MedCase-Structured: A Text-to-FHIR Dataset for Benchmarking Diagnostic Reasoning in Clinically Realistic EHR Settings

Valentina Bui Muti [1] **Eugénie Dulout** [1] **Ziquan Fu** [1]

Abstract

Large language models (LLMs) show promise for clinical reasoning and decision support, but evaluation in structured, electronic health record-congruent settings remains limited. Existing benchmarks often rely on static datasets or unstructured inputs that do not reflect the interoperable data formats used in clinical systems. We introduce a reusable pipeline for generating terminology-grounded HL7 FHIR R4 bundles from unstructured text, enabling controllable evaluation of clinical decision support systems over structured inputs. The pipeline combines staged LLM generation with terminology-grounded validation and repair to eliminate hallucinated codes and enforce structural and semantic consistency. Applying this approach to MedCaseReasoning, we construct **MedCase-Structured**, a synthetic dataset of 1,732 FHIR bundles derived from clinician-authored diagnostic cases, producing complete, valid bundles for 97.1% of attempted cases. Evaluation on **MedCase-Structured** reveals consistently lower diagnostic accuracy for LLMs on structured FHIR inputs than with plain text, highlighting the importance of deployment-aligned benchmarking.

1. Introduction

Large language models (LLMs) have demonstrated promising capabilities across a range of clinical reasoning and decision support tasks (Shool et al., 2025; Mansoor et al., 2025), motivating their use in clinical decision support systems (CDSS). The richness of patient data captured in electronic health records (EHRs) makes them a valuable input source for LLM-based CDSS. However, EHR data are heterogeneous and largely unstructured (Li et al., 2024a), making it challenging to effectively incorporate full patient context into LLM-based pipelines. As LLM-based CDSS become more prevalent, rigorous testing in structured, clinically meaningful, and deployment-aligned settings is essential.

Evaluating EHR-based CDSS tools presents two key challenges. First, real patient data are protected by strict privacy regulations, limiting access and reproducibility (Li et al., 2024a). Second, evaluation inputs must reflect the structure and standards of real clinical systems. Modern healthcare infrastructure increasingly relies on HL7's Fast Healthcare Interoperability Resources (FHIR) (HL7 International, b) for representing and exchanging patient data. While datasets such as MIMIC-IV (Johnson et al., 2023) are widely used for benchmarking clinical models (Li et al., 2024a), they are restricted to specific care settings and do not natively preserve EHR interoperability structures. Although derived representations such as MIMIC-IV-FHIR (Bennett et al., 2023) map these into FHIR format, they are retrospective transformations rather than outputs of deployed clinical systems. Additionally, it is meaningfully different to evaluate LLM based CDSS on FHIR-based inputs versus textual inputs, as understanding structured and coded information requires higher reasoning ability because of the absence of semantically continuous context that the LLMs are primarily pre-trained on. Recent work shows that both input representation and evaluation protocols significantly influence LLM performance in clinical tasks (Shool et al., 2025; Navarro et al., 2026; Yang et al., 2026), emphasizing the need for standardized, deployment-aligned benchmarks. Similarly, studies of FHIR-based systems highlight the difficulty of reasoning over structured patient data and the lack of realistic evaluation benchmarks (Lee et al., 2025).

These challenges highlight the need for publicly available synthetic clinical data that can represent patient context in EHR-congruent formats while supporting controlled evaluation. Tools such as Synthea (Walonoski et al., 2018) generate realistic patient records while bypassing privacy concerns and supporting export in FHIR-compatible formats. However, Synthea relies on predefined modules and heuristic rules, which may limit its ability to capture complex or atypical clinical scenarios and provide the fine-grained

[1]System Inc., New York, NY, USA. Correspondence to: Valentina Bui Muti <valentina@system.com>, Ziquan Fu <frank@system.com>.

Proceedings of the Workshop on Structured Data for Health at the 43rd International Conference on Machine Learning, Seoul, South Korea. Copyright 2026 by the author(s).

