification was attempted. The University of Murcia Institutional Research Ethics self-assessment indicated that formal review by the University Research Ethics Committee was not required [26].

Only structured data from the MIMIC-IV hosp and icu modules were used; the note module containing free-text clinical narratives was excluded. Patient-level values were used only in embedding-based preprocessing, indexing, context recovery, and clustering and were never included in prompts sent to external LLM services. GPT-4o (Azure OpenAI) and Llama 3.3 70B (Together AI) received only schema-level metadata and FHIR reference information, thereby minimizing privacy risks associated with third-party inference services. Azure OpenAI was used in accordance with PhysioNet recommendations for responsible use of MIMIC data [27].

Task and Evaluation Metrics

The task involved mapping selected structured clinical data elements from MIMIC-IV tables to HL7 FHIR resources and attribute paths. For each data instance corresponding to a MIMIC-IV data element (table column), the model was required to identify the most appropriate target HL7 FHIR resource and 3 attribute paths, producing a structured mapping output. For example, given the intime column from the icustays table, the reference mapping corresponds to the FHIR Encounter resource and the Encounter.location.period.start attribute path.

Model inputs included MIMIC-IV table and column names, descriptions when available, and contextual information retrieved from the FHIR reference corpus. Performance was evaluated by comparing model outputs against the R5-adapted MIMIC-IV on FHIR reference mappings [17].

Two evaluation scenarios were considered: (1) a baseline scenario with context-rich, well-documented data (17 tables and well-separated data elements) and (2) a low-context scenario with reduced structural and contextual information, in which the original table organization and table-level descriptions were removed, while data element descriptions were retained and all data elements were presented in a single unified table with randomized ordering.

Performance was evaluated at two levels: (1) resource identification and (2) data element–to-FHIR mapping. Resource identification performance was evaluated as top-k accuracy at the data element level. For each of the 65 data elements, retrieval was considered successful if at least one of its accepted reference FHIR resources appeared among its assigned top-k candidates. In the low-context condition without clustering, candidates were retrieved independently for each data element. In the clustering condition, candidates were retrieved for each recovered cluster, and the resulting list was assigned to every data element in that cluster. Thus, both low-context conditions were evaluated over the same 65 data elements, regardless of how the candidate lists were generated. Top-k accuracy was calculated as the number of data elements with an accepted resource in their assigned top-k list divided by 65.

Data element–to-FHIR mapping performance was evaluated using attribute-level mapping accuracy. Each of the 65 evaluated data elements was assigned a binary correct or incorrect outcome by comparing the predicted FHIR attribute path or paths with the corresponding reference mapping. A data element was considered correctly mapped when at least one predicted path matched one of the accepted reference paths according to the predefined matching rules.

For example, for the intime data element from the icustays table, a prediction of Encounter.location.period.start was considered correct because it matched the reference mapping, whereas a prediction such as Encounter.location.period.end was considered incorrect.

A predicted path was accepted when it exactly matched a reference path or represented a valid specialization of that path. For example, if the reference mapping was defined as Observation.value, a prediction such as Observation.valueQuantity was considered correct because it preserved the reference path while specifying the concrete value type. In contrast, if the reference mapping was Observation.valueQuantity, a prediction of the more general Observation.value was considered incorrect because it did not reach the required level of specificity. Empty outputs and predefined no-mapping placeholders were also considered incorrect when a reference mapping was available.

For data elements with multiple accepted reference mappings, matching any one of the valid alternatives was sufficient for the data element to be classified as correct. Additional predicted paths were not independently counted as false positives, since evaluation was performed at the data element level rather than at the individual path level. Consequently, the metric evaluates whether the model identified at least one acceptable mapping for each data element, rather than the precision of the complete set of paths generated for that element.

Resource-identification performance (top-k accuracy) was calculated as:

$$\frac{N_{\text{data elements with at least one accepted reference resource in their assigned Top-}k\text{ list}}}{N_{\text{data elements evaluated}}}$$

Resource identification accuracy was evaluated at the data-element level in all conditions (n=65). In the baseline scenario, the top-2 candidate list retrieved for each

MIMIC-IV table was assigned to its constituent data elements. In the low-context scenario with clustering, the top-5 list retrieved for each recovered cluster was assigned to its constituent data elements. Without clustering, a top-5 list was retrieved independently for each data element. Each element was counted as correct when at least one of its accepted reference resources appeared in its assigned candidate list.

Data element–to-FHIR mapping accuracy was calculated as:

$$\frac{N_{correct}}{N_{evaluated}}$$

where $N_{correct}$ represents the number of correctly mapped data elements and $N_{evaluated}$=65 represents the total number of evaluated elements.

Incorrect mappings were additionally categorized using a deterministic rule-based evaluator. Error categories were mutually exclusive and were assigned sequentially according to a predefined precedence order: NoMap, ResourceConfusion, ExtensionOveruse, Specificity, and TerminalAttributeConfusion (TAConfusion). Because the categorization was performed automatically using these rules, no manual adjudication or interrater disagreement resolution was required.

NoMap was assigned when no mapping was returned or when all predicted outputs corresponded to predefined no-mapping placeholders. ResourceConfusion was assigned when the FHIR resource or resources predicted by the model did not overlap with any of the accepted reference resources, indicating that the model selected an inappropriate clinical domain despite producing a plausible attribute path. ExtensionOveruse was assigned when the prediction used an FHIR extension, although the reference mapping could be represented using a standard FHIR element. Specificity was assigned when the predicted path terminated at a more general parent element than the reference mapping, indicating that the model identified the correct structural region but did not reach the required level of granularity. TAConfusion was assigned to remaining incorrect mappings after the preceding rules had been evaluated, typically representing cases in which the model identified the correct resource or structural neighborhood but selected an incorrect terminal FHIR attribute.

To assess run-to-run stability, baseline prompting configurations were run 5 times, and low-context temperature experiments 10 times per model and temperature setting. Repetitions were kept even at temperature 0, since identical decoding parameters do not guarantee identical outputs in hosted LLM inference. For each configuration, we report the mean attribute-level mapping accuracy and, for each error category, the mean number of errors and sample SD across runs. Corresponding 95% CIs were estimated using the Student *t* distribution:

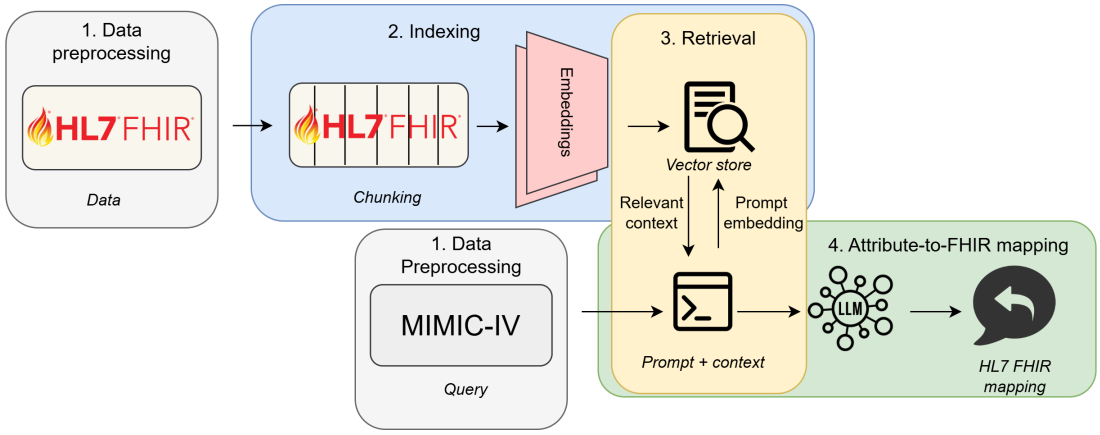
$$CI_{95\%} = \bar{A} \pm t_{0.975,\, n-1}\frac{s}{\sqrt{n}}$$

where $\bar{A}$ is the mean attribute-level mapping accuracy across repeated runs, S is the sample SD, and n is the number of repetitions. The Student *t* distribution was used only to construct 2-tailed 95% CIs, rather than for hypothesis testing. The corresponding degrees of freedom were 4 for the baseline experiments (n=5) and 9 for the low-context experiments (n=10). Because the baseline analysis included 5 repetitions per configuration, the resulting CIs were interpreted as descriptive estimates of uncertainty rather than as formal evidence of statistically significant differences between prompting strategies.

