SIIM CAIMI25
Conference on Artificial Intelligence in Medical Imaging
OCT 20-21 | UNIVERSITY OF CALIFORNIA, IRVINE

Leveraging Large Language Models to create Medical Research Databases: A Feasibility Study, Case Example, and Open-source App

Luke Altnether, UT Southwestern; Anish Goel; Ivan Pedrosa, MD, PhD; Yin Xi, PhD

Introduction/Background
The use of large language models (LLMs) to extract data from radiology reports has led to gains in efficiency compared to manual extraction. This process requires significant programming expertise, lacks a consistent methodology, and thus has little adoption in creating research databases. The purpose of this study was to develop an effective open-source pipeline for database creation that is statistically accurate and faster than manual extraction.

Methods/Intervention
A five-phase pipeline was developed: JSON schema creation, validated ground truth entry, prompt engineering, random sample validation, then running the task dataset. An open-source app was created to host all phases of the pipeline. To demonstrate the pipeline, Phi-4 was used to extract small renal masses (SRMs) and their attributes (side, size, clear cell likelihood score (ccLS), and suspicion for fat-poor AML (fpAML)) from 62 randomly selected non-templated radiology reports. Extraction accuracy was compared to a minimal accuracy of 90% with a one-sided proportion test. Timed manual extraction was done on a random sample of 30 reports.

Results/Outcome
All SRMs were identified (82/82, p = .002). Side, units, ccLS, and suspected fpAML were extracted with 100% (82/82) accuracy (p = .002). Extraction accuracy of SRM size, prior size, prior date, and location were 97.6% (80/82, p = .018), 96.3% (79/82, p = .042), 93.9% (77/82, p = .091), 92.7% (76/82, p = .266), respectively. The combined error rate across all data points was 2.17% (16/738, p < .001). The error rate for omitted observations was 43.8% (7/16) versus 56.2% (9/16) for hallucinations. The average inference time for Phi-4 was 2.96s per report compared to 106s for manual extraction.

Conclusion
Our methodology for LLM-based data extraction is repeatable and facilitated large gains in efficiency while maintaining high accuracy. The pipeline is entirely open-source and locally deployable, addressing both reproducibility and privacy concerns.

Statement of Impact
This study empowers clinical researchers to efficiently create accurate databases from unstructured medical text without programming expertise. The pipeline expands the scope of feasible retrospective research projects (particularly when data is unstructured) and promotes equitable access to advanced AI tools in academic settings.

