# NLP METHODS FOR EMBEDDING REAL-WORLD CLINICAL KNOWLEDGE IN LLMS: TOWARDS RESPONSIBLE, TRANSFORMATIVE, MEDICAL AI

Graciela Gonzalez-Hernandez, PhD
*Department of Computational Biomedicine, Cedars-Sinai Medical Center,*
*Los Angeles, CA, USA*
*Email: Graciela.GonzalezHernandez@csmc.edu*

Carl Berdahl, MD
*Cedars-Sinai Medical Center,*
*Los Angeles, CA, USA*
*Email: Carl.Berdahl@csmc.edu*

Brian Chapman, PhD
*UT Southwestern Medical Center,*
*Dallas, TX, USA*
*Email: brian.chapman@utsouthwestern.edu*

Wendy W. Chapman, PhD
*UT Southwestern Medical Center,*
*Dallas, TX, USA*
*Email: wendy.chapman@utsouthwestern.edu*

Sumeet Chugh, MD
*Center for Cardiac Arrest Prevention, Smidt Heart Institute, Cedars-Sinai Health System,*
*Los Angeles, CA, USA*
*Email: Sumeet.Chugh@csmc.edu*

Ian M. Campbell, MD, PhD
*Children's Hospital of Philadelphia,*
*Philadelphia, PA, USA*
*Email: campbellim@chop.edu*

Timothy Daskivich, MD
*Cedars-Sinai Medical Center,*
*Los Angeles, CA, USA*
*Email: Timothy.Daskivich@csmc.edu*

Hongfang Liu, PhD
*Dell Medical School, UT Austin,*
*Austin, TX, USA*
*Email: hongfang.liu@austin.utexas.edu*

Abeed Sarker, PhD
*Emory University,*
*Atlanta, GA, USA*
*Email: abeed@dbmi.emory.edu*

Davy Weissenbacher, PhD
*Department of Computational Biomedicine, Cedars-Sinai Medical Center,*
*Los Angeles, CA, USA*
*Email: Davy.Weissenbacher@cshs.org*

This session examines natural language processing methods for embedding, retrieving, structuring, evaluating, and governing real-world clinical knowledge in large language models. The seven accepted papers address clinical surveillance, patient-reported outcomes, longitudinal EHR reconstruction, synthetic clinical trajectories, medical search, evidence synthesis, and diagnostic reasoning. Across these settings, the papers move beyond final-answer accuracy toward explicit and auditable representations of timelines, patient states, reasoning traces, search behavior, and observability. They also highlight the complementarity of narrative notes, structured records, patient-generated information, case reports, and biomedical literature. The session identifies shared opportunities for cross-institutional benchmarks, provenance-aware evaluation, patient-centered knowledge integration, prospective human-AI assessment, and governance of updating and escalation.

Keywords: clinical NLP; large language models; medical knowledge; retrieval; clinical reasoning; evaluation; patient-centered data; responsible AI

## 1. Introduction

Large language models (LLMs) are shifting biomedical natural language processing from a field centered primarily on extraction, classification, and concept normalization toward one increasingly concerned with generation, reasoning, evidence retrieval, and interactive decision support. In medicine, realizing this potential requires methods that embed real-world clinical knowledge in LLM-based systems in ways that are explicit, evidence-grounded, current, and appropriate to the clinical context. Real-world clinical knowledge is distributed across guidelines, ontologies, electronic health record (EHR) narratives, structured observations, registries, biomedical literature, patient-reported outcomes, and patient communications.

These sources provide complementary but partial representations of clinical phenomena, patient trajectories, and care processes. They differ in authority, provenance, completeness, temporal resolution, accessibility, and relationship to a specific clinical decision. Moreover, information in the clinical record reflects not only patient states and clinical reasoning, but also documentation practices, institutional workflows, measurement processes, care fragmentation, and the circumstances under which data are collected. Embedding such knowledge in LLMs is therefore not simply a matter of supplying more text or improving answer accuracy. It requires methods for representing and linking heterogeneous evidence; retrieving, prioritizing, and updating relevant information; preserving provenance and temporal context; identifying uncertainty, missingness, and disagreement among sources; and evaluating whether model outputs, supporting evidence, and intermediate representations are clinically meaningful, well calibrated, and reliable across settings.

© 2026 The Authors. Open Access chapter published by World Scientific Publishing Company and distributed under the terms of the Creative Commons Attribution Non-Commercial (CC BY-NC) 4.0 License.

These requirements are central to developing medical LLMs that can support, rather than obscure, responsible clinical decision-making.

The PSB 2027 session, "NLP Methods for Embedding Real-World Clinical Knowledge in LLMs: Towards Responsible, Transformative Medical AI," was organized around these questions. Building on PSB's longstanding role in bringing together biomedical NLP, clinical informatics, knowledge representation, and translational biomedical research, the session sought work that treated medical knowledge as more than a static collection of facts. Our intent was to examine how clinical knowledge can be represented, grounded, updated, evaluated, and governed inside or alongside LLM-based systems. We were particularly interested in approaches that connected methodological innovation to real clinical data, patient-centered information, expert judgment, safety, fairness, and deployment constraints. The seven accepted papers collectively realize this agenda, although they approach it from different points in the knowledge pipeline.