Common Pipeline

Across both scenarios, the same four-stage workflow was applied: (1) data preprocessing, (2) indexing, (3) retrieval, and (4) LLM-based mapping. These stages incorporated multiple techniques, including data enrichment, data format transformation, text embeddings, clustering methods, and interaction with the LLM (Figure 3).

Figure 3. General pipeline defined for each scenario. FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven; LLM: large language model; MIMIC-IV: Medical Information Mart for Intensive Care IV.



Experimental Scenarios

The baseline and low-context scenarios differed in 4 aspects: metadata availability, the use of clustering for context recovery, the retrieval configuration, and the number of candidate HL7 FHIR resources provided to the LLM (top-k).

Table 1 summarizes both the fixed components and the scenario-specific variations.

Because the 2 scenarios differ in several components at once (metadata availability, input organization, context recovery, retrieval configuration, and number of candidate resources), the baseline versus low-context comparison evaluates the complete pipelines, not the isolated effect of individual components. Accordingly, the top-2 and top-5 resource identification results should not be interpreted as directly comparable retrieval metrics, because the value of k was part of the scenario-specific pipeline configuration. In the following, the 4-stage workflow is described.

Table 1. Comparison of the baseline and low-context experimental scenarios.

Process	Baseline scenario	Low-context scenario
Number of evaluated data elements	65 data elements	65 data elements
Column or data element name	Available	Available
Data element description	Available	Available
Table description	Available	Removed
Data element organization	Original table grouping	All elements pooled and randomly ordered
Context recovery	Not required	Unsupervised semantic clustering
Retrieval unit	MIMIC-IV[a] table	Recovered cluster+data element signals
HL7[b] FHIR[c] resource candidates (top-k)	k=2	k=5
HL7 FHIR resource description	Available	Available
HL7 FHIR resource attribute descriptions	Available	Available
Prompting strategy	Four strategies, temperature= 0	Selected strategy across temperature=0, 0.5, and 1
Evaluation gold standard	65 R5[d]-adapted mappings	65 R5-adapted mappings

[a]MIMIC-IV: Medical Information Mart for Intensive Care IV.
[b]HL7: Health Level Seven.
[c]FHIR: Fast Healthcare Interoperability Resources.
[d]R5: Release 5.

Stage 1: Data Preprocessing

Overview

This stage aimed to improve data quality through two steps: (1) enrichment of attribute descriptions and (2) adoption of a common data format.

The original MIMIC-IV documentation provided limited descriptions for many tables and columns. Semantic enrichment is particularly important when using embeddings with LLMs, as richer and more precise descriptions generally enable the models to produce more accurate and meaningful outputs. By using LLMs, these descriptions were enriched for each of the table data elements or columns Multimedia Appendix 3.

Previous studies showed how data presented to the model had a substantial impact on its performance [28]. Using JSON format for representing tabular information improves model interpretability and comprehension, leading to improved task performance. Accordingly, JSON was adopted as the standard data format in this study.

To quantify the contribution of description enrichment, an additional ablation analysis was conducted in the low-context scenario using the original MIMIC-IV data element descriptions while keeping the clustering, retrieval, model, and prompting configurations unchanged.

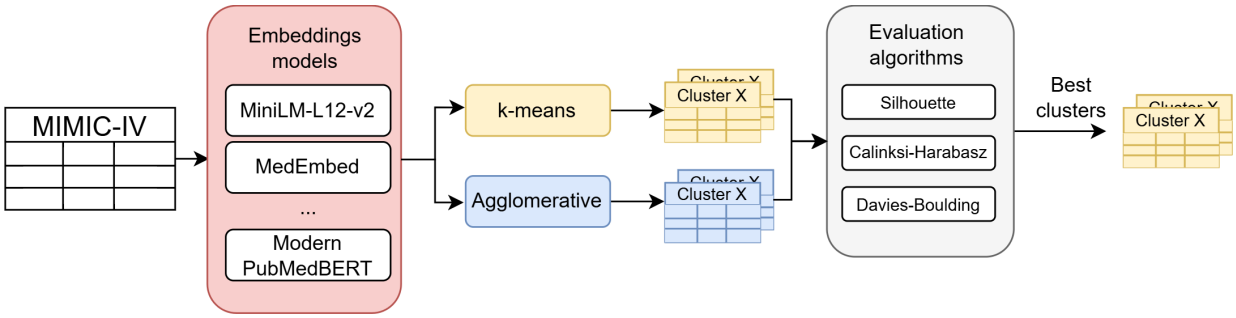
Stage 1a: Context Recovery via Clustering (Low-Context Scenario)

In the low-context scenario, input data were presented as a unified table, in which columns were randomly distributed and lacked table-level context. Under these conditions, single–data element retrieval offered insufficient context for reliable FHIR resource identification. Preliminary experiments that retrieved candidate resources independently per data element using only column names and/or descriptions produced unstable and frequently implausible results, reflecting the limited semantic information available at the data element level in sparse-metadata settings.

To mitigate this limitation, unsupervised semantic clustering was used to recover context by grouping related data elements before retrieval. These clusters then served as input units for subsequent retrieval and LLM-based mapping. Several biomedical embedding models were screened for context recovery and retrieval in the low-context scenario, including MiniLM-L12, all-mpnet-base-v2, MedEmbed-large, PubMedBERT, BioBERT, Gemma, and ModernPubMedBERT [29-34]. In addition, several clustering methods were evaluated, including K-Means, agglomerative

clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH), Ordering Points To Identify the Clustering Structure (OPTICS), and spectral clustering [35-40]. The best-performing configurations were selected using standard internal validity criteria [41-43]. Following this comparative evaluation, all-mpnet-base-v2 was selected as the embedding model, and agglomerative clustering was selected as the clustering approach, using complete linkage, cosine distance, and 21 clusters. Figure 4 summarizes this context-recovery workflow.

Figure 4. Clustering workflow for the unified MIMIC-IV (Medical Information Mart for Intensive Care IV) table. BERT: bidirectional encoder representations from transformers.

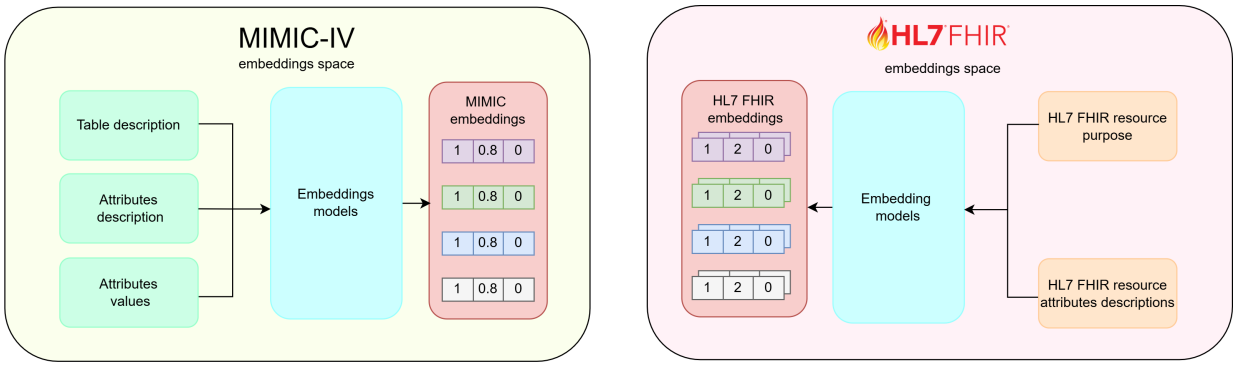


Stage 2: Indexing

In this stage, retrievable vector representations were created for both the HL7 FHIR reference corpus and the MIMIC-IV input data to be mapped. For MIMIC-IV, the indexed unit corresponded to the original table in the baseline scenario and to the recovered cluster in the low-context scenario, together with the available descriptive metadata and data element–level information. For the HL7 FHIR reference corpus, representations were built from resource-level purpose statements and element descriptions.

