SIIM CAIMI26 Conference on Artificial Intelligence in Medical Imaging
OCT 26-27 | University of Pennsylvania, Philadelphia, PA

A Schema-Constrained Hybrid Regex–Large Language Model Pipeline for Reliable Structured Extraction of Interventional Radiology Reports

Seema Singh Saharan, PhD, UCSF; Jhaydine Bandola; Rachel Johansson, MD; Pete Lokken, MD, MPH; Beck Olson; Rajesh Shah, MD; Alexandra So; Harsh Tripathi; Xiao Wu, MD

Introduction/Background
Narrative interventional radiology reports contain clinically important procedural information but are difficult to analyze at scale because documentation is heterogeneous and stored as free text. Large language models (LLMs) can interpret variable language, but unconstrained extraction may introduce unsupported information. We evaluated whether a schema-constrained hybrid pipeline could combine deterministic precision with selective LLM reasoning to structure trans arterial chemoembolization (TACE) reports while limiting hallucination.

Methods/Intervention
In this retrospective pilot, 26 expert-annotated TACE reports were mapped to a standardized interventional radiology reporting schema. Stage 1 used regular expressions, section-header normalization, semantic matching, and schema validation to extract diagnoses, procedural details, contrast administration, chemotherapeutic agents, vascular access, radiation metrics, and complications. Stage 2 sent only low-confidence or ambiguous fields to a healthcare-compliant GPT-4. Proposed values were accepted only after source-text verification and schema validation. Two interventional radiologists independently evaluated extracted fields against the expert reference and graded them as an exact match, an inexact match but contextually accurate, an inexact match and contextually inaccurate, data missing, or hallucinated data.

Results/Outcome
Accuracy was highest for vascular access (100%), postoperative diagnosis and complications (96.2%), compression duration (92.3%), chemotherapeutic agent (88.5%), and kerma-area product (84.6%). Accuracy was lower for closure device (73.1%), contrast volume (69.2%), and preoperative diagnosis (61.5%). The weakest fields involved workflow-dependent concepts, including sedation details, embolization position, chemotherapy dose, angiography, hemostasis technique, and access technique. The hallucination rate was 3%. These findings indicate that explicitly documented, schema-aligned fields were extracted more reliably than context-dependent procedural details.

Conclusion
A schema-constrained hybrid regex–LLM pipeline reliably extracted several well-defined elements from TACE reports while maintaining a low rate of unsupported output. Performance limitations were concentrated in fields requiring contextual or workflow-level interpretation. Larger, multi-site evaluation is needed to assess generalizability and improve semantic extraction.

Statement of Impact

Computable medical data is best when highly structured. Converting unstructured narrative procedure reports into computable structured data could support registries, quality improvement, clinical research, and outcomes assessment. Restricting LLM use to ambiguous fields and verifying proposed values against source text offers a practical design pattern for safer clinical information extraction.

