

Evaluating Large Language model against Structured EHR Diagnoses: Agreement Patterns and Discrepancies

Mohan Kashyap Pargi, MTech¹, Hanieh Razzaghi, PhD¹, L. Charles Bailey MD, PhD¹
Levon Uditjian, MD, MBI¹, Nhat Nguyen, PhD¹

¹Childrens Hospital of Philadelphia, Philadelphia, PA, USA

Background

Large language models (LLMs) like GPT-OSS show promise for validating clinician-assigned diagnosis codes (structured EHR codes) against clinical notes¹, but systematic agreement patterns with different structured diagnosis codes remain unknown. We sought to address this gap by prompting GPT-OSS with paired structured EHR codes and unstructured notes across two scenarios: (1) bulk evaluation of all codes per note vs. (2) per-code validation per clinical note. We analyzed concordance patterns to identify ICD-10-CM chapter codes with strong clinician-LLM alignment and systematic divergence, providing scalable guidance for LLM integration into clinical coding workflows.

Methods

We analyzed 122 clinical notes paired with clinician-assigned ICD-10-CM codes for corresponding visits from structured EHR data using GPT-OSS. Two prompting strategies were evaluated: (1) bulk validation, where the LLM received the full note and all structured codes in a single prompt to determine overall agreement or disagreement; and (2) per code validation, where the LLM reviewed the full note but evaluated each structured code individually. The prompting pipeline is illustrated in Figure 1. Inputs included unstructured clinical notes and structured ICD-10-CM codes, while outputs represented LLM-generated agreement or disagreement assessments. A post-processing step aggregated these outcomes to quantify concordance and discordance across each individual ICD-10-CM categories.

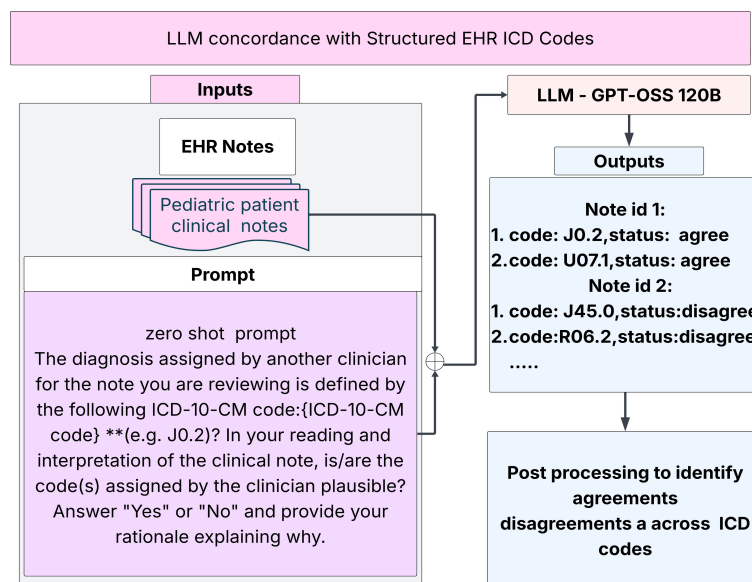


Figure 1. LLM- structured EHR concordance framework

Results

Agreement between LLM outputs and structured diagnosis codes are provided in Figure 2 for per code validation strategy, showing marked variation across chapters, ranging from 12.5% (D) to 100% (U). High agreement chapters (E,H,L,U) were dominated by well-defined conditions such as COVID-19 (U-codes), dermatologic (L), and ear/eye (H) disorders. In contrast, low-agreement chapter codes (D, Q, Z, R) reflected clinician-LLM discordance on Z-codes (family history, exposures, risk factors), rare congenital anomalies (e.g. Q23.4, Q25.0) and symptom codes (e.g R06.89, R05.9). Intermediate performance in respiratory (J = 63%) and digestive (K = 54%) chapters with examples such as asthma (J45.20) and Crohn's disease (K50.80) reveals chapter-specific variability. Overall GPT-OSS 4.0 demonstrated 31% agreement (49 Yes, 74 No) in bulk prompting compared to 52% agreement (170 Yes, 151 No) in per-code prompting. The increase in agreement with per-code prompting strategy suggests that granular validation allows the LLM to capture more nuanced concordance patterns.

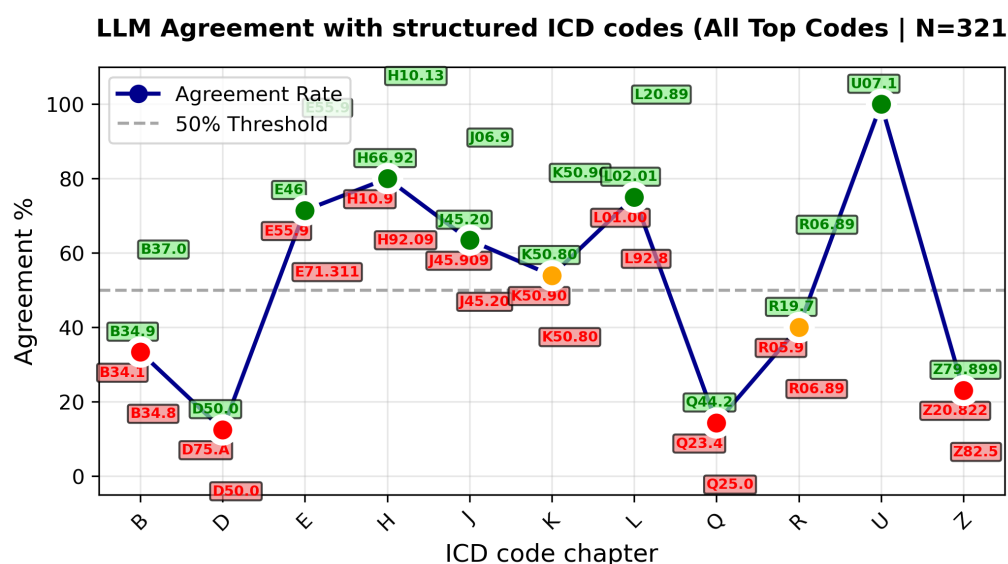


Figure 2. LLM Agreement % with structured ICD codes across ICD chapters

Conclusion

GPT-OSS 4.0 achieved over 50% overall concordance with structured data under the per-code evaluation framework. Notably, the LLM's strong performance on U-codes (COVID-19) at 100% contrasted with poor agreement on Z and Q-codes <25% suggests a preference for explicitly diagnosed conditions over symptoms, exposures, or administrative codes (often added for billing purposes). This pattern indicates LLMs may excel at validating clinical diagnoses directly supported by narrative evidence while flagging auxiliary codes for human review.

References

1. Li S, Zheng C, Wu J, Xu Q, Xu X, Wang H, Sun Y, Bai Z, Xu Y, Zhu L, Hu W. Verification is All You Need: Prompting Large Language Models for Zero-Shot Clinical Coding. IEEE Journal of Biomedical and Health Informatics. 2025 Jul 28.