Indexing was adapted to the characteristics of each scenario. In the baseline scenario, retrieval relied on a simpler combination of lexical and semantic representations [44-47]. In the low-context scenario, the all-mpnet-base-v2 representation selected during the context-recovery comparison was reused for indexing and downstream retrieval to maintain representational consistency. Figure 5 summarizes the indexing process.

Figure 5. High-level overview of embedding-space construction. FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven; MIMIC-IV: Medical Information Mart for Intensive Care IV.

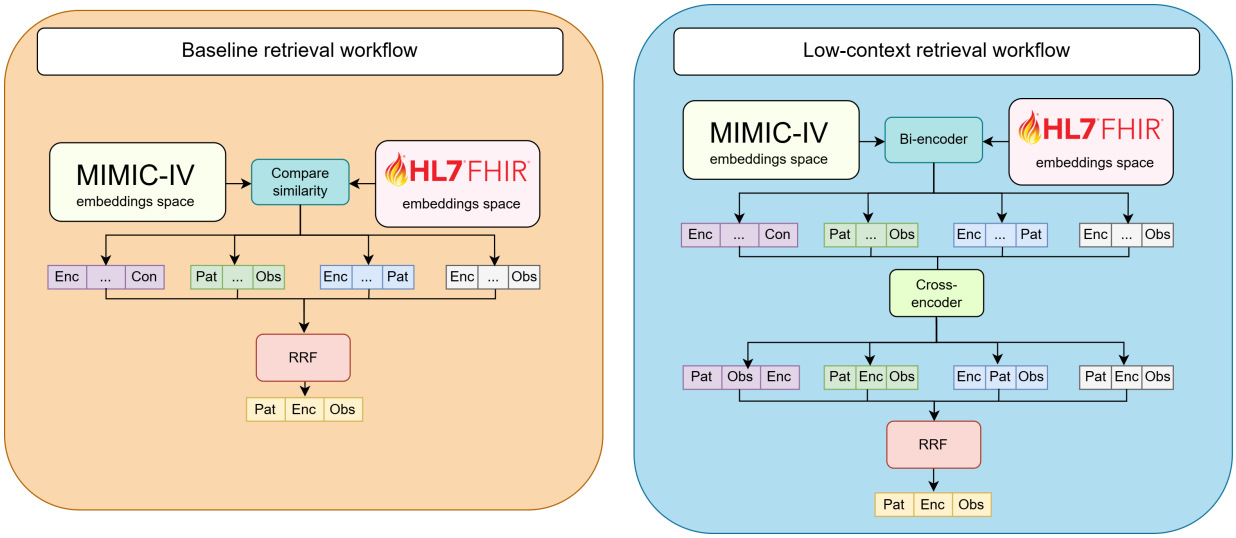


Stage 3: Retrieval

Once vector representations had been created for both the MIMIC-IV input data and the HL7 FHIR reference corpus, similarity-based retrieval [48] was used to identify candidate FHIR resources for each input unit. In the baseline scenario (Figure 6, left side), each MIMIC-IV table was matched directly against the HL7 FHIR resource corpus using a simpler retrieval configuration, and the top 2 candidate resources were provided as context for downstream mapping.

In the low-context scenario (Figure 6, right side), retrieval followed a more robust multistage configuration to compensate for sparse and heterogeneous metadata. Candidate resources were identified using both cluster-level and data element–level signals, and the resulting rankings were combined before selecting the top 5 resources for the LLM-based mapping stage. This design aimed to improve contextual grounding when resource identification could not be reliably resolved from isolated column descriptions alone.

Figure 6. Illustration of the embedding-based similarity comparison in the embedding space and the subsequent extraction of the most similar Health Level Seven Fast Healthcare Interoperability Resources after applying reciprocal rank fusion [49]. Enc: Encounter resource; Pat: Patient resource; Obs: Observation resource; Con: Condition resource. FHIR: Fast Healthcare Interoperability Resources; HL7: Health Level Seven; MIMIC-IV: Medical Information Mart for Intensive Care IV; RRF: reciprocal rank fusion.



Stage 4: Data Element–to-FHIR Mapping

Once candidate resources had been selected in the previous stage, they were provided to the LLM as contextual information to generate the corresponding data element–to-FHIR mapping predictions. Based on the experiments performed (Results section), the top 2 candidate resources were provided in the baseline scenario, whereas the top 5 candidate resources were used in the low-context scenario. Explicit output constraints were also applied to guide the model's responses.

For each data entity to be mapped (table or cluster data element or column), the prompt included: (1) the input entity to be mapped, including the available table or cluster-level context and column descriptions; (2) the candidate HL7 FHIR resources identified during the retrieval stage; and (3) supporting context derived from the reference resources, such as resource and attribute descriptions.

Table 2 presents an illustrative example of data element–to-FHIR mappings produced for the MIMIC-IV icustays table when paired with the HL7 FHIR Encounter resource.

Table 2. Illustrative example of data element–to-FHIR[a] mapping output.

icustays data elements or columns	FHIR[a] attribute paths
subject_id	Encounter.subject
first_careunit	Encounter.location
intime	Encounter.location.period.start
outtime	Encounter.location.period.end

[a]FHIR: Fast Healthcare Interoperability Resources.

Models and Prompt Strategies

GPT-4o [50] and Llama 3.3 70B [51] were selected as the 2 general-purpose LLMs, representing proprietary and open-source approaches, respectively. Due to the high cost of fine-tuning, both models were evaluated in their original, non–fine-tuned form. GPT-4o was accessed through Microsoft Azure OpenAI using API version 2025-01-01-preview, whereas Llama 3.3 70B Instruct was accessed through the Together AI API. Additional implementation details, including generation parameters, embedding and clustering configurations, retrieval settings, and software specifications, are provided in Multimedia Appendix 4. The 65 data elements included in the final evaluation set, their reference mappings, and their evaluation outcomes were not used to design or optimize the prompting strategies. Prompt design was performed independently of the final evaluation procedure. Model performance was assessed in terms of mapping accuracy and behavioral consistency across prompt and decoding configurations. A total of 4 prompting strategies were compared to evaluate the effect of prompt structure on data element–to-FHIR mapping performance. The first is self-reflexive prompting, in which an initial response was generated and then revised through self-review to reduce errors. The second is mixture of prompts (MoP), a prompting strategy that started from the same initial mapping template as the other configurations but applied several alternative reflection prompts during review. Each prompt independently reassessed the initial mapping under a different review instruction, keeping the template, input data, retrieved FHIR

context, and output structure unchanged. A synthesis judge then reconciled the resulting mappings into the final one, following predefined schema- and semantic-priority rules. The third is 5-Serial-Schema, a 5-step prompting approach in which outputs were iteratively refined using a predefined JSON schema and contextual guidance to improve mapping accuracy and reduce formatting errors. The fourth is 5-Serial-NoSchema, a 5-step prompting approach without a predefined schema, designed to test whether iterative prompting alone was sufficient to preserve output structure and formatting.

Prompting was treated as an experimental factor rather than as a fixed component of the pipeline. In the baseline scenario, all prompting strategies were evaluated under deterministic decoding (T=0) to isolate the effect of prompt structure. In the low-context scenario, self-reflexive prompting was retained as a common prompting configuration for both models and evaluated across multiple temperature settings ($T \in \{0, 0.5, 1.0\}$).