Supplemental Documents

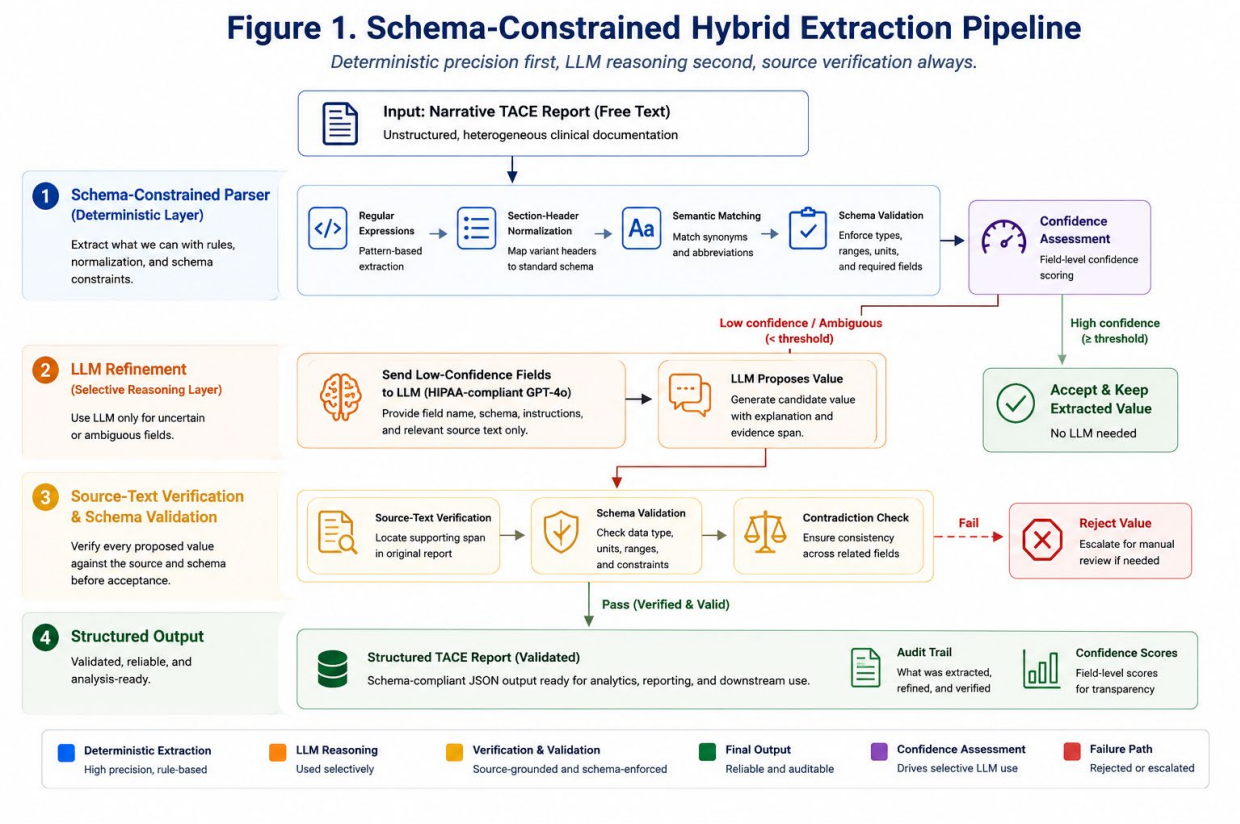

Figure 1. Trustworthy Extraction: Overview of the schema-constrained hybrid extraction pipeline. High-confidence fields are extracted deterministically, whereas only ambiguous fields are refined by an LLM and subsequently verified against the source text and schema before producing validated structured outputs.

