ClinBench: A Standardized Multi-Domain Framework for Evaluating Large Language Models in Clinical Information Extraction

Ismael Villanueva-Miranda, Zifan Gu, Donghan M. Yang, Kuroush Nezafati, Jingwei Huang, Peifeng Ruan, Xiaowei Zhan, Guanghua Xiao, Yang Xie

Quantitative Biomedical Research Center, Peter O'Donnell Jr. School of Public Health, UT Southwestern Medical Center, Dallas, TX 75390, United States

NEURAL INFORMATION PROCESSING SYSTEMS

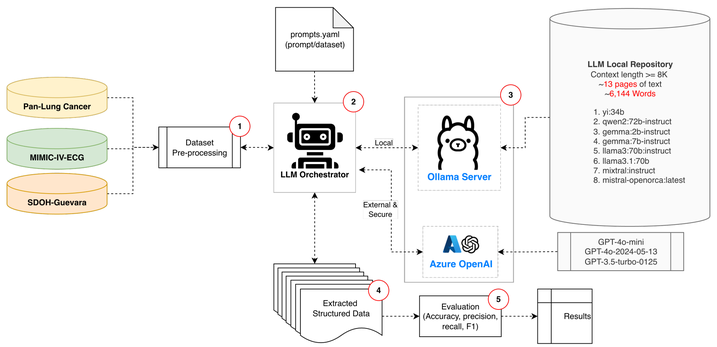
Motivation

The clinical NLP gap

Rapid LLM adoption outpaces systematic evaluation. Current benchmarks lack standardization across diverse clinical domains, terminology, and privacy constraints.

To address this gap, we introduce **ClinBench**, an open-source, multi-model, multi-domain framework for rigorous, reproducible evaluation of structured clinical information extraction.

Methodology: The ClinBench Framework



The ClinBench automated five-step benchmarking workflow.

(1) Ingestion of preprocessed clinical datasets
(2) YAML-configured, schema-guided LLM orchestrator processing
(3) Model inference (local/API), to
(4) JSON-validated data extraction
(5) Comprehensive performance evaluation

Framework Highlights:

- Standardized Input: Ingests consistent datasets
- Hybrid Orchestrator: Supports both Local (Ollama) and Secure API (Azure/OpenAI) models within the same pipeline.
- Schema Enforcement: Validates output structure via the strictjson library to ensure machine-readable data.

Methodological Rigor:

- **Reproducibility First:** All inference is performed at Temperature = 0 to ensure deterministic results.
- **Static vs. RAG:** We utilized Static Knowledge Injection via YAML rather than Retrieval-Augmented Generation (RAG). This isolates the LLM's reasoning capabilities, removing the RAG's performance as a confounding variable.

Datasets & Models Evaluated

Three Clinical Domains:

1. **Lung Cancer (TCGA):** Complex TNM staging from pathology reports.
2. **Atrial Fibrillation (MIMIC-IV-ECG):** Binary detection from ECG reports.
3. **SDOH (MIMIC-III):** Extraction of employment and housing status.

Dataset	Size	Variables
Lung Cancer (TCGA)	774	pT pN tumor_stage histologic_diagnosis
Atrial Fibrillation (MIMIC-IV-ECG)	700	AF Not AF
SDOH (MIMIC-III)	1405	**Employment** Employed, Unemployed, Unknown **Housing** Housing, Homeless, Unknown
Total	2879	12 variables