The common prompt structure used across experiments is shown in Textbox 1. It assigned the LLM the role of a mapping assistant and provided explicit instructions to map attribute semantics based on names, descriptions, and retrieved FHIR context, without providing raw or example patient data values to the model.

Textbox 1. High-level structure of the common mapping prompt used across experiments.

Different prompting strategies were applied to this shared prompt skeleton. Complete prompting templates are provided in Multimedia Appendices 5–9. FHIR R5: Fast Healthcare Interoperability Resources Release 5.
###################### CONTEXT ######################You are a domain expert assistant specializing in mapping clinical tabular data to FHIR resources (FHIR R5).(...)The documents you have to consider to answer the query is: <context>###################### TASK ######################Based on the provided table attribute information and standard FHIR attribute definitions, determine the top 3 most specific FHIR attributes that correspond to each table attribute, listed in order of relevance (most specific first).###################### OUTPUT FORMAT ######################<Output format>Each key corresponds to the table attribute name, and its value is an array of exactly three strings. If fewer than three matches are found, use "No additional attribute found" to fill the empty slots.###################### EXAMPLE ######################<FHIR example>###################### ADDITIONAL INSTRUCTIONS ######################- Consider the attribute name and descriptions when determining the best FHIR attributes.- Return only the final JSON object without additional commentary.- Use the FHIR R5 specification as the reference (https://www.hl7.org/fhir/).- Double-check your mappings before returning the final result.- You must use the provided functions to format your response.- Ensure the JSON is well-formed and adheres to the specified structure.

Results

Baseline Scenario: Prompting Strategy and Mapping Performance

In the baseline scenario, resource identification was trivial under context-rich metadata: the correct HL7 FHIR resource was retrieved for all MIMIC-IV tables considered (top 2 accuracy=100%). This result indicates that when input data are well described, the retrieval stage does not represent the main bottleneck. Instead, performance differences across configurations were primarily driven by the data element–to-FHIR mapping of the standardization stage, which requires finer-grained semantic alignment. Table 3 reports mean attribute-level mapping accuracy and sample SD across 5 repeated runs.

Table 3. Mean (SD) attribute-level mapping accuracy for each prompting strategy evaluated with GPT-4o and Llama 3.3 70B in the baseline scenario.[a]

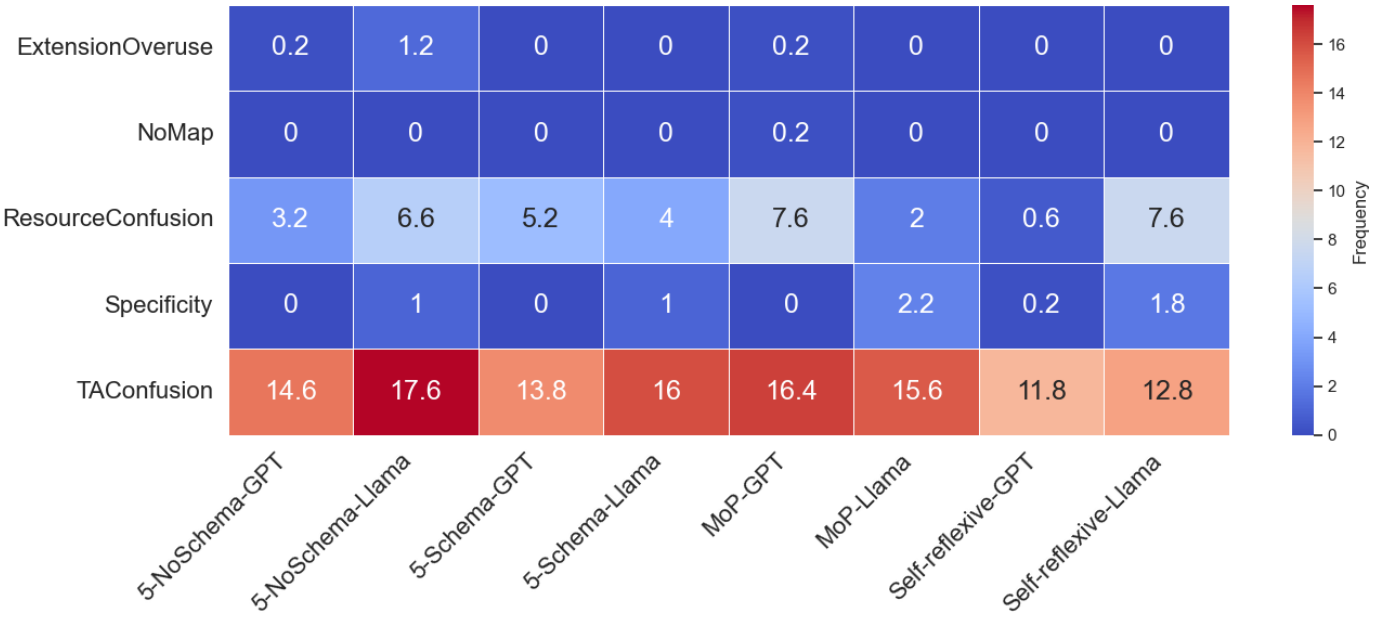
Prompts	GPT-4o, %, mean (SD)	Llama 3.3, %, mean (SD)
Self-reflexive	80.61 (2.06)	65.85 (1.69)
MoP[b]	62.46 (5.61)	69.54 (5.15)
5-Serial-Schema	70.77 (2.88)	67.69 (3.77)
5-Serial-NoSchema	72.31 (1.54)	59.38 (3.54)

[a]95% CIs for the best-performing configuration of each model are reported in the text.
[b]MoP: mixture of prompts.

Performance patterns differed between models (Table 3). GPT-4o achieved the highest mean attribute-level mapping accuracy with self-reflexive prompting (80.61%, 95% CI 78.05-83.18), whereas Llama 3.3 performed best with MoP (69.54%, 95% CI 63.15-75.93). Thus, a single self-review step was most effective for GPT-4o in the context-rich baseline setting, whereas Llama 3.3 benefited more from prompt diversification, indicating that prompting effectiveness was model-dependent.

To better understand model behavior, a detailed error analysis was conducted across all configurations. Figure 7 presents a heatmap summarizing the mean distribution of error types for each model and prompting strategy.

Figure 7. Heatmap of mean errors grouped by techniques and models. MoP: mixture of prompts; TAConfusion: TerminalAttributeConfusion.



In the baseline scenario (Table 4 and Figure 7), TerminalAttributeConfusion was the predominant error class across models and prompting strategies, indicating that mappings often reached the correct conceptual and structural neighborhood but failed to select the most appropriate terminal FHIR attribute.

ResourceConfusion was a secondary, strategy-sensitive failure mode, whereas Specificity errors and ExtensionOveruse were comparatively infrequent. Overall, the main limitation under context-rich conditions was fine-grained terminal-attribute disambiguation rather than structural mapping failure.

Table 4. Mean (SD) number of errors for each error category and prompting strategy in the baseline scenario across 5 repeated runs (n=5).

Error category	5-Serial-NoSchema		5-Serial-Schema		MoP[a]		Self-reflexive	
	GPT-4o, mean (SD)	Llama 3.3, mean (SD)	GPT-4o, mean (SD)	Llama 3.3, mean (SD)	GPT-4o, mean (SD)	Llama 3.3, mean (SD)	GPT-4o, mean (SD)	Llama 3.3, mean (SD)
TAConfusion[b]	14.6 (1.67)	17.6 (2.07)	13.80 (1.30)	16 (1.87)	16.4 (2.30)	15.6 (2.61)	11.8 (1.48)	12.8 (2.49)
ResourceConfusion	3.2 (1.1)	6.6 (1.82)	5.20 (0.84)	4 (1.22)	7.6 (1.95)	2 (0.71)	0.6 (1.34)	7.6 (3.13)
Specificity	0 (0)	1.0 (0.0)	0 (0)	1 (0)	0 (0)	2.2 (0.45)	0.2 (0.45)	1.8 (0.45)
ExtensionOveruse	0.2 (0.45)	1.2 (1.10)	0 (0)	0 (0)	0.2 (0.45)	0 (0)	0 (0)	0 (0)
NoMap	0 (0)	0 (0)	0 (0)	0 (0)	0.2 (0.45)	0 (0)	0 (0)	0 (0)

[a]MoP: mixture of prompts.
[b]TAConfusion: TerminalAttributeConfusion.