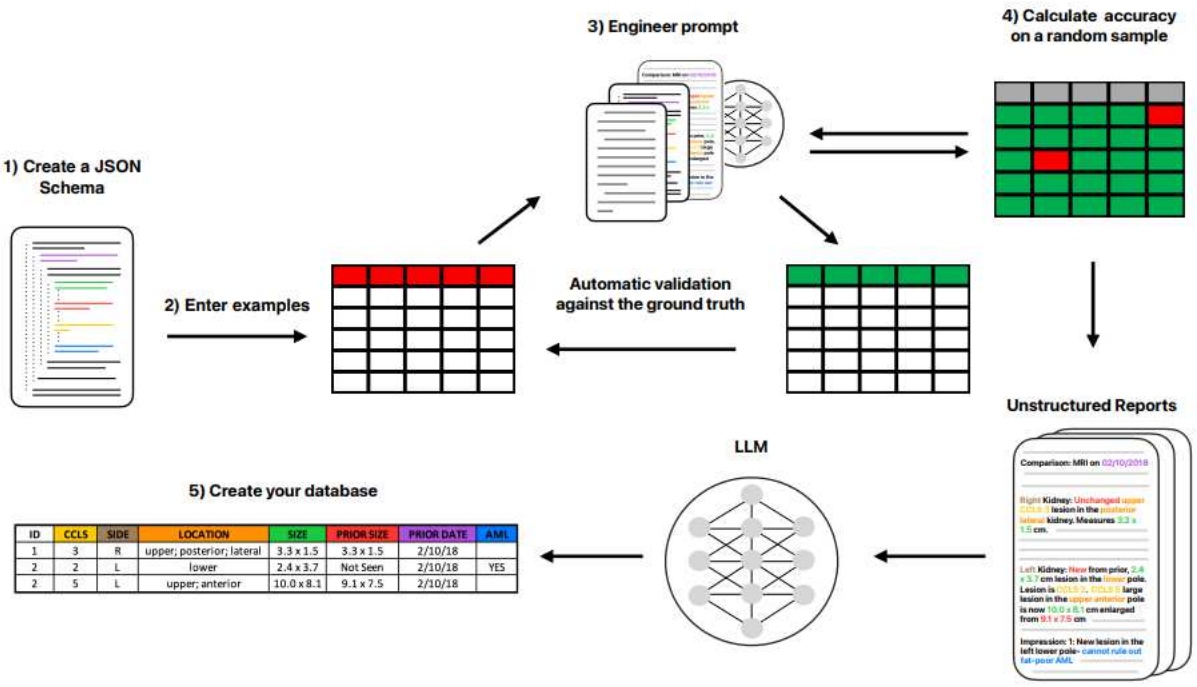

Figure 1: A schematic demonstrating the general pipeline that was developed. First a JSON schema is created. Then, representative examples are manually entered. The user can then engineer a prompt with automatic validation against the examples. Once a sufficiently high accuracy has been reached, the user can then test the prompt on a random sample to estimate the true accuracy. If this accuracy is sufficiently high, then the user can create the database.

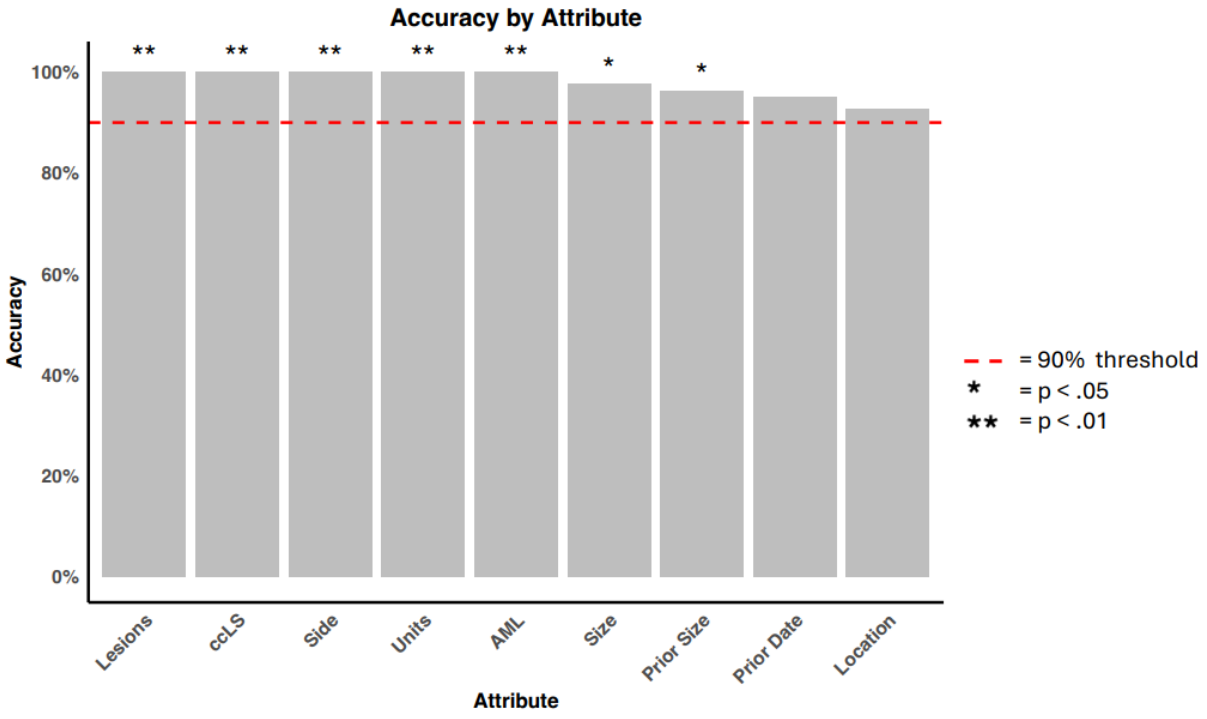

Figure 2: A bar chart representing the manually verified extraction accuracy across individual fields. The 90% accuracy threshold is denoted in red. Significance was calculated using a one sided proportion test.