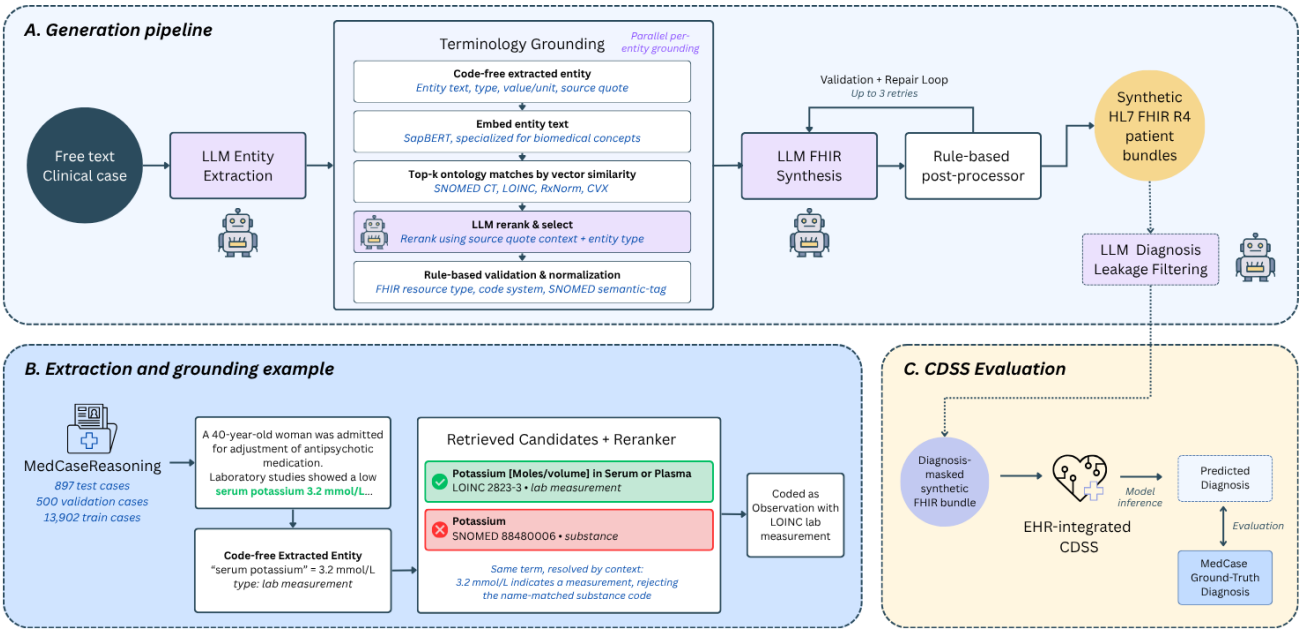

Figure 1. Overview of MedCase-Structured. (A) Free-text clinical cases are converted into terminology-grounded HL7 FHIR R4 bundles. (B) An example MedCaseReasoning (Wu et al., 2025) case shows extraction, ontology-aware grounding, and rejection of a name-matched but semantically invalid code. (C) Diagnosis-masked bundles are used for EHR-congruent CDSS evaluation against ground-truth diagnosis.

control required to stress-test model reasoning. Recent approaches using LLMs for text-to-FHIR transformation (Li et al., 2024b; Frei et al., 2026) offer improved patient-level control; however, they primarily focus on faithful reconstruction of existing clinical records rather than generating diverse evaluation datasets.

Taken together, these limitations highlight a key gap: existing approaches do not provide flexible and controllable methods for generating terminology-grounded, FHIR-formatted patient cases that can systematically evaluate model reasoning under diverse and challenging conditions.

To address this gap, we introduce a reusable pipeline for generating synthetic HL7 FHIR R4 patient bundles from unstructured text, with an emphasis on controllability and downstream evaluation. A central component of the pipeline is a terminology-grounded validation and repair step that identifies and corrects hallucinated clinical codes against standard clinical terminologies, while enforcing structural and semantic consistency across generated FHIR resources. This enables interoperable and scalable evaluation of LLM-based clinical systems.

We further introduce **MedCase-Structured**[1], a structured diagnostic reasoning dataset constructed by applying our pipeline to MedCaseReasoning (Wu et al., 2025). **MedCase-Structured** encodes each case as a terminology-validated FHIR R4 patient bundle, preserving the diagnostic complexity of the original narratives while representing it in an interoperable format. The dataset is designed to evaluate structured diagnostic reasoning over FHIR-style clinical context, with the source case reports providing diagnostically rich and challenging presentations.

[1] Dataset available on Hugging Face.

2. Related Work

Clinical data transformation and interoperability. Prior work focuses on transforming heterogeneous clinical data into standardized formats such as FHIR. Traditional approaches rely on rule-based NLP systems (Wang et al., 2018), often combining multiple tools for entity extraction and normalization. More recent LLM-based methods, including FHIR-GPT (Li et al., 2024b) and Infherno (Frei et al., 2026), convert clinical text into structured FHIR resources. However, these approaches primarily reconstruct existing clinical records and remain limited in resource coverage, rather than generating diverse or controllable patient data for downstream evaluation. Synthetic generators such as Synthea (Walonoski et al., 2018) provide large-scale FHIR-compatible patient data, but offer limited control over clinical complexity and patient-level variation.

LLMs for structured EHR and FHIR-based reasoning. Benchmarks such as EHRStruct (Yang et al., 2026) and FHIR-AgentBench (Lee et al., 2025) evaluate LLMs on structured EHR and FHIR-based tasks, showing that models struggle with knowledge-driven reasoning, retrieval over

complex patient records, and sensitivity to input formats and evaluation settings. However, these benchmarks operate on fixed datasets and cannot generate new patient scenarios, vary clinical complexity, or systematically probe model behavior under controlled conditions.