Low-Context Scenario

Self-reflexive prompting was selected as a common configuration to enable a controlled comparison of model behavior and temperature effects under identical prompting conditions. Although MoP achieved the highest baseline performance for Llama 3.3, using a common prompting strategy avoided confounding model and temperature effects with model-specific prompt optimization.

Effect of Description Enrichment

To quantify the contribution of LLM-based description enrichment, an additional ablation experiment was conducted in the low-context scenario. GPT-4o was evaluated using self-reflexive prompting and the same clustering-based context recovery and retrieval configuration, but replacing the enriched data element descriptions with the original MIMIC-IV descriptions. This analysis was designed to isolate the effect of semantic enrichment on downstream attribute-level mapping performance while keeping the remaining pipeline components unchanged. The results across the 3 temperature settings are summarized in Table 5.

Description enrichment consistently improved attribute-level mapping accuracy across all temperature settings, with absolute gains of 8.15-12.02 percentage points and a mean improvement of 10.67 (SD 2.19) percentage points (Table 5). This indicates that richer semantic descriptions provided additional grounding even when clustering-based context recovery and retrieval are kept unchanged.

Table 5. Effect of LLM[a]-based data-element description enrichment on GPT-4o attribute-level mapping accuracy in the low-context scenario.[b]

Temperature	Original descriptions (n=10), %, mean (SD)	LLM-enriched descriptions (n=10), %, mean (SD)	Absolute difference
T=0	64.44 (1.43)	76.46 (2.41)	+12.02 pp[c]
T=0.5	65.08 (1.27)	73.23 (2.92)	+8.15 pp
T=1	64.92 (2.27)	76.77 (1.14)	+11.85 pp

[a]LLM: large language model.
[b]Results are reported across 3 temperature settings using self-reflexive prompting and the same clustering-based context recovery and retrieval configuration. The absolute difference is expressed in percentage points between the enriched and original-description conditions. Each configuration was evaluated across 10 repeated runs.
[c]pp: percentage points.

Context Recovery and Retrieval Performance

A controlled comparison was conducted to assess the effect of clustering-based context recovery under otherwise identical experimental conditions. Clustering increased top 5 resource-identification accuracy from 50% to 94% and mean attribute-level mapping accuracy from 54.87% (SD 1.88%) to 76.77% (SD 1.14%; Table 6). These results show that, under limited contextual information, improved candidate-resource retrieval substantially benefits downstream mapping performance.

Table 6. Controlled comparison of resource-identification and downstream attribute-level mapping performance with and without clustering-based context recovery in the low-context scenario.[a]

Low-context retrieval setting	Resource identification (top 5), n/N (%)	Attribute-level mapping accuracy (%), mean (SD)	Best observed attribute-level mapping accuracy (%)
No clustering (attribute-level queries)	33/65 (50)	54.87 (1.88)	56.32
Clustering (cluster-level queries)	61/65 (94)	76.77 (1.14)	78.19

[a]Both conditions were evaluated using GPT-4o with self-reflexive prompting at temperature=1 under otherwise identical experimental settings (n=10 repeated runs per condition). Attribute-level mapping performance is reported as mean accuracy (SD), together with the best observed accuracy across runs. Retrieval queries were constructed per data element without clustering and per recovered cluster with clustering; top-5 accuracy was evaluated per data element in both conditions (n=65).

Temperature Robustness and Error Analysis

Once the low-context scenario configuration was established, both models were evaluated using the self-reflexive prompting strategy across different temperature settings (Table 7). Self-reflexive was selected as the common configuration to preserve comparability between models; therefore, the low-context Llama 3.3 results should be interpreted as performance under this shared prompting configuration rather than as the model-specific optimum.

Table 7. Low-context mean (SD) attribute-level mapping accuracy for GPT-4o and Llama 3.3 70B across temperature settings in the low-context scenario (n=10 runs per configuration).

Temperature	GPT-4o (%), mean (SD)	Llama 3.3 (%), mean (SD)
0	76.46 (2.41)	58.31 (2.23)
0.5	73.23 (2.92)	56.31 (1.81)
1	76.77 (1.14)	54.31 (4.17)

GPT-4o consistently outperformed Llama 3.3 across temperature settings and remained comparatively stable with mean attribute-level mapping accuracy ranging from 73.23% (SD 2.92) to 76.77% (SD 1.14 ; Table 7). Llama 3.3 showed a monotonic decrease in mean accuracy as temperature increased, from 58.31% (SD 2.23) at temperature=0 to 54.31% (SD 4.17) at temperature=1, together with greater run-to-run variability. These results were interpreted descriptively; no inferential hypothesis test was performed to compare temperatures.

Error profiles differed between models (Figure 8 and Table 8). For GPT-4o, ResourceConfusion was consistently the dominant error category across temperatures, whereas TerminalAttributeConfusion remained substantially less frequent. Llama 3.3 showed substantial contributions from both error types, with the overall error burden increasing as temperature increased. ExtensionOveruse, Specificity, and NoMap remained comparatively infrequent for both models. Overall, GPT-4o was predominantly limited by resource-level ambiguity, whereas Llama 3.3 was affected by both resource-level and terminal-attribute disambiguation errors.

Figure 8. Heatmap of mean errors grouped by techniques and models in low-context scenario. TAConfusion: TerminalAttributeConfusion.

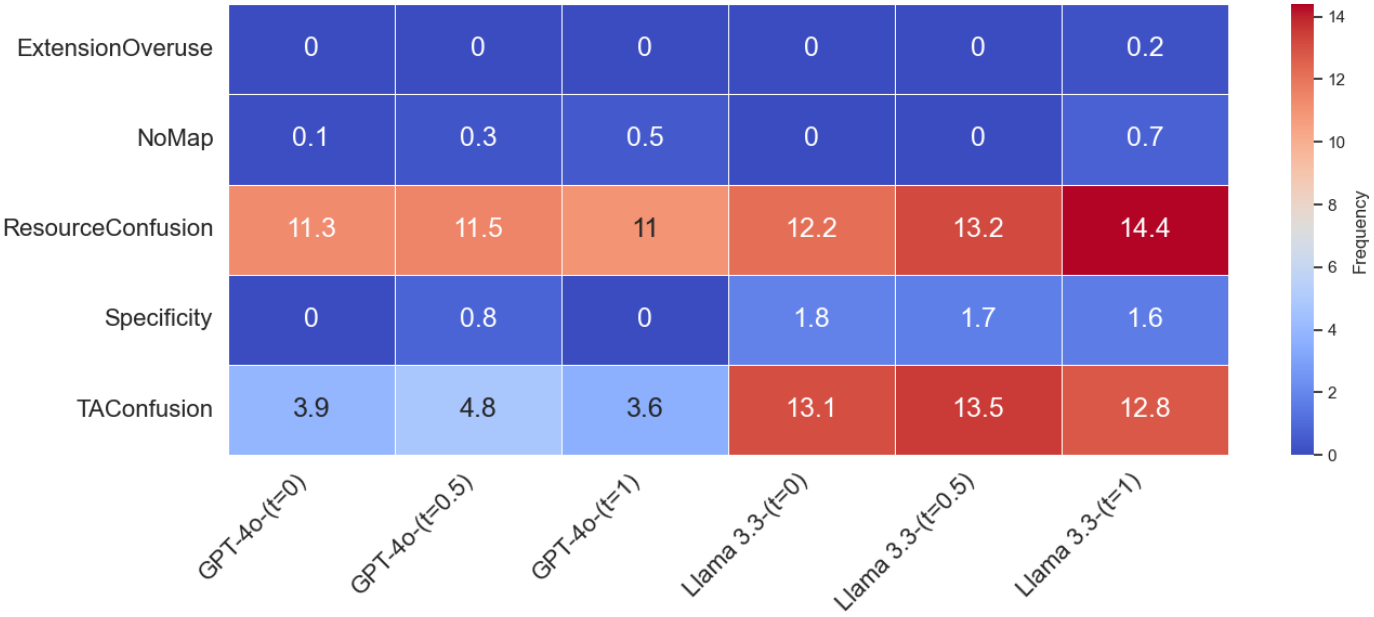


Table 8. Mean number of errors and corresponding SD for each error category across 10 repeated runs at each temperature setting in the low-context scenario.[a]