## 2. Accepted papers

The accepted papers span clinical surveillance, patient-reported outcomes, longitudinal EHR reconstruction, synthetic clinical trajectories, medical search, evidence synthesis, and diagnostic reasoning. They illustrate two complementary directions: making clinical information more explicit and usable for downstream reasoning, and evaluating or training models to retrieve, justify, and align their answers with clinical evidence. The papers are summarized below in that order.

Weissenbacher and colleagues from Cedars-Sinai Medical Center, Cedars-Sinai Health System, and the University of Pennsylvania develop LLM-based systems for identifying in-hospital cardiac arrest events and their locations from routine EHR notes. The study compares guideline-enriched prompting, supervised fine-tuning, and self-correction against ICD-code-based extraction. Fine-tuning substantially improves performance on a curated corpus, while a two-pass correction workflow is more robust on a large, unselected validation cohort. The practical contribution is a scalable approach for prioritizing encounters for clinician adjudication and supporting cardiac-arrest surveillance and registry development[1].

Habibi and colleagues from Emory University School of Medicine introduce an observability-aware framework for evaluating language-based clinical phenotyping when the reference standard and model input come from different measurement channels. Using patient-reported depression outcomes, the study distinguishes model-recognition failure from limitations in what the analyzed clinical text can observe. The addition of patient portal messages improves sensitivity, while content-based abstention mainly enriches the positive population rather than reliably reducing selective risk. The work provides a way to interpret apparent model errors in light of missing or unrecorded patient evidence[2].

Kumar and colleagues from the National Institutes of Health and Carnegie Mellon University formulate clinical timeline reconstruction as retrieval-augmented temporal grounding. Narrative text supplies event semantics, while structured EHR rows provide partial temporal evidence from labs, medications, vitals, and procedures. Their scaffolded workflow extracts central events, builds an initial timeline, attaches additional events, and calibrates timestamps using structured data. Multimodal alignment generally improves temporal performance, although uncertainty remains

substantial and many text-derived events have no structured counterpart. The paper shows both the value and the limits of treating structured EHR data as temporal evidence[3].

Zhou and colleagues from Stanford University, the University of Texas Dell Medical School, the University of Oklahoma Health Sciences Center, and LSU Health Shreveport introduce Monopath Directed Acyclic Graphs, which convert clinical case reports into ordered patient states linked by explicit clinical transitions. Using lung-cancer case reports, they evaluate graph fidelity, embedding structure, and graph-conditioned synthetic-case generation. Physicians rate graph-conditioned cases as more temporally valid, actionable, and safe for decision-making than text-only generations, although clinical validity itself does not significantly differ. The contribution is a structured intermediate representation for preserving decision-relevant clinical trajectories, especially in rare or atypical cases[4].

Zhan and colleagues from the University of Minnesota introduce MedSR, a framework that trains compact language models to generate auditable reason-search-answer trajectories. The method combines an external search tool, LoRA adaptation, group relative policy optimization, and rule-based rewards without a separately trained reward model. Across medical question-answering benchmarks and retrieval corpora, learned search behavior can outperform untrained retrieval prompting, but the benefit depends on the model, corpus, retrieval depth, reward design, and adapter capacity. The paper treats search as a learned component of medical reasoning rather than a fixed preprocessing step[5].

Liu and colleagues from the National Institutes of Health and the University of Illinois Urbana-Champaign evaluate whether contemporary chatbots retrieve the same primary clinical studies selected by Cochrane reviewers. Across models, user roles, and repeated queries, retrieval recall varies substantially, with researcher-oriented prompts performing better than clinician- or patient-oriented prompts. The analysis also identifies a bias toward larger clinical trials, even after accounting for publication year, citations, and open-access status. The study contributes an evidence-retrieval benchmark that exposes selection and coverage bias, extending evaluation beyond citation fabrication[6].

Wu and colleagues from Stanford University, the University of Southern California, and the University of California, San Francisco introduce MedCaseReasoning, an open-access dataset of nearly 14,000 diagnostic question-answer cases derived from clinical case reports and paired with explicit reasoning statements. The authors show that current reasoning models have important deficits in both diagnostic accuracy and recall of documented reasoning, and that fine-tuning on the reasoning traces improves both measures. The contribution is an evaluation and training resource that treats faithful diagnostic reasoning, rather than only the final diagnosis, as a core dimension of medical LLM performance[7].

## 3. Discussion

Several points of convergence emerge across the papers. First, the accepted work moves toward explicit and auditable knowledge representations. Timelines, state-transition graphs, reasoning traces, search trajectories, and observability annotations expose information that would otherwise remain implicit in narrative text or latent in model parameters. These representations do not

eliminate ambiguity, but they make it possible to ask more precise questions about provenance, temporal order, missing evidence, and the relationship between a model's intermediate behavior and its final output.

