From Free Text to Usable Labels: Privacy-Ensuring Open-weights LLM-Enhanced Clinical Report Extraction for Fetal MRI

Mingxuan Liu[1], Yijin Li[2], Juncheng Zhu[3], Hongjia Yang[1], Yiming Huang[4], Haoxiang Li[1], Yifei Chen[1], Xuguang Bai[1], Yi Liao[3], Haibo Qu[3], Qiyuan Tian[1]

[1] School of Biomedical Engineering, Tsinghua University
[2] School of Biological Science and Medical Engineering, Beihang University
[3] West China Second University Hospital, Sichuan University
[4] Department of Computer Science and Engineering, University of California San Diego
arktisx@foxmail.com

Purpose: Fetal Magnetic Resonance Imaging (MRI) is crucial for evaluating fetal abnormalities, but findings are often documented in unstructured, free-text reports. This format hinders large-scale data analysis, the creation of patient cohorts for research, and the development of medical artificial intelligence (AI) systems. While proprietary large language models (LLMs) like GPT can automate the extraction of structured information, their high cost and data privacy concerns limit their widespread use in clinical settings. This study aims to address this gap by developing and evaluating FetalExtract-LLM, a novel model for structured information extraction from fetal MRI reports, built by instruction-tuning an open-weights, privacy-preserving LLM for secure, in-house deployment.

Methods: The FetalExtract-LLM development pipeline comprises four stages (Figure 1). First, a dataset of free-text fetal MRI reports was retrospectively collected. Second, the reports were preprocessed to separate the “Observations” and “Impressions” sections. Third, the Qwen-3-235B-A22B model was used to generate a training corpus of structured data from the unstructured reports using a predefined JSON schema and ten curated examples for in-context learning (ICL). Subsequently, the DeepSeek-R1-0528-Qwen3-8B model was fine-tuned on this generated corpus via instruction tuning to create the final FetalExtract-LLM. The performance of FetalExtract-LLM was evaluated on a clinician-annotated test set and compared against proprietary models, including GPT-4.1, Claude Sonnet 4, and Gemini 2.5 Pro Preview.

Results: FetalExtract-LLM demonstrated high performance (Table 1), achieving an average F1-score of 0.987 in per-key matching, which is on par with the performance of leading proprietary LLMs. Additional performance metrics for our model were a

JSONable accuracy of 1.000, a domain compliance accuracy of 0.966, a positive finding accuracy of 0.829, and an exact matching accuracy of 0.770.

Conclusions: FetalExtract-LLM provides a viable, privacy-preserving, and cost-effective solution for structuring radiological data, thereby facilitating advanced research, cohort building, and the development of clinical AI tools without compromising patient data confidentiality. This approach effectively addresses the challenge of making large-scale radiological data more accessible and usable for secondary applications.

Table 1. Comparative Performance Evaluation of Different LLMs.

Models	Average F1-score	JA	DCA	PFA	EMA
Proprietary LLMs					
GPT-4.1	0.803	1.000	0.989	0.800	0.098
Claude Sonnet 4	0.965	1.000	0.996	0.900	0.702
Gemini 2.5 Pro Preview	0.949	1.000	0.996	0.936	0.502
Open-weights LLMs					
Qwen3-235B-A22B	0.954	1.000	0.955	0.836	0.732
NuExtract-2-8B	0.580	1.000	0.272	0.757	0.000
Qwen3-8B	0.672	0.898	0.049	0.771	0.004
DeepSeek-R1-0528-Qwen3-8B	0.927	0.996	0.962	0.786	0.498
Open-weights LLMs + Instruction Tuning with Simulation Data					
Qwen3-8B	0.887	1.000	0.722	0.719	0.120
DeepSeek-R1-0528-Qwen3-8B	0.945	1.000	0.693	0.655	0.121
Open-weights LLMs + Instruction Tuning with Clinical Data					
Qwen3-8B	0.987	1.000	0.966	0.806	0.754
FetalExtract-LLM	0.987	1.000	0.966	0.829	0.770

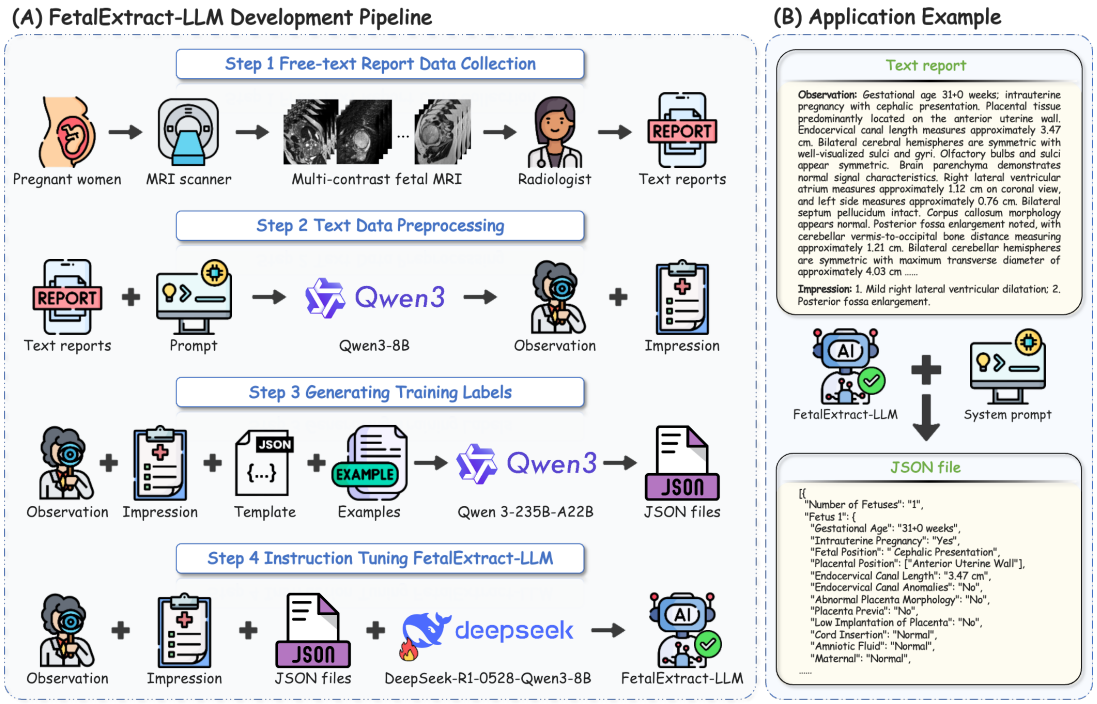

Fig. 1. (A) FetalExtract-LLM Development Pipeline. (B) Application Example.