Low-context scenario error type	Temperature=0		Temperature=0.5		Temperature=1	
	GPT-4o, mean (SD)	Llama 3.3, mean (SD)	GPT-4o, mean (SD)	Llama 3.3, mean (SD)	GPT-4o, mean (SD)	Llama 3.3, mean (SD)
ExtensionOveruse	0 (0)	0 (0)	0 (0)	0 (0)	0 (0)	0.2 (0.63)
Specificity	0 (0)	1.8 (0.79)	0.8 (1.69)	1.7 (0.95)	0.0 (0)	1.6 (1.26)
NoMap	0.1 (0.52)	0 (0)	0.3 (0.48)	0 (0)	0.5 (0.53)	0.7 (1.89)
ResourceConfusion	11.3 (0.67)	12.2 (1.03)	11.5 (0.97)	13.2 (1.14)	11 (0)	14.4 (2.01)
TAConfusion[b]	3.90 (1.29)	13.1 (1.29)	4.8 (1.23)	13.5 (0.97)	3.6 (0.7)	12.8 (1.48)

[a]Error categories were mutually exclusive; therefore, the sum of the category means corresponds to the mean number of incorrectly mapped data elements per run.
[b]TAConfusion: TerminalAttributeConfusion.

Discussion

Principal Findings

This study evaluated the ability of LLMs to support the mapping of structured clinical data to HL7 FHIR under different levels of contextual information. Across both scenarios, the main unresolved challenge was fine-grained semantic disambiguation: distinguishing between semantically adjacent FHIR resources and selecting the most appropriate terminal element path within the correct structural context.

In the baseline scenario, where table structure and descriptive metadata were available, resource identification was not a limiting factor; instead, performance depended primarily on attribute-level disambiguation and prompt design. Prompting effects were model dependent: GPT-4o benefited most from self-reflexive prompting, whereas Llama 3.3 performed best with MoP (Tables 3 and 4).

In the low-context scenario, removal of table-level information shifted the main bottleneck toward candidate resource identification. Clustering-based context recovery substantially improved resource retrieval and downstream mapping (Table 6), demonstrating the importance of reconstructing contextual information when source metadata are limited.

Temperature affected the models differently: GPT-4o remained comparatively stable, whereas Llama 3.3 showed decreasing accuracy, and greater variability as temperature increased (Tables 7 and 8). However, low-context experiments used self-reflexive prompting as a common configuration for both models rather than each model's optimal baseline strategy. Because MoP performed best for Llama 3.3 in the baseline scenario, its low-context results may underestimate performance achievable with model-specific prompt optimization.

Across both scenarios, residual errors were dominated by ResourceConfusion and TerminalAttributeConfusion, indicating that fine-grained semantic disambiguation remained the central challenge. Schema-guided prompting, metadata enrichment, context recovery, and retrieval therefore appear primarily as enabling mechanisms that constrain the

mapping space and provide the grounding needed to resolve this ambiguity.

Previous studies have demonstrated the potential of LLMs for FHIR-related transformation tasks, including schema matching between medical forms and FHIR elements and the generation of FHIR resources or bundles from unstructured clinical text [19-22]. In contrast, our study evaluates structured clinical data across multiple candidate FHIR resources and explicitly separates resource identification from attribute-level mapping. Our findings extend previous observations on the importance of FHIR schema complexity and prompt formulation [20-22] by showing that candidate-resource retrieval becomes a major determinant of downstream mapping when source-schema context is reduced. Direct quantitative comparison with previously reported non-LLM mapping approaches is not straightforward because existing systems differ substantially in task definition, target data, FHIR resources, degree of manual configuration, and evaluation criteria. Prior FHIR harmonization pipelines generally rely on manually specified mappings, templates, or transformation rules rather than automated data element–to-FHIR attribute-path prediction over a common benchmark. Therefore, this study does not aim to establish superiority over non-LLM approaches, but rather to characterize the capabilities and limitations of LLMs for supporting this mapping task under controlled experimental conditions.

The results support the use of LLMs as decision support tools for data element–to-FHIR mapping rather than as fully autonomous standardization systems. Intended users include clinical informatics, interoperability, and data engineering professionals with sufficient FHIR expertise to validate ambiguous mappings. Human supervision is particularly relevant at 2 stages: candidate resource selection, where retrieval errors may propagate downstream, and attribute-level disambiguation, where closely related FHIR elements or local implementation conventions often require expert judgment. In operational use, source elements with insufficient documentation or ambiguous retrieval results should be flagged for expert review rather than automatically mapped.

Several limitations should be acknowledged. First, the experiments were conducted on a limited subset of tables and attributes from MIMIC-IV using a single reference mapping, limiting generalizability to other institutions, schema designs, and documentation practices. Second, the evaluation focused on resource identification and data element–to-attribute mapping rather than complete ETL workflows, FHIR resource instantiation, terminology coding, or clinical validation. Third, the overall pipeline remains sensitive to retrieval quality, since downstream mapping remained dependent on the candidate-resource retrieval, allowing retrieval errors to propagate through the pipeline.

Future work should extend the pipeline beyond structural matching toward full resource instantiation, semantic validation, and integration into broader ETL workflows. Particular attention should be given to improving candidate-resource retrieval and fine-grained attribute disambiguation through stronger grounding in FHIR definitions, implementation guides, and controlled terminologies. External evaluation across datasets, institutions, and schema designs should assess generalizability and the contribution of individual pipeline components, including clustering, retrieval, schema-guided prompting, temperature, and model-specific prompting strategies. Controlled comparisons with rule-based, embedding-based, and manually configured mapping approaches using the same benchmark and evaluation criteria are also needed.

Future studies should additionally investigate confidence or consistency measures derived from repeated LLM responses to identify ambiguous mappings requiring expert review. Finally, the applicability of context recovery to other clinical information models and common data models, including OMOP CDM and openEHR, should be investigated.

Conclusions

This study showed that LLMs can support the mapping of structured clinical data to HL7 FHIR, although performance depends on the contextual information available to the model. Under context-rich conditions, prompt design influenced attribute-level mapping performance, with the optimal strategy differing between GPT-4o and Llama 3.3 70B. Under reduced context, clustering-based context recovery substantially improved candidate-resource retrieval and downstream mapping.

Across both scenarios, ResourceConfusion and TerminalAttributeConfusion remained the dominant errors, indicating that fine-grained semantic disambiguation is the principal remaining challenge. These findings support the use of LLMs as assistive components within human-supervised FHIR mapping workflows rather than as fully autonomous mapping systems.

Generative AI Use

The authors declare the use of generative AI in the research and writing process. According to the GAIDeT taxonomy (2025), the following tasks were delegated to GenAI tools under full human supervision: literature search and systematization, code generation, code optimization, text generation, proofreading and editing, summarizing text, formulation of conclusions, reformatting, quality assessment, identification of limitations, recommendations, and publication support. The GenAI tool used was: ChatGPT (OpenAI), GPT-5.6 Sol. Responsibility for the final manuscript lies entirely with the authors. For GenAI tools, authors do not bear responsibility for the final outcomes.