Second, the papers treat clinical knowledge as heterogeneous and complementary. Narrative notes can describe phenotype and clinical meaning; structured records can anchor time; patient-reported information can reveal symptoms that were not documented by clinicians; case reports can preserve rare trajectories; and the biomedical literature can supply evidence beyond the local record. The resulting systems are strongest when they combine these sources without assuming that any one source is complete. This also means that retrieval and representation choices are inseparable from questions of access, selection, missingness, and bias.

Third, the papers collectively argue for evaluation beyond end-task accuracy. Relevant questions include whether a patient state was temporally reconstructed, whether evidence was observable in the input, whether retrieved studies align with expert selection, whether a reasoning trace reflects clinically meaningful considerations, and whether generated content supports a safe next action. Human review remains important across these settings, but the papers also show the need for structured and reproducible evaluation protocols. Performance on curated datasets can differ from performance in broad, unselected clinical populations, as illustrated by the cardiac-arrest validation cohort, the observability analysis of undocumented symptoms, and the retrieval study's dependence on user role and study characteristics.

The collection also identifies important gaps. Most studies remain single-task or single-domain evaluations, leaving open questions about transportability across institutions, populations, specialties, and documentation practices. The field lacks shared standards for representing provenance, uncertainty, temporal validity, and conflicts among sources. Patient-generated knowledge is represented in one important study, yet systematic methods for incorporating patient priorities, community narratives, and lived experience into model behavior remain underdeveloped. The session proposal also emphasized continual updating, fairness, governance, and deployment monitoring, while the accepted papers provide only an initial view of how these requirements can be operationalized. Prospective evaluation of human-AI teams, workflow effects, escalation behavior, and longitudinal model drift is a clear next step.

These convergences suggest several opportunities. Shared benchmarks could combine evidence retrieval, longitudinal reconstruction, diagnostic reasoning, observability, and safety-oriented judgment rather than treating them as isolated tasks. Such benchmarks should include multiple information channels and explicitly record what evidence was available, retrieved, omitted, or contradicted. Cross-institutional studies could test whether structured clinical knowledge representations generalize across documentation cultures and patient populations. Evaluation suites should pair automated measures with expert assessment of actionability, calibration, fairness, and safe failure. Finally, knowledge-aware systems need governance mechanisms that make updates, source provenance, model limitations, and escalation paths visible to the people responsible for deploying and using them.

## 4. Conclusion

Taken together, the seven papers mark a transition from medical LLMs that merely produce answers toward systems that can expose, retrieve, structure, justify, and be evaluated against the clinical knowledge behind those answers. The session's central insight is that responsible medical LLM development requires treating knowledge representation, evidence access, reasoning behavior, and evaluation as a connected design problem. The next phase of work will require connecting these components into robust, continuously monitored, and clinically accountable systems. By bringing together methods for knowledge representation, retrieval, reasoning, evaluation, and real-world implementation, this session aims to help define what responsible and transformative medical LLMs should know, how they should demonstrate that knowledge, and how the field should govern their use in practice.

## References

1. Weissenbacher D, Vo J, Uy-Evanado A, Reinier K, Chugh H, Vishnu K, Venkatesan M, O'Connor K, Chugh SS, Gonzalez-Hernandez G. Beyond ICD Codes: Fine-Tuning LLMs for In-Hospital Cardiac Arrest Identification from EHR Notes. In: Pacific Symposium on Biocomputing 2027.
2. Habibi S, Kavalieratos D, Parikh RB, Bozkurt S. Learning from Imperfect Patient-Reported Outcomes: An Observability-Aware Framework for Language-Based Clinical Phenotyping. In: Pacific Symposium on Biocomputing 2027.
3. Kumar S, Noroozizadeh S, Kim J, Weiss JC. Text Knows What, Tables Know When: Clinical Timeline Reconstruction via Retrieval-Augmented Multimodal Alignment. In: Pacific Symposium on Biocomputing 2027.
4. Zhou A, Fanous A, Bikia V, Deng Y, Natarajan S, Xu S, Agarwal AA, Garg A, Fanous N, Huang L, Tolbert P, Luu J, Alsentzer E, Daneshjou R. Monopath DAGs: Structured Patient Trajectories from Clinical Case Reports. In: Pacific Symposium on Biocomputing 2027.
5. Zhan Z, Wang J, Song Y, Zeng M, Yu K, Cui M, Lin M, Xie F, Zhang R. MedSR: Aligning Language Models to Reason and Search for Medical Question Answering with Group Relative Policy Optimization. In: Pacific Symposium on Biocomputing 2027.
6. Liu Q, Jin Q, Menke JD, Kahnt T, Lu Z. Do AI Chatbots Find What Experts Would? Effects of Model, User Role, and Sample Size on Study Retrieval for Medical Questions. In: Pacific Symposium on Biocomputing 2027.
7. Wu K, Wu E, Thapa R, Wei K, Zhang A, Suresh A, Tao JJ, Sun MW, Lozano A, Zou J. MedCaseReasoning: Evaluating and Learning Diagnostic Reasoning from Clinical Case Reports. In: Pacific Symposium on Biocomputing 2027.