

EP13.31. AI-Enabled Structured Data Extraction from Endometriosis Ultrasound Reports



Jodie Avery¹, Alison Deslandes¹, Haiyi Li², Yutong Li², Yiheng Chi², Shay Freger³, Mathew Leonardi^{1,3}, Yuan Zhang^{1,4}, M. Louise Hull¹ and Hsaing-Ting Chen²

¹ Robinson Research Institute, University of Adelaide, Adelaide, Australia

² School of Computer and Mathematical Sciences, University of Adelaide, Adelaide, Australia

³ Department of Obstetrics and Gynecology, McMaster University, Hamilton, Canada

⁴ Australian Institute for Machine Learning, University of Adelaide, Adelaide, Australia

Introduction

Extraction of structured data from ultrasound reports is a critical but labour-intensive bottleneck in clinical imaging research, limiting scalability and reproducibility. Large language models (LLMs) offer the potential to rapidly transform unstructured narrative reports into analysable datasets. In this study, we evaluated the performance of three locally deployed LLMs in extracting structured data from endometriosis transvaginal ultrasound (eTVUS) reports.

Methods

- 49 eTVUS reports extracted by three LLMs: Llama3-8b, Mistral-7b and GPT-oss:20b
- All models hosted on a local server to ensure data security
- The same reports extracted by a human research assistant into the same structured spreadsheet, blinded to the LLM extraction
- An expert sonographer reviewed all extractions to assess for errors
- Individual cell coding correlated to assess for matches

Results

LLMs outperformed the human extractor on numeric and date fields. The human extractor outperformed the LLMs on text, categorical and Boolean fields (Figure 1).

98.40%

Human extractor accuracy (SD 2.13%)

86.02% GPT-oss:20b (SD 6.87%)

80.53% Llama3-8b (SD 4.58%)

78.89% Mistral-7b (SD 4.68%)

Take home message

LLM performance did not surpass human extraction, but the complementary error profiles support a hybrid human-AI workflow. LLMs can efficiently automate structured data generation and highlight inconsistencies, while human oversight ensures semantic accuracy. This approach has the potential to substantially reduce manual workload and enable scalable, high-quality dataset generation for imaging research.

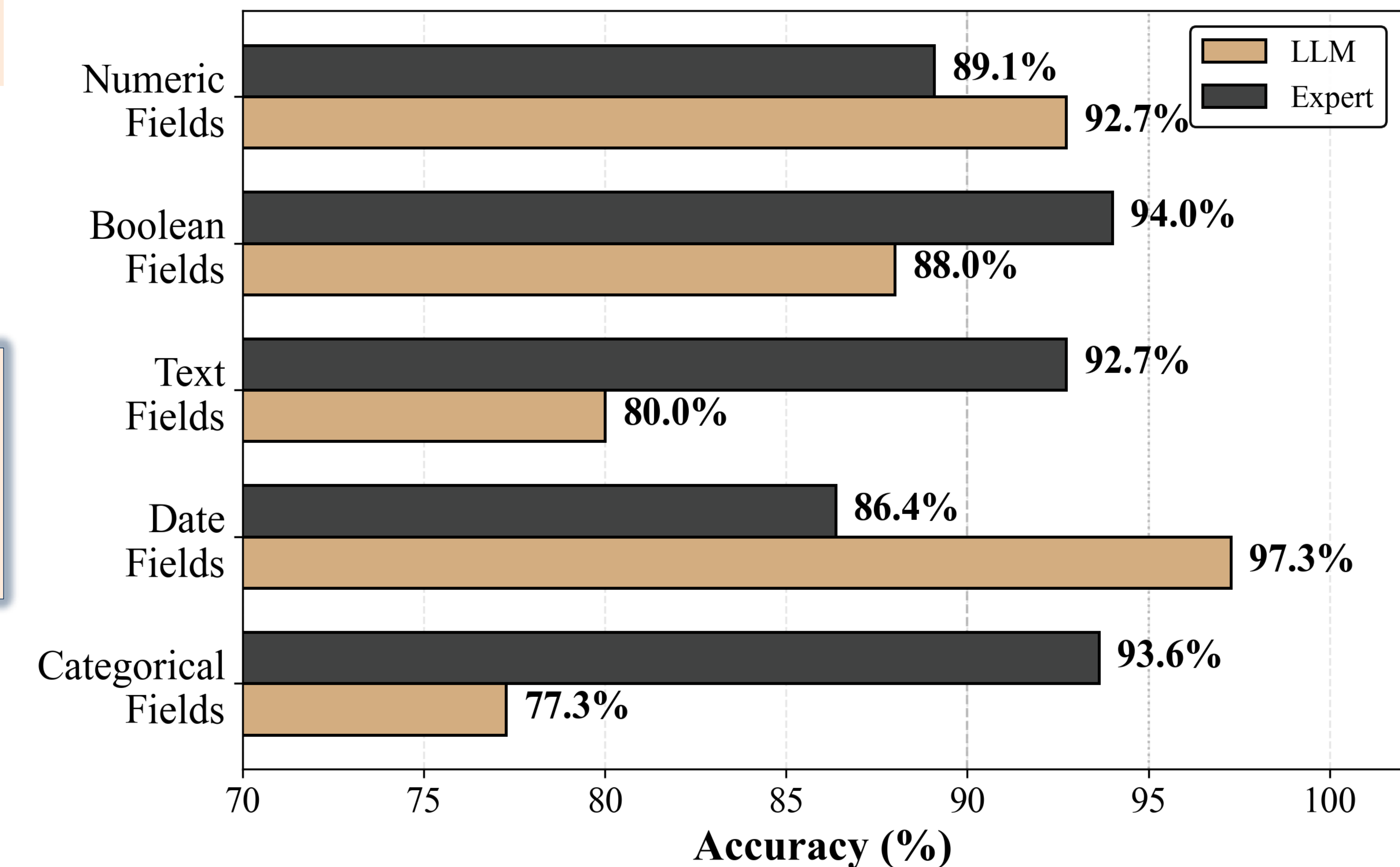


Figure 1. Performance breakdown by data field type, LLM against the human extractor.

Obstetrics &
Gynecology



Robinson
Research
Institute