11 LLMs Benchmarked: Including proprietary (GPT-4o, GPT-3.5) and open-source (LLaMA 3.1-70b, Mixtral, Qwen2, Gemma).

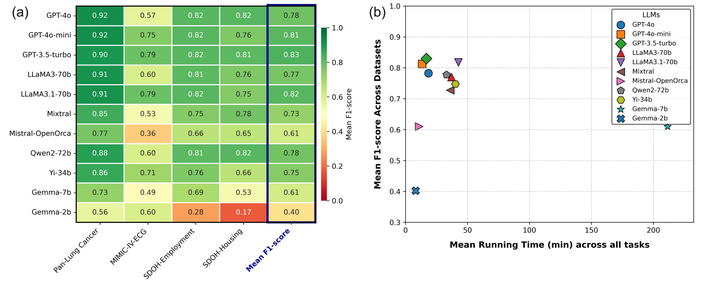
Organization	Model	Family	Size	Release Date	Context Length	Architecture
Meta (2)	Llama 3.1: 70B	Llama	70B	2024-07-23	128K	Transformer (decoder-only)
	Llama 3: 70B Instruct	Llama	70B	2024-04-18	128K	Transformer (decoder-only)
OpenAI (3)	GPT-4o Mini	GPT-4o	Not specified	2024-07-18	128K	Transformer with multi-modality
	GPT-4o (2024-05-16)	GPT-4o	Not specified	2024-05-13	128K	Multimodal Transformer
	GPT-3.5-Turbo-1109	GPT-3.5	Not specified	2023-11-06	16K	Transformer
Others (2) (Alibaba Group & 01.AI)	Qwen2: 72B Instruct	Qwen2	72B	2024-06-07	128K	Transformer (dense and MoE variants)
	Yi: 34B	Yi	34B	2023-11-02	200K	Transformer (decoder-only with RoPE)
Mistral AI (2)	Mixtral: Instruct	Mixtral	22B	2024-04-17	64K	Transformer (decoder-only, sparse MoE)
	Mistral-OpenOrca	OpenOrca	7B	2024-05-13	8K	Transformer (decoder-only)
Google DeepMind (2)	Gemma: 7B Instruct	Gemma	7B	2024-02-21	8K	Transformer (decoder-only, multi-query attention)
	Gemma: 2B Instruct	Gemma	2B	2024-02-21	8K	Transformer (decoder-only)

Evaluation Methodology

ClinBench ensures rigorous benchmarking by validating LLM outputs against expert-curated ground truth from TCGA (Lung Cancer) and MIMIC (AF, SDOH) datasets. Performance is assessed using **Weighted F1 Score as the primary metric and Runtime for efficiency**, with a specific emphasis on Sensitivity (Recall) to prioritize clinical safety in capturing positive cases.

Results

Results comparison of the 11 LLMs across four clinical information-extraction tasks. **(a)** A heatmap shows F1-scores per model and dataset, with the rightmost column summarizing the mean F1 across all tasks (greener cells indicate better performance). **(b)** A scatter plot relates each model's mean F1 to its mean running time, highlighting performance–efficiency trade-offs and showing that some models (e.g., GPT-4o-mini, GPT-3.5-turbo) achieve relatively high F1 with lower runtime.



Token Cost Analysis for API-Based Models

Dataset	Model	Prompt	Tokens Completion	Total	Costs Prompt	Costs Completion	Total Cost
SDOH	gpt-3.5-turbo-1106	753,207	33,044	786,251	$0.75	$0.07	$0.82
	gpt-4o-2024-05-13	753,207	28,506	781,713	$3.77	$0.43	$4.19
	gpt-4o-mini	753,207	32,932	786,139	$0.11	$0.02	$0.13
ECG	gpt-3.5-turbo-1106	631,568	3,869	635,437	$0.63	$0.01	$0.64
	gpt-4o-2024-05-13	631,568	3,855	635,423	$3.16	$0.06	$3.22
	gpt-4o-mini	631,568	13,322	644,890	$0.09	$0.01	$0.10
Lung	gpt-3.5-turbo-1106	1,747,615	61,349	1,808,964	$1.75	$0.12	$1.87
	gpt-4o-2024-05-13	1,747,615	61,656	1,809,271	$8.74	$0.92	$9.66
	gpt-4o-mini	1,747,615	61,539	1,809,154	$0.26	$0.04	$0.30

Ablation Study on Prompting Strategy

Removing the structured YAML/JSON constraints resulted in an important performance drop (e.g., -73% F1 for GPT-3.5), proving that our schema enforcement is essential for clinical reliability.

Model	F1 Score ClinBench	F1 Score Unstructured Prompt	Performance Drop
gpt-4o	0.92	0.46	-50%
gpt-4o-mini	0.92	0.38	-59%
gpt-3.5-turbo	0.90	0.24	-73%

UT Southwestern Medical Center
QBRC Quantitative Biomedical Research Center