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

Automating NLP on Clinical CSV Files Using a Customizable LLM Workflow Tool

Shiven Velagapudi, UT Southwestern; Yee S. Ng, MD; Yin Xi, PhD

Introduction/Background

Clinical and research workflows frequently rely on structured data exported as CSV files from EHR systems. Processing these datasets often involves time-intensive manual review, inconsistent text analysis, and limited reproducibility. While general-purpose LLMs offer powerful capabilities for natural language processing (NLP), they are not readily usable at scale for structured clinical data. ProcessCSV addresses this gap by providing an intuitive web platform that enables batch processing of clinical CSV files using customizable LLM prompts.

Methods/Intervention

ProcessCSV allows users to upload structured clinical datasets and define LLM prompts with column-level precision using a $ placeholder for dynamic row-wise substitution. The system processes selected columns using remote-hosted LLMs via a REST API, appending enhanced output to the dataset. Built using Django and pandas, the platform includes a session-based interface for uploading, previewing, and batch-processing datasets of varying sizes. The system supports temperature control (0.0-1.0) for reproducible outputs and processes CSV files up to 2.5 MB with sequential row-by-row processing to ensure reliable LLM API interactions. Integration with an Ollama server enables flexible use of models (e.g., LLaMA variants), and the system can be configured for local deployment to support privacy-compliant workflows when required.

Results/Outcome

Initial testing demonstrates ProcessCSV can process over 1,000 records within 40 minutes using a single prompt, compared to estimated 40+ hours of manual effort. Users can standardize terminology, extract symptoms, or generate summaries by modifying prompt instructions. All original data is preserved, and enhanced outputs are added as new columns, supporting clean downstream analysis. Key anticipated outcomes include improved reproducibility, accelerated clinical coding, and scalable integration into research pipelines.

Conclusion

ProcessCSV automates clinical data analysis workflows by applying LLM-based enhancements directly within structured CSV files. It transforms tasks previously requiring hours of manual review into scalable, reproducible pipelines compatible with large-scale clinical research.

Statement of Impact

ProcessCSV empowers research teams and informatics professionals to unlock the potential of LLMs for structured clinical data, reducing manual burden, enabling standardization across sites, and supporting HIPAA-compliant deployment in clinical environments while maintaining institutional data sovereignty.

Keywords

Clinical Data Processing; Large Language Models; Natural Language Processing; Healthcare Informatics; Automated Workflow; Structured Data Enhancement