Funding

This work was supported by MCIN/AEI/10.13039/501100011033 under grant RYC2020-030190-I, in part by the Horizon Europe HORIZON-HLTH-2021-TOOL-06-03 under grant 101057603 and in part by

Horizon-HLTH-2022-Tool-12-Two-Stage under grant 101080875. The funders had no role in the study design, data analysis, interpretation of the results, preparation of the manuscript, or decision to submit the work for publication.

Data Availability

The MIMIC-IV (Medical Information Mart for Intensive Care IV) dataset used in this study is available through PhysioNet under credentialed access requirements, which include completion of the required training and acceptance of the applicable data-use conditions. A publicly available demonstration dataset can be accessed for reference and experimentation purposes [52]. The various prompts and complete results of the data transformations are available on GitHub [53]. The version of the code corresponding to the experiments reported in this study is identified by commit 398555133b969c97a9b82e4c891a026692d46143.

Authors' Contributions

Conceptualization: AR-T, CMC, PCDA
Data curation: AR-T
Formal analysis: AR-T, CMC
Funding acquisition: CMC
Investigation: AR-T, CMC, PCDA
Methodology: AR-T, CMC, PCDA
Software: AR-T
Supervision: CMC, PCDA
Writing – original draft: AR-T, CMC
Writing – review & editing: PCDA

Conflicts of Interest

None declared.

Multimedia Appendix 1

Release 4 to Release 5 adaptations affecting the reference mappings included in the evaluation.
[DOCX File (Microsoft Word File), 30 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Initial 119 MIMIC-IV data elements and rationale for selection in the final FHIR mapping evaluation set.
[DOCX File (Microsoft Word File), 47 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Original MIMIC-IV descriptions vs LLM-enriched descriptions.
[DOCX File (Microsoft Word File), 43 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Experiment configurations.
[DOCX File (Microsoft Word File), 31 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Prompting template.
[DOCX File (Microsoft Word File), 28 KB-Multimedia Appendix 5]

Multimedia Appendix 6

User-input template.
[DOCX File (Microsoft Word File), 28 KB-Multimedia Appendix 6]

Multimedia Appendix 7

Self-reflection prompt.
[DOCX File (Microsoft Word File), 28 KB-Multimedia Appendix 7]

Multimedia Appendix 8

MoP: Mixture of prompts.
[DOCX File (Microsoft Word File), 30 KB-Multimedia Appendix 8]

Multimedia Appendix 9

Serial prompting configurations.

[DOCX File (Microsoft Word File), 29 KB-Multimedia Appendix 9]

References

1. Vayena E. Value from health data: European opportunity to catalyse progress in digital health. Lancet. Feb 20, 2021;397(10275):652-653. [doi: 10.1016/S0140-6736(21)00203-8] [Medline: 33571452]
2. Horgan D, Hackett J, Westphalen CB, et al. Digitalisation and COVID-19: the perfect storm. Biomed Hub. 2020;5(3):1341-1363. [doi: 10.1159/000511232] [Medline: 33564668]
3. Iacob N, Simonelli F. Towards a European health data ecosystem. Eur J risk regul. Dec 2020;11(4):884-893. [doi: 10.1017/err.2020.88]
4. Hussein R, Scherdel L, Nicolet F, Martin-Sanchez F. Towards the European Health Data Space (EHDS) ecosystem: a survey research on future health data scenarios. Int J Med Inform. Feb 2023;170:104949. [doi: 10.1016/j.ijmedinf.2022.104949] [Medline: 36521422]
5. SNOMED International. URL: https://www.snomed.org [Accessed 2026-09-14]
6. ICD-11: International Classification of Diseases 11th Revision. World Health Organization. URL: https://icd.who.int [Accessed 2026-09-14]
7. LOINC. URL: https://loinc.org [Accessed 2026-09-14]
8. Bender D, Sartipi K. HL7 FHIR: an agile and restful approach to healthcare information exchange. In: Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems. 2013:326-331. [doi: 10.1109/CBMS.2013.6627810] ISBN: 978-1-4799-1053-3
9. Ahmadi N, Peng Y, Wolfien M, Zoch M, Sedlmayr M. OMOP CDM can facilitate data-driven studies for cancer prediction: a systematic review. Int J Mol Sci. Oct 5, 2022;23(19):11834. [doi: 10.3390/ijms231911834] [Medline: 36233137]
10. Kalra D, Beale T, Heard S. The openEHR Foundation. Stud Health Technol Inform. 2005;115:153-173. [Medline: 16160223]
11. Pedrera-Jiménez M, García-Barrio N, Frid S, et al. Can OpenEHR, ISO 13606, and HL7 FHIR work together? An agnostic approach for the selection and application of electronic health record standards to the next-generation health data spaces. J Med Internet Res. Dec 28, 2023;25:e48702. [doi: 10.2196/48702] [Medline: 38153779]
12. Marfoglia A, Nardini F, Arcobelli VA, Moscato S, Mellone S, Carbonaro A. Towards real-world clinical data standardization: a modular FHIR-driven transformation pipeline to enhance semantic interoperability in healthcare. Comput Biol Med. Mar 2025;187:109745. [doi: 10.1016/j.compbiomed.2025.109745] [Medline: 39904102]
13. Carbonaro A, Giorgetti L, Ridolfi L, et al. From raw data to research-ready: a FHIR-based transformation pipeline in a real-world oncology setting. Comput Biol Med. Oct 2025;197(Pt B):111051. [doi: 10.1016/j.compbiomed.2025.111051] [Medline: 40945218]
14. Pontegra software technologies BV. Ignifyr. URL: https://ignifyr.io [Accessed 2026-09-14]
15. Williams E, Kienast M, Medawar E, et al. A standardized clinical data harmonization pipeline for scalable AI application deployment (FHIR-DHP): validation and usability study. JMIR Med Inform. Mar 21, 2023;11:e43847. [doi: 10.2196/43847] [Medline: 36943344]
16. Zong N, Stone DJ, Sharma DK, et al. Modeling cancer clinical trials using HL7 FHIR to support downstream applications: a case study with colorectal cancer data. Int J Med Inform. Jan 2021;145:104308. [doi: 10.1016/j.ijmedinf.2020.104308] [Medline: 33160272]
17. Bennett AM, Ulrich H, van Damme P, Wiedekopf J, Johnson AEW. MIMIC-IV on FHIR: converting a decade of in-patient data into an exchangeable, interoperable format. J Am Med Inform Assoc. Mar 16, 2023;30(4):718-725. [doi: 10.1093/jamia/ocad002] [Medline: 36688534]
18. Stuhlmiller TJ, Rabe AJ, Rapp J, et al. A scalable method for validated data extraction from electronic health records with large language models. medRxiv. Preprint posted online on Feb 26, 2025. [doi: 10.1101/2025.02.25.25322898]
19. Yoon D, Han C, Kim DW, et al. Redefining health care data interoperability: empirical exploration of large language models in information exchange. J Med Internet Res. May 31, 2024;26:e56614. [doi: 10.2196/56614] [Medline: 38819879]
20. Pipalia A. Comparative analysis of NLP techniques for automated matching of medical intake forms to the FHIR data schema: embedding similarity and language models [Master's thesis]. Oregon Health & Science University; 2024. [doi: 10.6083/bpxhc43715]
21. Li Y, Wang H, Yerebakan HZ, Shinagawa Y, Luo Y. FHIR-GPT enhances health interoperability with large language models. NEJM AI. Aug 2024;1(8). [doi: 10.1056/aics2300301] [Medline: 40746832]
22. Delaunay J, Girbes D, Cusido J. Evaluating the effectiveness of large language models in converting clinical data to FHIR format. Appl Sci. 2025;15(6):3379. [doi: 10.3390/app15063379]