To our knowledge, no prior work provides controllable, text-driven generation of clinically realistic FHIR records designed specifically for evaluating diagnostic reasoning. Our work addresses this gap by enabling on-demand generation of structured patient data from unstructured inputs for evaluation in clinically realistic settings.

3. Method

Generating FHIR from free text using LLMs can produce hallucinated or invalid clinical codes, structurally inconsistent resources, and diagnostic leakage that biases evaluation. We address these issues with a multi-stage synthetic patient generator that converts unstructured English free-text into structurally valid, terminology-grounded HL7 FHIR R4 patient bundles. The generator targets resources needed to represent diagnostic case context rather than attempting full longitudinal EHR reconstruction.

Unlike agent-based approaches in which the model dynamically decides when to invoke tools (Frei et al., 2026), our pipeline calls LLMs at fixed stages: clinical information extraction, candidate reranking during terminology grounding, FHIR synthesis, and semantic leak detection. These stages are supported by deterministic vocabulary retrieval, structural and clinical-consistency validation, and rule-based post-processing. The extraction model identifies concepts from free-text without emitting clinical codes. Code assignment is deferred to a grounding stage that embeds, retrieves, and reranks candidates against target terminologies. This separation enables extraction validation, terminology grounding, and completeness checks on a flat intermediate representation before FHIR synthesis. After FHIR synthesis, validation errors are fed back through a repair loop, while post-processing handles completeness and normalization.

We use different LLMs for each stage: GPT-5.4-mini (OpenAI, 2026b) for extraction, Gemini 3.5 Flash (Google DeepMind, 2026b) for candidate reranking, and Anthropic's Claude (claude-sonnet-4-5) (Anthropic, 2025) for FHIR synthesis and semantic leak detection. All stages use greedy decoding (temperature 0). Extraction and reranking use JSON-schema-constrained outputs, while synthesis and leak detection use token caps of 64K and 4K, respectively. Reasoning parameters use provider defaults, and extraction and reranking responses are cached to reduce residual nondeterminism.

3.1. Extraction

The first LLM stage parses each free-text case into a typed intermediate representation of the patient. It captures demographics and the clinical findings, including symptoms, findings, vitals, labs, medications, procedures, and history, with a verbatim source quote retained for every extracted item. Resource-specific attributes mentioned in prose, such as analyte for laboratory results or dose for medications, are also extracted. Pertinent negatives are flagged so they are not later asserted as positive findings, and the primary diagnosis is captured separately to support optional diagnosis hiding.

3.2. Terminology Grounding

Each extracted concept is routed by type to one or more target terminologies seen in Table 1: SNOMED CT (National Library of Medicine), LOINC (McDonald et al., 2003), RxNorm (Nelson et al., 2011), and CVX (Centers for Disease Control and Prevention, 2018). Codes are drawn from an internally curated terminology store that aggregates OMOP (Voss et al., 2015) and other interoperable standards.

We first embed each concept description with SapBERT embeddings (Liu et al., 2021) and retrieve nearest-neighbor candidates from the routed terminologies using a FAISS index (Johnson et al., 2017) of preferred-term embeddings. An LLM reranker then selects the best candidate or rejects all candidates, using the verbatim source quote as context. For example, a quantitative laboratory value such as "serum potassium 3.2 mmol/L" causes a like-named code to be rejected when it has the wrong concept type, as shown in Figure 1B. Deterministic checks enforce consistency between the selected code and the target FHIR resource, following FHIR R4 coding guidance (HL7 International, a). A code-system preference selects the expected terminology for each resource, and a semantic-axis constraint rejects codes from incompatible SNOMED CT (National Library of Medicine) hierarchy, as shown in Table 1. Concepts with no acceptable candidate remain uncoded, with the verbatim source quote carried as free-text. A final validation pass checks emitted codes against the curated terminology store and reduces unverifiable codes to text only.

3.3. FHIR Synthesis and Validation

The second LLM stage assembles the grounded intermediate representation into HL7 FHIR R4 resources. The prompt specifies the mapping from each scenario field to its resource type and attributes, as shown in Table 1, to maintain structural conformance and clinical consistency. Structural resources for `Patient` and `Encounter` are also constructed. Each generated resource is checked by a three-layer validator: FHIR R4 schema conformance, FHIR

Table 1. Code-system preference and accepted SNOMED CT axis per FHIR resource. Target terminologies include SNOMED CT (National Library of Medicine), LOINC (McDonald et al., 2003), RxNorm (Nelson et al., 2011), and CVX (Centers for Disease Control and Prevention, 2018). Codes from other hierarchies are rejected and re-picked; ">" denotes fallback and "—" an unconstrained axis.