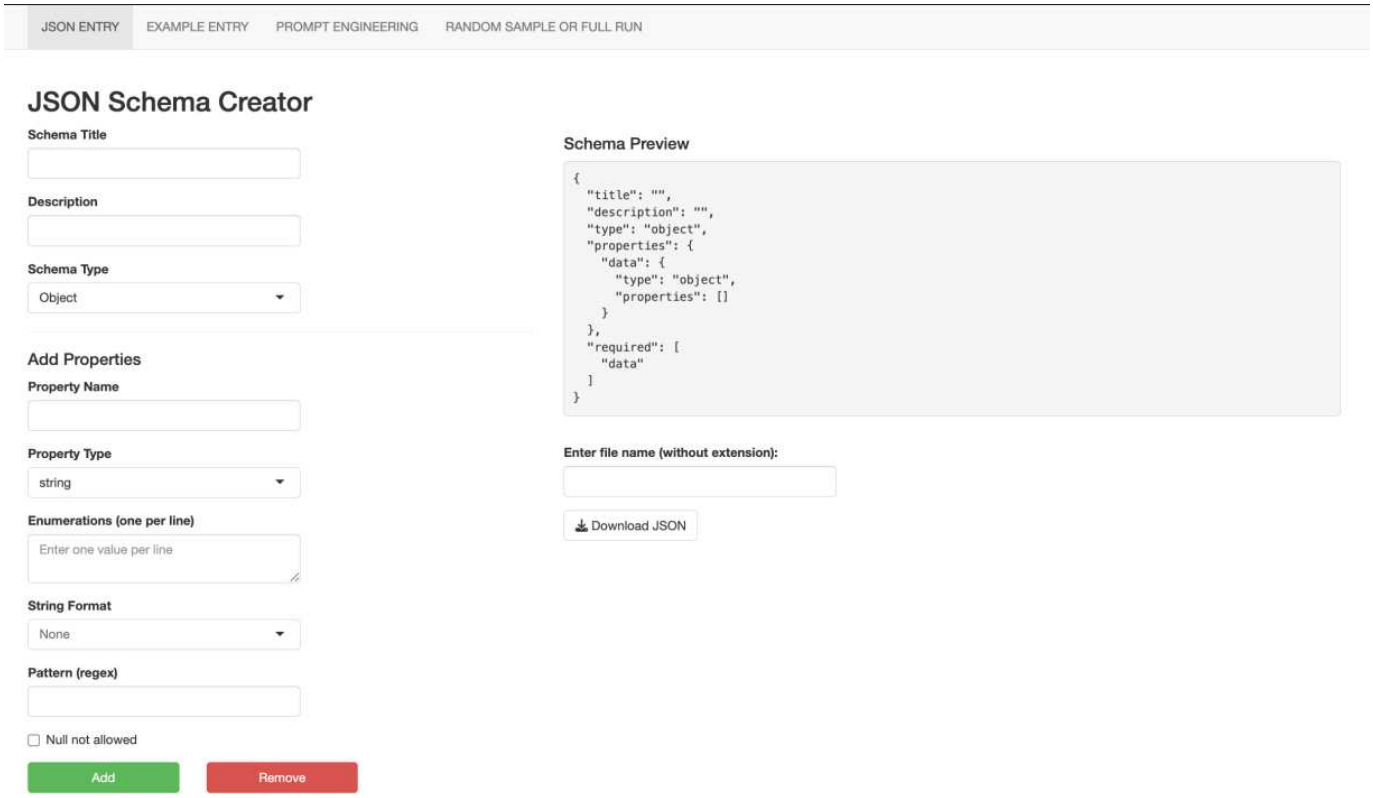

Figure 3: App user interface which facilitates the pipeline. The given panel allows for fully customizable JSON schema creation. Additional panels are denoted on the top pane but not shown in this figure.

Keywords

Artificial Intelligence; Informatics; Data Science