ARKE: An Ontology-Driven Framework for Standardizing Radiology Procedure Terminology Using LLMs and RAG

<Sumin Lee[1,2*]>, < Kyulee Jeon[1,2*]>, < Yiju Park[1,2]>, <Min Seong Kim[3]>, <Juhyeon Jin[1]>, <Changhoon Han[1]>, <Soonho Yoon[4]>,<Seng Chan You[1,2†]>

[1]Department of Biomedical Systems Informatics, Yonsei University College of Medicine [2]Institute for Innovation in Digital Healthcare, Yonsei University, Seoul, Korea [3]Department of Material Sciences and Engineering, Yonsei University College of Engineering [4]Department of Radiology, Seoul National University College of Medicine, Seoul, South Korea

Background

Standardizing medical terminologies is a prerequisite for scalable observational research using the OMOP Common Data Model (CDM) [1]. Radiology remains particularly fragmented due to institution-specific procedure codes that embed modality, anatomy, and contrast details in highly variable formats [2]. This heterogeneity impedes imaging data reuse and limits phenotype portability across Observational Health Data Science and Informatics (OHDSI) networks.

General-purpose vocabularies like SNOMED CT and billing systems (e.g., CPT-4, Korea's EDI) offer broad coverage, but lack the attribute-level granularity required for radiology-specific phenotyping [3]. In contrast, the LOINC/RSNA Radiology Playbook provides a structured ontology by linking LOINC codes with RadLex attributes across 18 semantic dimensions, including modality, body region, and imaging technique [4, 5]. Despite this, mapping local terms to this schema remains a manual, resource-intensive task that does not scale well across institutions.

Prior automated approaches—such as Term Frequency-Inverse Document Frequency (TF-IDF) scoring, rule-based heuristics, and embedding-based similarity—have shown limited effectiveness in capturing the ontological structure and multi-attribute semantics of radiology procedures [6]. Moreover, unconstrained large language models (LLMs) are prone to hallucinations and semantic drift in clinical contexts, further limiting their reliability [7, 8].

To address these limitations, we introduce ARKE (Automated Radiology Knowledge Encoding)—an ontology-driven framework that integrates knowledge graph-based rule retrieval, constrained LLM prompting, and retrieval-augmented generation (RAG) to automate high-fidelity mapping of local imaging procedure terms to standardized LOINC-RadLex codes.

Methods

For this study, we collected 2,126 imaging procedure terms from a tertiary hospital in South Korea. These local terms were split into training (1,822 cases) and validation (304 cases) sets. The training set was used for developing ARKE to ensure that each local term was mapped to the most optimal LOINC-RadLex code, as shown in Table 1, while the validation set was utilized for optimization and evaluation.

As illustrated in Figure 1, the ARKE framework consists of four sequential steps. First, we constructed a knowledge graph based on the February 2025 release of the LOINC/RSNA Radiology Playbook. Each LOINC-RadLex code was represented as a node, while its associated "Parts" served as edges. These Parts encode 18 key semantic attributes—such as modality, anatomical region, contrast usage, laterality, and imaging view—forming a structured ontology to support downstream reasoning and matching. Second, local imaging procedure terms were translated into English, cleaned (e.g., symbol removal), and converted into RadLex PartType-based structured JSON using the GPT-4o-2024-08-06 model. The prompting was designed to extract attribute values conforming to the Playbook schema, enabling consistent attribute-level alignment between local terms and standard codes. Third, for each structured local term, we retrieved the top 10 candidate LOINC-RadLex codes using a graph-based rule-matching algorithm. Similarity was computed using four complementary methods: Jaccard similarity, F1 score, simple overlap (i.e., matched vs. unmatched attribute count), and a weighted matching function that emphasized high-importance attributes. Fourth, the top 10 candidates were provided to the LLM through a RAG pipeline, where they were explicitly presented as contextual input. The model then applied a chain-of-thought prompting strategy to compare attribute-level consistency across candidates and select the most appropriate final code.

For evaluation, four reviewers independently mapped the 304 cases in the validation set and constructed a reference through cross-checking. During this process, 67 non-procedural or ambiguous terms were excluded, resulting in 237 valid cases used for performance assessment. The ARKE framework was evaluated against this reference set using four metrics: accuracy, hit rate, mean reciprocal rank (MRR), and normalized discounted cumulative gain at rank 10 (NDCG@10).

Table 1. Example of radiology procedure codes standardization

Local Code term	**EDI Code term**	**SNOMED Code term**	**LOINC-RadLex Code term**		
			LOINC Code term	**RadLex PartType**	**RadLex Part term**
CT Research Liver – LBW (contrast)	RC4018 [Abdominal CT (contrast)]	429862006 [CT of liver with contrast]	24815-3 [CT Liver W contrast IV]	Modality	CT
				Anatomic location	Abdomen
					Liver
				Pharmaceutical	Contrast
					IV
				Timing	W
CT Abdomen + Pelvis (contrast)		419394006 [CT of abdomen and pelvis]	36813-4 [CT Abdomen and Pelvis W contrast IV]	Modality	CT
				Anatomic location	Abdomen
					Pelvis
CT Research liver + pelvis_RPP (contrast)				Pharmaceutical	Contrast
					IV
				Timing	W