Vocabulary	FHIR resource	Accepted SNOMED axis
SNOMED CT	`Condition`	disorder, finding, situation, morphologic abnormality
SNOMED CT	`Observation` (symptom/finding)	finding, disorder, morphologic abnormality
LOINC > SNOMED CT	`Observation` (lab/test)	finding, disorder, morphologic abnormality, observable entity
SNOMED CT > LOINC	`Observation` (social)	social context
SNOMED CT	`Procedure`	procedure
RxNorm	`MedicationRequest`	—
CVX	`Immunization`	—
SNOMED CT	`AllergyIntolerance`	—
SNOMED CT	`FamilyMemberHistory`	—
LOINC	`DiagnosticReport`	—

invariants, and formatting and consistency checks. Validation errors are returned to the synthesis model for up to three repair attempts. A final deterministic post-processing pass enforces missing resources and normalizes units, dates, and status fields.

3.4. Diagnosis Hiding

We support configurable suppression of diagnostic conclusions during bundle generation. The assembled bundle is filtered using one of four modes: NONE removes all diagnostic conclusions; HIDDEN removes only the primary diagnosis; EXPLICIT retains only patient-stated conditions; FULL retains all extracted diagnoses. In NONE and HIDDEN modes, code- and substring-based filtering is followed by an LLM semantic scan over all narrative fields to redact residual diagnostic context, including abbreviations, implied conclusions, and synonyms absent from the synonym list.

4. MedCase-Structured

In this section, we introduce **MedCase-Structured**, a synthetic, case-report-derived FHIR dataset for structured diagnostic reasoning.

Table 2. MedCaseReasoning (Wu et al., 2025) conversion outcomes across dataset splits. Cases are partitioned into successfully generated MedCase-Structured examples, bundles generated but missing extracted element, and exclusion classes: imaging-dependent cases, and non-human / multi-patient reports.

Split	Original Total	Imaging Excluded	Subject Excluded	Incomplete Bundles	Success Cases
Test	897	769	17	1	**110**
Val	500	428	7	2	**63**
Train	13,092	11,262	222	49	**1,559**

4.1. Dataset

The dataset is constructed by applying our reusable text-to-FHIR pipeline to MedCaseReasoning (Wu et al., 2025), an open-access dataset of approximately 14,500 diagnostic cases sourced from publicly available case reports and designed to evaluate LLM alignment with clinician-authored diagnostic reasoning. Each case includes a final diagnosis. The original dataset is split into 13,092 training, 500 validation, and 897 test cases.

We exclude cases that are non-human, involve multiple patients, or reference imaging details, which are not currently supported by the generator. The remaining 1,784 cases are processed through the pipeline. This conservative filtering reduces coverage but ensures that released cases satisfy the structural, terminology-grounding, and diagnosis-hiding constraints required for evaluation.

Table 2 shows the final dataset statistics. Of the pipeline-processed cases, 1,732 are successfully converted into valid FHIR representations (97.1% success rate).

4.2. Pipeline Failure Modes

Among the 1,732 shipped bundles, terminology coverage remains the primary limitation. 18.2% of distinct extracted clinical concepts (3,775/20,794) were preserved as text but did not receive a normalized code, reflecting code-retrieval limitations concentrated in the categories shown in Table 4.

Separately, 52 attempted cases (2.9% of the 1,784 processed cases) did not produce complete bundles and were excluded from the final **MedCase-Structured** output. In each, a clinical element extracted from the narrative was dropped during grounding or validation. In 42 cases, the element could not be grounded to a code whose semantic axis matches its FHIR resource type and was suppressed. In the remaining 10, an assigned code failed FHIR or terminology validation.

4.3. Evaluation

We evaluate the diagnostic accuracy of popular LLM models on **MedCase-Structured** and compare it to that on the

Table 3. Terminology coverage in MedCase-Structured. Target terminologies include SNOMED CT (National Library of Medicine), LOINC (McDonald et al., 2003), RxNorm (Nelson et al., 2011), and CVX (Centers for Disease Control and Prevention, 2018). Resource-level coverage is the fraction of code-eligible resource instances carrying at least one terminology code; concept-level coverage is the fraction of distinct concepts (deduplicated by display text) carrying a code; distinct codes counts unique codes emitted.

	Resource-level		**Concept-level**		
Terminology	N	Coverage	N	Coverage	**Distinct Codes**
SNOMED CT	19,517	85.0%	15,573	83.5%	4,619
LOINC	15,562	90.0%	3,903	77.6%	1,753
RxNorm	3,681	84.7%	1,280	74.5%	634
CVX	46	93.5%	38	94.7%	24
Overall	**38,806**	**87.0%**	**20,794**	**81.8%**	**7,030**

Table 4. Sources of uncoded (text-only) concept gaps. Counts are distinct never-coded concepts; percentages are of unique concepts in each terminology (denominators in Table 3). Target terminologies include SNOMED CT (National Library of Medicine), LOINC (McDonald et al., 2003), RxNorm (Nelson et al., 2011), and CVX (Centers for Disease Control and Prevention, 2018).