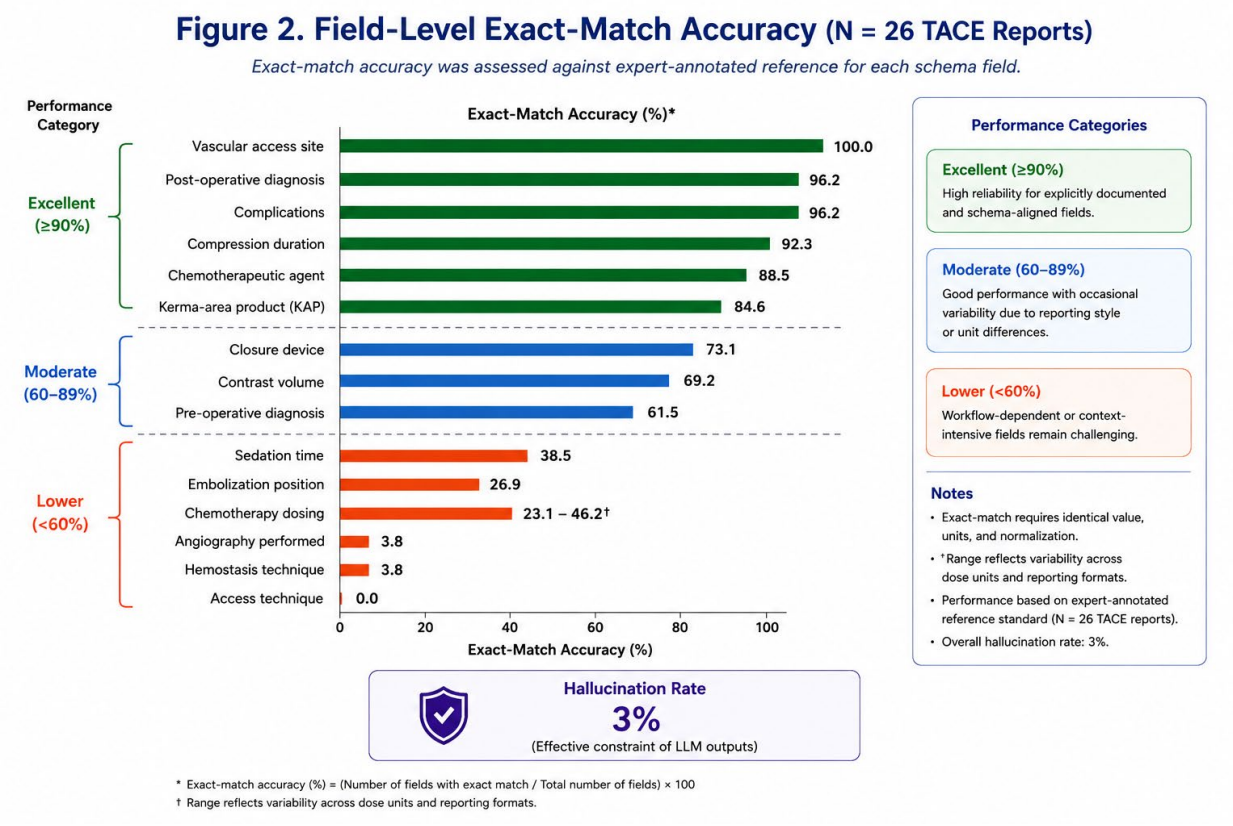

Figure 2. Extraction Accuracy: Exact-match accuracy by schema field for automated extraction from TACE reports. Accuracy was assessed against expert annotations across 26 reports. Most explicitly documented fields achieved high accuracy, while context-dependent procedural details showed lower performance. The overall hallucination rate was 3%.

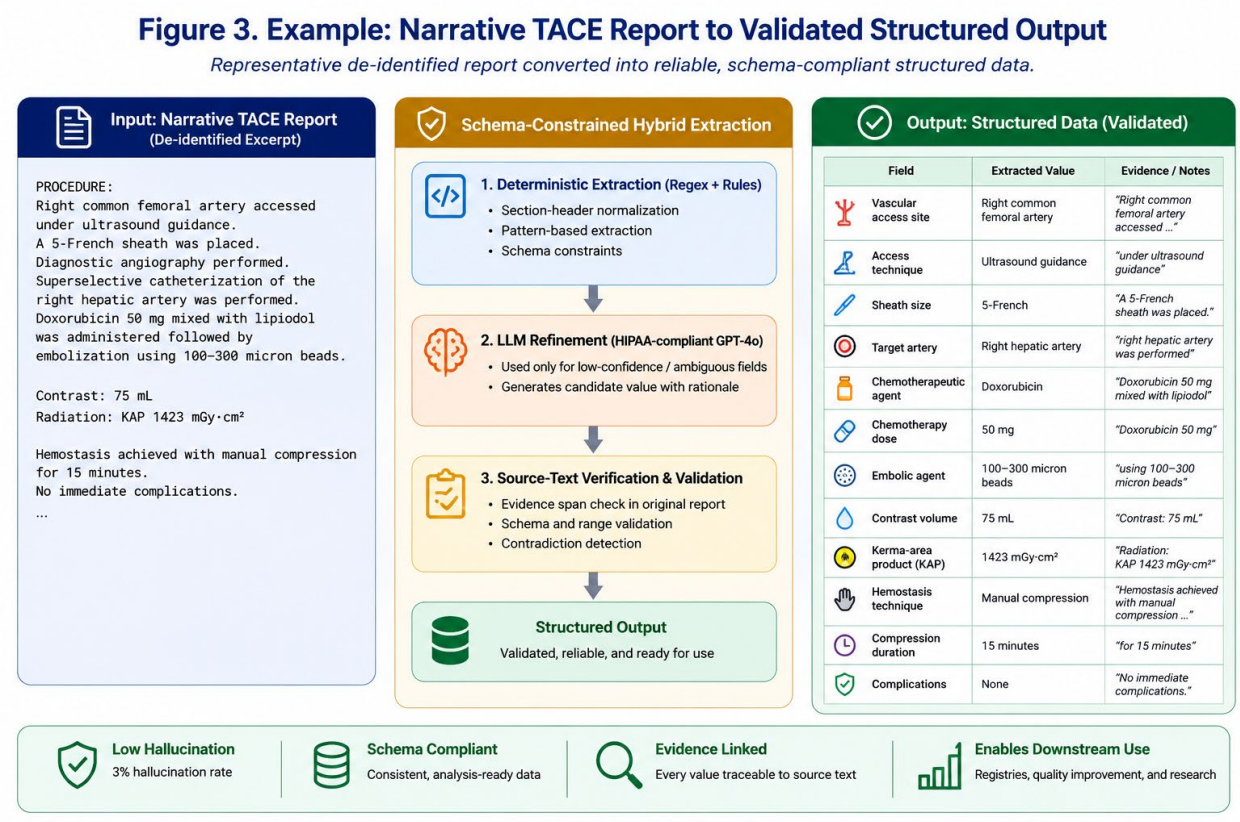

Figure 3. Structured Workflow: Representative workflow illustrating conversion of a narrative TACE procedure report into validated structured data. The schema-constrained hybrid pipeline first performs deterministic extraction of high-confidence fields, followed by selective LLM refinement of ambiguous elements and source-text verification with schema and consistency checks. The resulting structured output is evidence-linked, traceable to the original report, and suitable for downstream clinical registries, quality improvement initiatives, and research applications while reducing the risk of unsupported or hallucinated extractions.

Keywords

Interventional radiology; large language models; natural language processing; information extraction; structured reporting; clinical informatics