A Clinical SKOS Ontology and Evaluation Benchmark for LLM Query Generation over ICU Knowledge Graphs

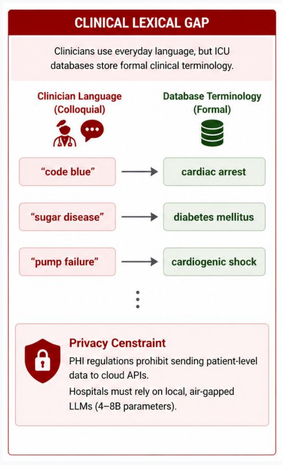
Khurrum Ali
Master of Science in Computer Science
Luddy School of Informatics, Computing and Engineering

01. Introduction

Clinicians often query ICU data using everyday clinical language such as **“code blue”**, **“sugar disease”** or **“pump failure”**, while databases store formal terms such as cardiac arrest, diabetes mellitus, and cardiogenic shock.

LLMs can help bridge this lexical gap, but clinical privacy constraints often require local air-gapped LLMs rather than cloud models. This creates a challenge: can small local LLMs generate ontology-grounded clinical queries reliably?

This work evaluates that question using:
1. ClinSKOS-ICU
2. A 421-concept SKOS ontology
3. ClinNLU - A benchmark for clinical query generation over an ICU knowledge graph.



02. Objective

Evaluate whether LLMs can use a SKOS ontology to translate colloquial ICU language into formal query terms, and identify whether local LLMs truly use the ontology or bypass it.

Main research question:

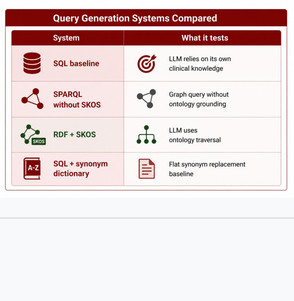
Can privacy-preserving local LLMs perform ontology-grounded clinical query generation without hardcoding formal medical terms?

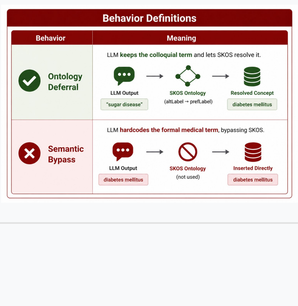
03. Methodology

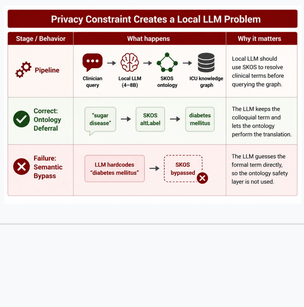
Resources

Dataset: eICU ICU cohort
Knowledge graph: Hospital, Patient, Diagnosis, OrganSystem, Drug
Ontology: ClinSKOS-ICU
- 421 ICU concepts
- 60+ curated colloquial mappings
- SKOS links using *altLabel*, *prefLabel*, *broader*, and *exactMatch*

Benchmark: ClinNLU
- 102 lexical grounding questions
- 150 semantic reasoning questions





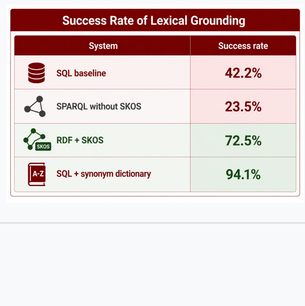


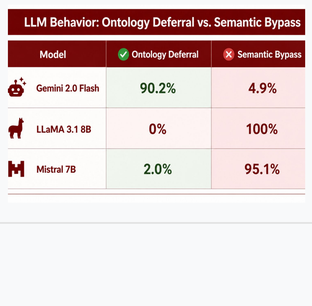
04. Analysis

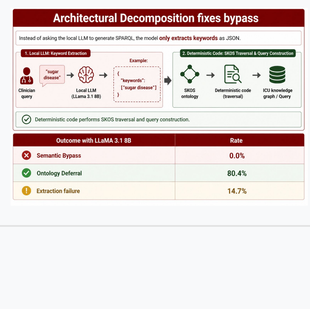
SKOS improves lexical grounding, but dictionary preprocessing scores highest on simple synonym tasks.

Interpretation:

SKOS substantially improves grounding over ungrounded SQL and SPARQL without SKOS. However, the SQL+dictionary condition performs best on simple one-to-one synonym substitution because the dictionary directly rewrites the query before generation.
The paper argues that this does not make dictionaries a replacement for ontologies, because dictionaries lack formal governance, hierarchy, interoperability, and cross-standard mapping.







05. Conclusion

This work shows that ontology-grounded clinical query generation requires more than simply giving an LLM access to a knowledge graph.

The paper makes three main contributions:

- ClinSKOS-ICU and ClinNLU
- A 421-concept ICU SKOS ontology.
- A clinical NLU benchmark for evaluating lexical grounding and semantic reasoning.
- Semantic Bypass as a new failure mode
- Local LLMs frequently hardcode formal medical terms instead of using the ontology.
- This creates an illusion of success while bypassing the intended safety mechanism.
- Architectural Decomposition as a practical solution
- Local LLMs are restricted to grammar-constrained JSON extraction.
- SKOS traversal and SPARQL generation are handled by deterministic code.
- This reduces Semantic Bypass from 100% to 0% for LLaMA 3.1 8B and achieves 80.4% ontology deferral.

Acknowledgement

I thank Dr. David Leake and Dr. Damir Cavar for their guidance on this work. I also acknowledge Indiana University Bloomington’s Luddy School of Informatics, Computing, and Engineering for computational support, and MIT / the eICU Collaborative Research Database team for making the clinical dataset available.

FIND ME :

KGLLM WORKSHOP| LREC 2026