Terminology	**Never-coded**	**% of unique**	**Dominant gap categories**
SNOMED CT	2,571	16.5%	Dermatologic/morphologic findings; normal exam findings; axis-mismatched symptoms
LOINC	875	22.4%	Specialty examinations; pathology stains; qualitative serologies
RxNorm	327	25.5%	Drug-class mentions without an ingredient; non-RxNorm substances; missing ingredients
CVX	2	5.3%	Negligible
Total	**3,775**	**18.2%**	

corresponding questions in text format from the original MedCaseReasoning dataset. We average the results over 3 runs and report the mean accuracy. The detailed setup of the experiment is shown in Section B.

Table 5 shows the results of our evaluation. LLMs perform consistently worse in diagnostic reasoning when dealing with structured FHIR inputs compared to plain text patient descriptions. We also see insignificant improvement from in-context learning, showing that the ability to understand and reason with structured FHIR data might not be easily learned through simple few shot examples. The experiment results indicate that diagnostic reasoning on structured EHR data is a far more challenging task than simple text-based reasoning.

Table 5. Comparison of LLM diagnostic accuracy on the FHIR-based MedCase-Structured (MCS) dataset and the subset of the corresponding questions in the text-based MedCaseReasoning (MCR) (Wu et al., 2025). Evaluated models are GPT-5.4 (OpenAI, 2026a), Gemini-3.1-Pro (Google DeepMind, 2026a), and Claude-Opus-4.6 (Anthropic, 2026). Δ is the accuracy difference (MCS $-$ MCR).

Model	MCR	**MCS**	Δ
GPT-5.4			
w/ zero-shot	69.39	57.37	*–12.02*
w/ 1-shot	67.27	55.45	*–11.82*
w/ 5-shot	70.91	58.18	*–12.73*
Gemini-3.1-Pro			
w/ zero-shot	53.03	42.42	*–10.61*
w/ 1-shot	48.18	45.45	*–2.73*
w/ 5-shot	46.36	40.91	*–5.45*
Claude-Opus-4.6			
w/ zero-shot	61.82	54.55	*–7.27*
w/ 1-shot	61.82	56.82	*–5.00*
w/ 5-shot	61.37	56.30	*–5.07*

5. Conclusion

We introduce a reusable pipeline to convert clinical case narratives into terminology-grounded HL7 FHIR R4 bundles and release **MedCase-Structured**, a synthetic FHIR dataset built for evaluating CDSS in structured, EHR-congruent diagnostic reasoning settings.

Our results show that LLMs achieve consistently lower diagnostic accuracy when operating over structured FHIR inputs compared to plain text descriptions. This suggests that structured FHIR inputs may introduce additional challenges for diagnostic reasoning. These findings highlight the importance of evaluating CDSS on deployment-aligned data formats, as performance on simplified or unrelated inputs may not reflect behavior in clinical environments.

Our pipeline has several limitations. It currently supports a limited subset of FHIR resources and represents temporal information through repeated, date-aware resources rather than full longitudinal patient trajectories. Because imaging evidence is not yet supported, the released dataset emphasizes cases whose diagnostic context can be represented in text-grounded FHIR resources. Terminology grounding also remains challenging when clinical descriptions are unsupported, ambiguous, or too specific to map cleanly to a single standardized concept. Future work should broaden terminology retrieval, add axis-aware grounding, extend resource and longitudinal modeling, and incorporate context-aware validation and imaging support to improve robustness.

Impact Statement

This work aims to improve evaluation of CDSS in EHR-native settings by generating structured, clinically realistic

synthetic patient data for controlled and interoperable benchmarking.

Synthetic data may not fully capture real-world complexity, and errors in generation or terminology grounding may propagate into downstream evaluations. These datasets should therefore complement, not replace, real-world clinical validation.

Acknowledgment

We thank David Kang for helpful discussions and feedback on early drafts of this paper. We are grateful to the anonymous reviewers for their constructive comments. This work was supported by System Inc.

References

Anthropic. Introducing Claude Sonnet 4.5, 2025. URL https://www.anthropic.com/news/claude-sonnet-4-5.

Anthropic. Claude Opus 4.6, 2026. URL https://www.anthropic.com/news/claude-opus-4-6.

Bennett, A. M., Ulrich, H., van Damme, P., Wiedekopf, J., and Johnson, A. E. W. MIMIC-IV on FHIR: converting a decade of in-patient data into an exchangeable, interoperable format. *Journal of the American Medical Informatics Association*, 30(4):718–725, April 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad002. URL https://doi.org/10.1093/jamia/ocad002.

Centers for Disease Control and Prevention. IIS | Code Sets | CVX | Vaccines | CDC, September 2018. URL https://www2a.cdc.gov/vaccines/iis/iisstandards/vaccines.asp?rpt=cvx.

Frei, J., Feldhus, N., Raithel, L., Roller, R., Meyer, A., and Kramer, F. Infherno: End-to-end Agent-based FHIR Resource Synthesis from Free-form Clinical Notes. In Croce, D., Leidner, J., and Moosavi, N. S. (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 163–174, Rabat, Marocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-382-1. doi: 10.18653/v1/2026.eacl-demo.13. URL https://aclanthology.org/2026.eacl-demo.13/.