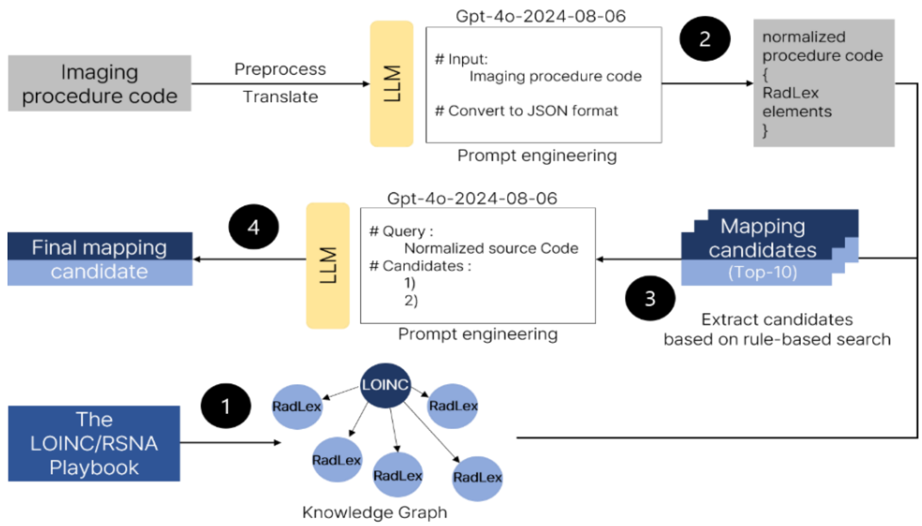

Figure 1. Workflow of ARKE framework

Results

The ARKE framework successfully generated one-to-one mappings for all 237 cases. Table 2 summarizes the performance of ARKE under various similarity-based retrieval methods. Across all metrics, LOINC-RadLex based mapping consistently outperformed direct LOINC mapping using cosine similarity. Retrieval performance improved as more candidates were considered; top-10 retrieval generally outperformed top-1 selection. Notably, the weighted match method achieved the highest overall results, including a Top-5 hit rate of 0.6835 and NDCG@10 of 0.6123. The final candidate selected after prompting achieved an accuracy of 0.6245, with both the simple overlap and weighted match methods yielding the best performance.

Table 2. Performance of ARKE framework

Method		Final candidate	Top-1	Top-3	Top-5	Top-10		
		Accuracy	**Accuracy**	**Hit Rate**	**Hit Rate**	**Hit Rate**	**MRR**	**NDCG@10**
Direct LOINC mapping	Cosine similarity	0.5359	0.2785	0.5021	0.5945	0.7511	0.4170	0.4964
LOINC-RadLex mapping	Jaccard similarity	0.6160	0.4641	0.6245	0.6624	0.7679	0.5560	0.6065
	Simple overlap	0.6245	0.4810	0.6118	0.6624	0.7553	0.5628	0.6086
	Weighted match	0.6245	0.4768	0.6203	0.6835	0.7722	0.5626	0.6123
	F1-score	0.6203	0.4641	0.6245	0.6624	0.7679	0.5560	0.6065

In several instances, the final candidate selected by ARKE demonstrated higher semantic specificity compared to the reference. For example, the input term "US Preop Kidney Doppler for Renal mass target Biopsy" was mapped to "US Guidance for biopsy of Kidney," which more precisely reflects the procedural intent than the reference term "US.doppler Renal vessels." Similarly, for "GYN US Routine General Doppler Limited," ARKE selected "US.doppler Pelvis vessels limited," while the reference remained at the broader category "US.doppler Abdomen and Pelvis," without indicating anatomical specificity or scope limitation.

Conclusion

In this study, we developed and evaluated ARKE, a novel framework that integrates ontology-guided retrieval, structured LLM prompting, and RAG to automate the standardization of radiology procedure terms. By structuring local terms into attribute-rich representations aligned with the LOINC-RadLex ontology, ARKE enables precise and scalable terminology mapping with minimal manual intervention.

The structured approach adopted in ARKE may help inform future efforts to integrate imaging procedure data into the OMOP CDM, supporting more consistent phenotyping and cross-institutional data reuse. While initial results against a silver-standard reference were promising, further validation using curated gold standards and multi-institutional datasets is necessary to assess generalizability across diverse coding practices and clinical environments.

References

1. Hripcsak G, Ryan PB, Duke JD, Shah NH, Park RW, Huser V, et al. Characterizing treatment pathways at scale using the OHDSI network. Proceedings of the National Academy of Sciences. 2016;113(27):7329-36.
2. Wang KC, Patel JB, Vyas B, Toland M, Collins B, Vreeman DJ, et al. Use of radiology procedure codes in health care: the need for standardization and structure. Radiographics. 2017;37(4):1099-110.
3. Vreeman DJ, McDonald CJ, editors. Automated mapping of local radiology terms to LOINC. AMIA Annual Symposium Proceedings; 2005.
4. Chepelev LL, Kwan D, Kahn CE, Filice RW, Wang KC. Ontologies in the New Computational Age of Radiology: RadLex for Semantics and Interoperability in Imaging Workflows. RadioGraphics. 2023;43(3):e220098.
5. Vreeman DJ, Abhyankar S, Wang KC, Carr C, Collins B, Rubin DL, et al. The LOINC RSNA radiology playbook-a unified terminology for radiology procedures. Journal of the American Medical Informatics Association. 2018;25(7):885-93.
6. OHDSI community. Usagi tool. [online]. Accessed 20 Jun 2025. Available from: https://www.ohdsi.org/analytic-tools/usagi/
7. Pan S, Luo L, Wang Y, Chen C, Wang J, Wu X. Unifying large language models and knowledge graphs: A roadmap. IEEE Transactions on Knowledge and Data Engineering. 2024;36(7):3580-99.
8. Wang J, Hu X, Hou W, Chen H, Zheng R, Wang Y, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. arXiv preprint arXiv:230212095. 2023.