23. Hier DB, Carrithers MA, Platt SK, Nguyen A, Giannopoulos I, Obafemi-Ajayi T. Preprocessing of physician notes by LLMs improves clinical concept extraction without information loss. Information. 2025;16(6):446. [doi: 10.3390/info16060446]
24. Schmiedmayer P, Rao A, Zagar P, et al. LLMonFHIR: a physician-validated, large language model-based mobile application for querying patient electronic health data. JACC Adv. Jun 2025;4(6 Pt 1):101780. [doi: 10.1016/j.jacadv.2025.101780] [Medline: 40373519]
25. Johnson AEW, Bulgarelli L, Shen L, et al. MIMIC-IV, a freely accessible electronic health record dataset. Sci Data. Jan 3, 2023;10(1):1. [doi: 10.1038/s41597-022-01899-x] [Medline: 36596836]
26. Research ethics committee [Webpage in Spanish]. University of Murcia. URL: https://www.um.es/web/comision-etica-investigacion/comision-etica/ceic/formularios [Accessed 2026-09-16]
27. Responsible use of MIMIC data with online services like GPT. PhysioNet. URL: https://physionet.org/news/post/gpt-responsible-use [Accessed 2026-09-14]
28. Singha A, Cambronero J, Gulwani S, Le V, Parnin C. Tabular representation, noisy operators, and impacts on table structure understanding tasks in LLms. NeurIPS. URL: https://neurips.cc/virtual/2023/81303 [Accessed 2026-10-08]
29. Reimers N, Gurevych I. Sentence-BERT: sentence embeddings using siamese BERT-networks. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019:3982-3992. URL: https://www.aclweb.org/anthology/D19-1 [Accessed 2026-10-08] [doi: 10.18653/v1/D19-1410]
30. Balachandran A. abhinand/MedEmbed-large-v0.1. Hugging Face. 2024. URL: https://huggingface.co/abhinand/MedEmbed-large-v0.1 [Accessed 2026-09-14]
31. NeuML/pubmedbert-base-embeddings. Hugging Face. 2023. URL: https://huggingface.co/NeuML/pubmedbert-base-embeddings [Accessed 2026-09-14]
32. Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics. Feb 15, 2020;36(4):1234-1240. [doi: 10.1093/bioinformatics/btz682] [Medline: 31501885]
33. EmbeddingGemma model card. Google AI for Developers. 2025. URL: https://ai.google.dev/gemma/docs/embeddinggemma/model_card [Accessed 2026-09-14]
34. Lokeshch19/modernpubmedbert. Hugging Face. 2025. URL: https://huggingface.co/lokeshch19/ModernPubMedBERT [Accessed 2026-09-14]
35. Chong B. K-means clustering algorithm: a brief review. Acad J Comput Inf Sci. 2021;4(5):37-40. [doi: 10.25236/AJCIS.2021.040506]
36. Müllner D. Fastcluster: fast hierarchical, agglomerative clustering routines for R and python. J Stat Softw. 2013;53(9):1-18. [doi: 10.18637/jss.v053.i09]
37. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. KDD'96: Proceedings of the Second International Conference on Knowledge Discovery and Data Mining. 1996:226-231. [doi: 10.5555/3001460.3001507]
38. Zhang T, Ramakrishnan R, Livny M. BIRCH: an efficient data clustering method for very large databases. ACM SIGMOD Rec. 1996;25(2):103-114. [doi: 10.1145/235968.233324]
39. Ankerst M, Breunig MM, Kriegel HP, Sander J. OPTICS: ordering points to identify the clustering structure. ACM SIGMOD Record. ;28(2):49-60. [doi: 10.1145/304182.304187]
40. von Luxburg U. A tutorial on spectral clustering. Stat Comput. Dec 2007;17(4):395-416. [doi: 10.1007/s11222-007-9033-z]
41. Rousseeuw PJ. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J Comput Appl Math. Nov 1987;20:53-65. [doi: 10.1016/0377-0427(87)90125-7]
42. Caliński T, Harabasz J. A dendrite method for cluster analysis. Communications in Statistics. Jan 1974;3(1):1-27. [doi: 10.1080/03610927408827101]
43. Davies DL, Bouldin DW. A cluster separation measure. IEEE Trans Pattern Anal Mach Intell. Feb 1979;1(2):224-227. [doi: 10.1109/TPAMI.1979.4766909] [Medline: 21868852]
44. Salton G, Buckley C. Term-weighting approaches in automatic text retrieval. Inf Process Manag. Jan 1988;24(5):513-523. [doi: 10.1016/0306-4573(88)90021-0]
45. Robertson S, Zaragoza H. The probabilistic relevance framework: BM25 and beyond. Found Trends Inf Retr. Sep 17, 2009;4(1-2):1-174. [doi: 10.1561/1500000019]
46. Cer D, Yang Y, Kong SY. Universal sentence encoder for English. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2018:169-174. URL: http://aclweb.org/anthology/D18-2 [Accessed 2026-10-08] [doi: 10.18653/v1/D18-2029]

47. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. Adv Neural Inf Process Syst. 2013;26:3111-3119.
48. Gunawan D, Sembiring CA, Budiman MA. The implementation of cosine similarity to calculate text relevance between two documents. J Phys: Conf Ser. Mar 1, 2018;978(1):012120. [doi: 10.1088/1742-6596/978/1/012120]
49. Cormack GV, Clarke CLA, Buettcher S. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In: SIGIR '09: The 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval. 2009:758-759. [doi: 10.1145/1571941.1572114] ISBN: 978-1-60558-483-6
50. GPT-4o system card. OpenAI. 2024. URL: https://openai.com/index/gpt-4o-system-card [Accessed 2026-09-14]
51. meta-llama/Llama-3.3-70B-Instruct. Hugging Face. 2024. URL: https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct [Accessed 2026-09-14]
52. Johnson AEW, Bulgarelli L, Pollard TJ, Horng S, Celi LA, Demo M. MIMIC-IV clinical database demo. version 2.2. PhysioNet. 2023. URL: https://physionet.org/content/mimic-iv-demo/2.2 [Accessed 2026-10-08]
53. alvumu / InteroperabilityFHRIAPI. GitHub. URL: https://github.com/alvumu/InteroperabilityFHIRAPI [Accessed 2026-09-22]

Abbreviations

BIRCH: Balanced Iterative Reducing and Clustering using Hierarchies
CDM: common data model
DBSCAN: Density-Based Spatial Clustering of Applications with Noise
EHDS: European Health Data Space
EHR: electronic health record
ETL: extract, transform, and load
EU: European Union
FHIR: Fast Healthcare Interoperability Resources
HL7: Health Level Seven
ICD: International Classification of Diseases
ICU: intensive care unit
LLM: large language model
LOINC: Logical Observation Identifiers Names and Codes
MIMIC-IV: Medical Information Mart for Intensive Care IV
MoP: mixture of prompts
OMOP: Observational Medical Outcomes Partnership
OPTICS: Ordering Points To Identify the Clustering Structure
RQ: research question
SNOMED CT: Systematized Nomenclature of Medicine Clinical Terms

Edited by Ivan Steenstra; peer-reviewed by Gunnar Piho, Wael Zakaria, Yiqing Wang; submitted 05.May.2026; final revised version received 25.Sep.2026; accepted 25.Sep.2026; published 13.Nov.2026

Please cite as:
Riquelme-Tornel Á, Costa del Amo PL, Martínez-Costa C
Evaluating Large Language Models for Standardizing Clinical Data Using Health Level Seven Fast Healthcare Interoperability Resources: Development and Evaluation Study
JMIR Med Inform 2026;14:e100368
URL: https://medinform.jmir.org/2026/1/e100368
doi: 10.2196/100368

© Álvaro Riquelme-Tornel, Pedro Luis Costa del Amo, Catalina Martinez-Costa. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 13.Nov.2026. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.