Google DeepMind. Gemini 3.1 Pro - Model Card, 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.5 Flash - Model Card, 2026b. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

HL7 International. Terminology-module - FHIR v4.0.1, a. URL https://hl7.org/fhir/R4/terminology-module.html.

HL7 International. FHIR R4 (v4.0.1), b. URL https://hl7.org/fhir/R4/index.html.

Johnson, A. E. W., Bulgarelli, L., Shen, L., Gayles, A., Shammout, A., Horng, S., Pollard, T. J., Hao, S., Moody, B., Gow, B., Lehman, L.-w. H., Celi, L. A., and Mark, R. G. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1): 1, January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01899-x. URL https://www.nature.com/articles/s41597-022-01899-x.

Johnson, J., Douze, M., and Jégou, H. Billion-scale similarity search with GPUs, February 2017. URL http://arxiv.org/abs/1702.08734. arXiv:1702.08734 [cs].

Lee, G., Bach, E., Yang, E., Pollard, T., Johnson, A., Choi, E., jia, Y., and Lee, J. H. FHIR-AgentBench: Benchmarking LLM Agents for Realistic Interoperable EHR Question Answering, November 2025. URL http://arxiv.org/abs/2509.19319. arXiv:2509.19319 [cs].

Li, L., Zhou, J., Gao, Z., Hua, W., Fan, L., Yu, H., Hagen, L., Zhang, Y., Assimes, T. L., Hemphill, L., and Ma, S. A scoping review of using Large Language Models (LLMs) to investigate Electronic Health Records (EHRs), May 2024a. URL http://arxiv.org/abs/2405.03066. arXiv:2405.03066 [cs].

Li, Y., Wang, H., Yerebakan, H. Z., Shinagawa, Y., and Luo, Y. FHIR-GPT Enhances Health Interoperability with Large Language Models. *NEJM AI*, 1(8):AIcs2300301, July 2024b. doi: 10.1056/AIcs2300301. URL https://ai.nejm.org/doi/abs/10.1056/AIcs2300301.

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., and Collier, N. Self-Alignment Pretraining for Biomedical Entity Representations, April 2021. URL http://arxiv.org/abs/2010.11784. arXiv:2010.11784 [cs].

Mansoor, I., Abdullah, M., Rizwan, M. D., and Fraz, M. M. Reasoning with large language models in medicine: a systematic review of techniques, challenges and clinical integration. *Health Information Science and Systems*, 14 (1):6, November 2025. ISSN 2047-2501. doi: 10.1007/s13755-025-00403-0. URL https://doi.org/10.1007/s13755-025-00403-0.

McDonald, C. J., Huff, S. M., Suico, J. G., Hill, G., Leavelle, D., Aller, R., Forrey, A., Mercer, K., DeMoor, G., Hook,

J., Williams, W., Case, J., and Maloney, P. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical Chemistry*, 49(4):624–633, April 2003. ISSN 0009-9147. doi: 10.1373/49.4.624.

National Library of Medicine. SNOMED CT United States Edition. URL https://www.nlm.nih.gov/healthit/snomedct/us_edition.html.

Navarro, D. F., Magrabi, F., and Coiera, E. Evaluation format, not model capability, drives triage failure in the assessment of consumer health AI, March 2026. URL http://arxiv.org/abs/2603.11413. arXiv:2603.11413 [cs].

Nelson, S. J., Zeng, K., Kilbourne, J., Powell, T., and Moore, R. Normalized names for clinical drugs: RxNorm at 6 years. *Journal of the American Medical Informatics Association : JAMIA*, 18(4): 441–448, 2011. ISSN 1067-5027. doi: 10.1136/amiajnl-2011-000116. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC3128404/.

OpenAI. Introducing GPT-5.4, June 2026a. URL https://openai.com/index/introducing-gpt-5-4/.

OpenAI. Introducing GPT-5.4 mini and nano, June 2026b. URL https://openai.com/index/introducing-gpt-5-4-mini-and-nano/.

Shool, S., Adimi, S., Saboori Amleshi, R., Bitaraf, E., Golpira, R., and Tara, M. A systematic review of large language model (LLM) evaluations in clinical medicine. *BMC medical informatics and decision making*, 25(1):117, March 2025. ISSN 1472-6947. doi: 10.1186/s12911-025-02954-4.

Voss, E. A., Makadia, R., Matcho, A., Ma, Q., Knoll, C., Schuemie, M., DeFalco, F. J., Londhe, A., Zhu, V., and Ryan, P. B. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association: JAMIA*, 22(3):553–564, May 2015. ISSN 1527-974X. doi: 10.1093/jamia/ocu023.

