Fine-Tuning Large Language Models on EHR Data for Early Endometriosis Diagnosis in Adolescents

Background. Endometriosis is a chronic condition affecting approximately 10% of women, with adolescents often experiencing diagnostic delays up to three times longer than adults.[1] Contributing factors include a lack of non-invasive diagnostic methods and symptom overlap with other conditions. Genetic variants may influence the risk of developing endometriosis[2-4], but integrating genomic data into clinical decision-making remains underexplored. This project aims to combine large language models (LLMs) with genetic biomarkers to develop multi-modal, fine-tuned models for early and accurate detection of endometriosis in adolescents.

Methods. We collected 7,221 clinical notes from the Mount Sinai Data Warehouse (MSDW) for 125 patients (ages 13–19) diagnosed with endometriosis. A subset of 700 notes from 26 patients was annotated by medical experts using standardized keyword-based criteria. These notes were segmented, tokenized, and encoded. The LLMs LLaMA, Gemma, GatorTron, and GPT-4o were fine-tuned for clinical natural language processing (NLP) tasks and evaluated using accuracy, precision, recall, and F1 metrics. In addition to the clinical notes, we obtained access to the Mount Sinai Million Health Discoveries Program for genetic data and are in the process of conducting a genome-wide association study (GWAS) to identify endometriosis-associated markers. These genomic features, including SNPs and polygenic risk scores (PRS), will be integrated as structured inputs alongside unstructured clinical text to improve predictive accuracy and support personalized diagnosis.

Results. Across nine clinical symptom categories, GPT-4o-chat achieved the strongest overall performance, with an average F1 of 0.486 and accuracy of 0.871. Gemma performed moderately (F1: 0.261, Acc: 0.869), while LLaMA showed comparable accuracy (0.905) but a lower F1 (0.258). GatorTron obtained the lowest F1 (0.189) despite competitive accuracy (0.887). For endometriosis classification specifically, all models achieved relatively high accuracy (>0.62), though GPT-4o-chat led with the highest F1 (0.771). These results indicate that GPT-4o-chat provided the best balance between precision and recall. However, performance varied widely by symptom, with models excelling in structured signals like pelvic tenderness but struggling on more non-specific categories such as GI symptoms. LLaMa exhibited the longest training and evaluation times.

Discussion. This study is one of the first systematic evaluations of open-source, trainable LLMs for NLP-based detection of adolescent endometriosis from unstructured EHR data. GPT-4o-chat demonstrated the strongest performance while maintaining computational efficiency, underscoring its potential scalability in resource-constrained clinical settings. By fine-tuning LLMs using EHR notes, these models can uncover subtle diagnostic patterns that are often missed in conventional workflows. Incorporating genetic information following GWAS will further enhance predictive power by enabling the models to capture both symptomatic presentation and underlying biological risk. Together, these findings highlight the promise of multi-modal frameworks to support earlier, more accurate diagnoses, guide personalized treatment strategies, and advance equitable care.

References

1. DiVasta, A. D., Vitonis, A. F., Laufer, M. R., & Missmer, S. A. (2018). Spectrum of symptoms in women diagnosed with endometriosis during adolescence vs adulthood. American journal of obstetrics and gynecology, 218(3), 324.e1–324.e11.
2. Mackenzie, S. C., et al. (2024). Genome-wide association reveals a locus in neuregulin 3 associated with gabapentin efficacy in women with chronic pelvic pain. iScience, 27(8), 110370.
3. Dimitrakov, J., & Guthrie, D. (2009). Genetics and phenotyping of urological chronic pelvic pain syndrome. The Journal of urology, 181(4), 1550–1557.
4. Li, Y. Z., & Ji, R. R. (2024). Gene therapy for chronic pain management. Cell reports. Medicine, 5(10), 101756.