CLIF-Query: Privacy-Preserving Natural Language Querying of CLIF Databases

Kaveri Chhikara MS[1], William F. Parker MD, PhD[1], CLIF Consortium
[1]University of Chicago, Chicago, IL

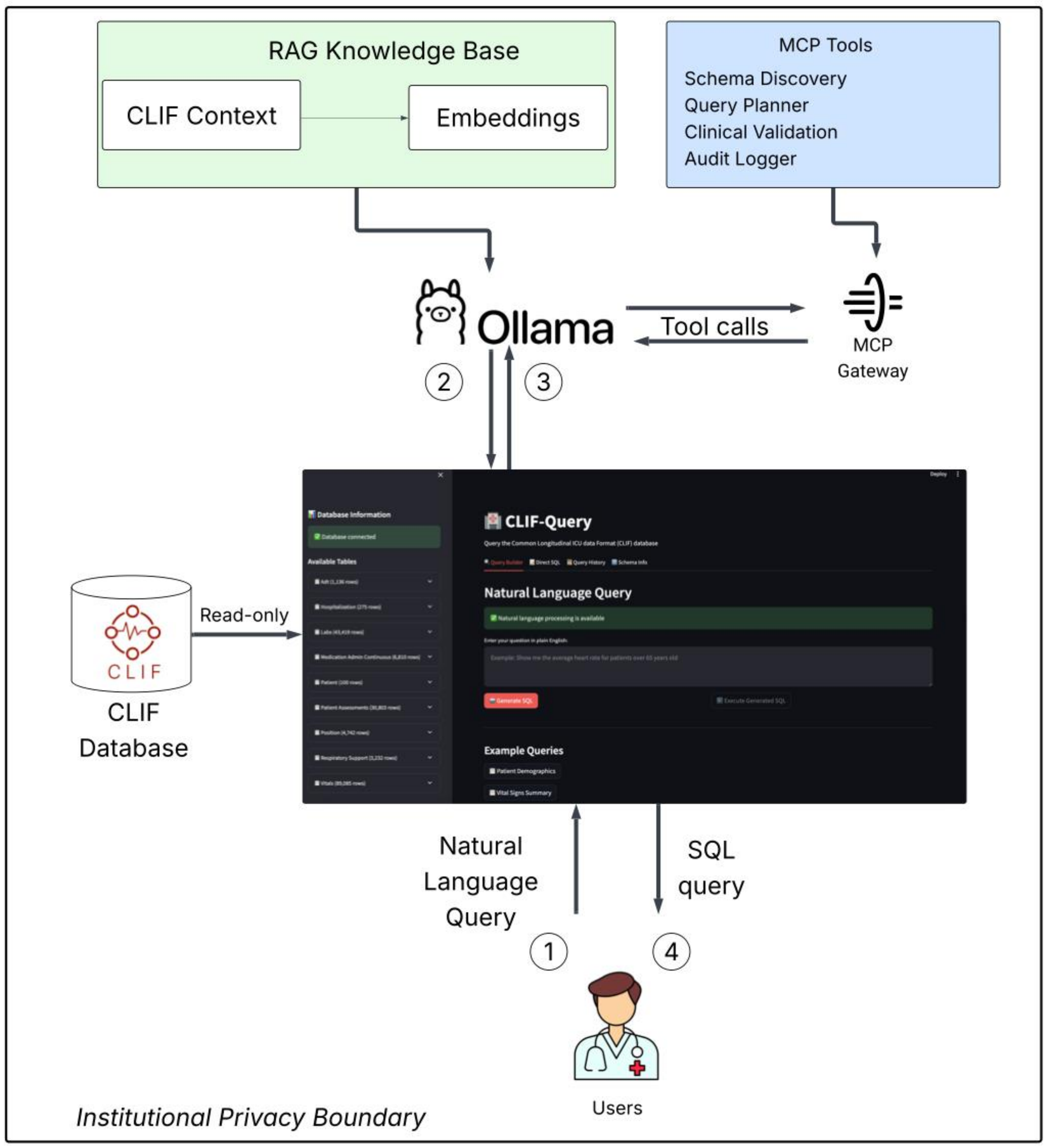
Introduction

The Common Longitudinal ICU data Format (CLIF) [1], developed by the CLIF Consortium across 17 health systems and 46 hospitals, harmonizes longitudinal EHR for more than 800,000 ICU hospitalizations across over 22 clinical domains (e.g., vital signs laboratory results, medications, respiratory support). Despite this standardization, broad adoption is limited by the SQL expertise typically required and by privacy concerns associated with cloud NLP tools. We present CLIF-Query, a locally deployed, privacy preserving natural-language interface that generates clinically faithful SQL over CLIF datasets while maintaining institutional data governance.

Architecture

CLIF-Query combines schema-aware retrieval, agentic planning, and strict execution controls.

- Knowledge Base: A local retrieval-augmented knowledge base is derived from CLIF DDLs: schema elements and relationships are embedded (SentenceTransformers all-MiniLM-L6-v2) in a local vector index and enriched with critical-care annotations (e.g., invasive mechanical ventilation, ICU locations, device settings, temporal patterns).
- Model Context Protocol: MCP agents orchestrate schema discovery, query planning with medical reasoning, and clinical validation.
- Local LLM Translation: Natural-language requests are translated to SQL by a local LLM (Ollama/Qwen2.5-Coder) and gated by a validator that enforces normalization, allow/deny lists, and parameter binding.
- Safety & Governance: Queries execute read-only in DuckDB over parquet with Arrow/Pandas post-processing; all interactions are audit-logged. No external APIs are used.



Results

- In multi-institutional testing on two consortium-relevant datasets
  - MIMIC-IV (364,627 patients; 546,028 hospitalizations; 55.5M vitals) and
  - Site A (97,254 patients; 166,814 hospitalizations; 94.4M vitals)
  - CLIF-Query correctly answered multi-step clinical questions.
- For example, “Identify patients on invasive mechanical ventilation for ≥24 hours since first ICU admission” produces SQL code joining respiratory_support (device_category='IMV') with adt (location_category='icu’) and applies correct temporal logic.
- Preliminary evaluation demonstrated 96% query-level accuracy, sub-second latency (median 7 ms), and successful coverage across 12–16 CLIF domains.

Conclusion

- This approach is **low-cost, efficient and privacy preserving**
- By grounding text-to-SQL in the CLIF schema and executing within an on-premise privacy boundary, CLIF-Query lowers the barrier to rigorous, multi-site ICU analytics across the CLIF Consortium.
- Ongoing work includes expanding domain annotations, curating gold-standard query sets, and user studies to assess clinician workload and trust.

References:

1. Rojas JC, Lyons PG, Chhikara K, et al. A common longitudinal intensive care unit data format (CLIF) for critical illness research. Intensive Care Med. 2025. doi:10.1007/s00134-025-07848-7
2. Attrach, Rafi Al, Pedro Moreira, Rajna Fani, Renato Umeton, and Leo Anthony Celi. "Conversational LLMs Simplify Secure Clinical Data Access, Understanding, and Analysis." arXiv preprint arXiv:2507.01053 (2025)