Walonoski, J., Kramer, M., Nichols, J., Quina, A., Moesel, C., Hall, D., Duffett, C., Dube, K., Gallagher, T., and McLachlan, S. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238, March 2018. ISSN 1527-974X. doi: 10.1093/jamia/ocx079. URL https://doi.org/10.1093/jamia/ocx079.

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S., and Liu, H. Clinical Information Extraction Applications: A Literature Review. *Journal of biomedical informatics*, 77: 34–49, January 2018. ISSN 1532-0464. doi: 10.1016/j.jbi.2017.11.011. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5771858/.

Wu, K., Wu, E., Thapa, R., Wei, K., Zhang, A., Suresh, A., Tao, J. J., Sun, M. W., Lozano, A., and Zou, J. MedCaseReasoning: Evaluating and learning diagnostic reasoning from clinical case reports, May 2025. URL http://arxiv.org/abs/2505.11733. arXiv:2505.11733 [cs].

Yang, X., Zhao, X., and Shen, Z. EHRStruct: A Comprehensive Benchmark Framework for Evaluating Large Language Models on Structured Electronic Health Record Tasks, April 2026. URL http://arxiv.org/abs/2511.08206. arXiv:2511.08206 [cs].

A. MedCase-Structured Sample

Listing 1 shows the truncated version of a representative structured patient bundle from MedCase-Structured. The full dataset is available at https://huggingface.co/datasets/system-technologies/MedCase-Structured.

Listing 1. Example FHIR R4 Patient Bundle from MedCase-Structured.


```
{
  "resourceType": "Bundle",
  "type": "collection",
  "entry": [
    {
      "fullUrl": "urn:uuid:d0597131-f20e-47b3-96c4-0c37d5998555",
      "resource": {
        "resourceType": "Patient",
        "id": "d0597131-f20e-47b3-96c4-0c37d5998555",
        "name": [
          {
            "use": "official",
            "given": [
              "Synthetic"
            ],
            "family": "Patient"
          }
        ],
        "gender": "male",
        "birthDate": "1974-01-15"
      }
    },
    {
      "fullUrl": "urn:uuid:17116138-6e0a-4ed3-8235-c9c93ad129d9",
      "resource": {
        "resourceType": "Encounter",
        "id": "17116138-6e0a-4ed3-8235-c9c93ad129d9",
        "status": "finished",
        "class": {
          "system": "http://terminology.hl7.org/CodeSystem/v3-ActCode",
          "code": "AMB",
          "display": "ambulatory"
        },
        "type": [
          {
            "coding": [
              {
                "system": "http://snomed.info/sct",
                "code": "185347001",
                "display": "Encounter␣for␣problem"
              }
            ],
            "text": "Encounter␣for␣problem"
          }
        ],
        "subject": {
          "reference": "Patient/d0597131-f20e-47b3-96c4-0c37d5998555"
        },
        "period": {
          "start": "2026-06-25T09:00:00-05:00",
          "end": "2026-06-25T10:00:00-05:00"
        },
        "reasonCode": [
          {
            "coding": [
              {
                "system": "http://snomed.info/sct",
                "code": "70819003",
                "display": "Erythema"
              }
            ],
            "text": "Severe␣redness␣around␣right␣elbow"
          }
        ]
      }
    },
    {
      "fullUrl": "urn:uuid:bc787ee0-4ff6-4260-a49c-ff48d9154d9b",
      "resource": {
        "resourceType": "Condition",
        "id": "bc787ee0-4ff6-4260-a49c-ff48d9154d9b",
        "clinicalStatus": {
          "coding": [
            {
              "system": "http://terminology.hl7.org/CodeSystem/condition-clinical",
              "code": "active",
              "display": "Active"
            }
          ],
          "text": "Active"
        },
        "verificationStatus": {
          "coding": [
            {
              "system": "http://terminology.hl7.org/CodeSystem/condition-ver-status",
              "code": "confirmed",
              "display": "Confirmed"
            }
          ]
        },
        "category": [
          {
            "coding": [
              {
                "system": "http://terminology.hl7.org/CodeSystem/condition-category",
```

```
                "code": "problem-list-item",
                "display": "Problem␣List␣Item"
              }
            ],
            "text": "Problem␣List␣Item"
          }
        ],
        "code": {
          "coding": [
            {
              "system": "http://snomed.info/sct",
              "code": "70819003",
              "display": "Erythema"
            }
          ],
          "text": "Severe␣redness␣around␣right␣elbow"
        },
        "bodySite": [
          {
            "coding": [
              {
                "system": "http://snomed.info/sct",
                "code": "368148009",
                "display": "Right␣elbow␣region"
              }
            ],
            "text": "right␣elbow"
          }
        ],
        "subject": {
          "reference": "Patient/d0597131-f20e-47b3-96c4-0c37d5998555"
        },
        "onsetDateTime": "2026-06-18",
        "recordedDate": "2026-06-25",
        "severity": {
          "text": "severe"
        }
      }
    },
    {
      "fullUrl": "urn:uuid:3235971f-df4a-483a-8134-6e34f3bd8103",
      "resource": {
        "resourceType": "Condition",
        "id": "3235971f-df4a-483a-8134-6e34f3bd8103",
        "clinicalStatus": {
          "coding": [
            {
              "system": "http://terminology.hl7.org/CodeSystem/condition-clinical",
              "code": "active",
              "display": "Active"
            }
          ],
          "text": "Active"
        },
        "verificationStatus": {
          "coding": [
            {
              "system": "http://terminology.hl7.org/CodeSystem/condition-ver-status",
              "code": "confirmed",
              "display": "Confirmed"
            }
          ]
        },
        "category": [
          {
            "coding": [
              {
                "system": "http://terminology.hl7.org/CodeSystem/condition-category",
                "code": "problem-list-item",
                "display": "Problem␣List␣Item"
              }
            ],
            "text": "Problem␣List␣Item"
          }
        ],
        "code": {
          "coding": [
            {
              "system": "http://snomed.info/sct",
              "code": "90673000",
              "display": "Burning␣sensation"
            }
          ],
          "text": "Intense␣burning␣sensation␣of␣right␣elbow"
        },
        "bodySite": [
          {
            "coding": [
              {
                "system": "http://snomed.info/sct",
                "code": "368148009",
                "display": "Right␣elbow␣region"
              }
            ],
            "text": "right␣elbow"
          }
        ],
        "subject": {
          "reference": "Patient/d0597131-f20e-47b3-96c4-0c37d5998555"
        },
        "onsetDateTime": "2026-06-18",
        "recordedDate": "2026-06-25",
        "severity": {
          "text": "intense"
        }
```

```
      }
    },
    {
      "fullUrl": "urn:uuid:3f5011c0-ee84-4a74-8541-ddb995bb8ddc",
      "resource": {
        "resourceType": "Condition",
        "id": "3f5011c0-ee84-4a74-8541-ddb995bb8ddc",
        "clinicalStatus": {
          "coding": [
            {
              "system": "http://terminology.hl7.org/CodeSystem/condition-clinical",
              "code": "active",
              "display": "Active"
            }
          ],
          "text": "Active"
        },
        "verificationStatus": {
          "coding": [
            {
              "system": "http://terminology.hl7.org/CodeSystem/condition-ver-status",
              "code": "confirmed",
              "display": "Confirmed"
            }
          ]
        },
        "category": [
          {
            "coding": [
              {
                "system": "http://terminology.hl7.org/CodeSystem/condition-category",
                "code": "problem-list-item",
                "display": "Problem␣List␣Item"
              }
            ],
            "text": "Problem␣List␣Item"
          }
        ],
        "code": {
          "coding": [
            {
              "system": "http://snomed.info/sct",
              "code": "403599000",
              "display": "Stinging␣of␣skin"
            }
          ],
          "text": "Intense␣stinging␣sensation␣of␣right␣elbow"
        },
        "bodySite": [
          {
            "coding": [
              {
                "system": "http://snomed.info/sct",
                "code": "368148009",
                "display": "Right␣elbow␣region"
              }
            ],
            "text": "right␣elbow"
          }
        ],
        "subject": {
          "reference": "Patient/d0597131-f20e-47b3-96c4-0c37d5998555"
        },
        "onsetDateTime": "2026-06-18",
        "recordedDate": "2026-06-25",
        "severity": {
          "text": "intense"
        }
      }
    },
// ... additional resources omitted for brevity
}
```


B. Experimental Setup

We use commercialized API endpoints provided by OpenAI, Google, and Anthropic to prompt corresponding LLMs for diagnostic reasoning on MedCaseReasoning and MedCase-Structured. We set reasoning parameters to medium, max generation tokens to 800, and temperature to 1.0 across all experiments. For few-shot learning cases, we randomly sample cases from the training split to build the few shot learning prompts for each run.

For evaluation, we use a Claude Sonnet 4.6 model as the LLM judge to compare the ”diagnosis” field to the ground truth diagnosis string. We prompt the judge to assess whether the predicted diagnosis is clinically equivalent to the ground truth and output a final binary decision.

B.1. Diagnostic Reasoning Prompt

System Prompt

You are a careful physician solving clinical diagnostic reasoning cases. Use only the provided case information. Return valid JSON only.

User Prompt

You will receive a {*FHIR Bundle JSON for a clinical case* OR *plain text clinical case description*}. Determine the most likely final diagnosis.
Return exactly this JSON schema: "diagnosis": "single most likely diagnosis", "reasoning": "brief explanation using the case evidence"
{`fhir_bundle`}
Now solve the target case.
Target case: {`case_input`}

B.2. Judge Prompt

System Prompt

You judge whether a predicted diagnosis is clinically equivalent to the ground truth. Accept synonyms, spelling variants, and equivalent specificity. Return valid JSON only.

User Prompt

Ground truth diagnosis: {`ground_truth`}
Predicted diagnosis: {`prediction`}