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

Anonymous Authors[1]

Abstract

Large language models (LLMs) show promise for clinical reasoning and decision support, but evaluation in realistic, electronic health record-congruent settings remains limited. Existing benchmarks often rely on static datasets or unstructured inputs that do not reflect the structured, interoperable data formats used in clinical systems. We introduce a pipeline for generating clinically realistic HL7 FHIR R4 bundles from unstructured text, enabling controllable evaluation of clinical decision support systems. The pipeline combines staged LLM generation with terminology-grounded validation and repair to reduce hallucinated codes and enforce structural and semantic consistency. Applying this approach to MedCaseReasoning, we construct **MedCase-Structured**, a synthetic dataset aligned with clinician-authored diagnostic cases, achieving valid FHIR generation for 82.5% of 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 and benchmarking in clinically realistic, end-to-end 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) 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. 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 highly realistic, publicly available, and EHR-congruent synthetic clinical data. 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 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

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

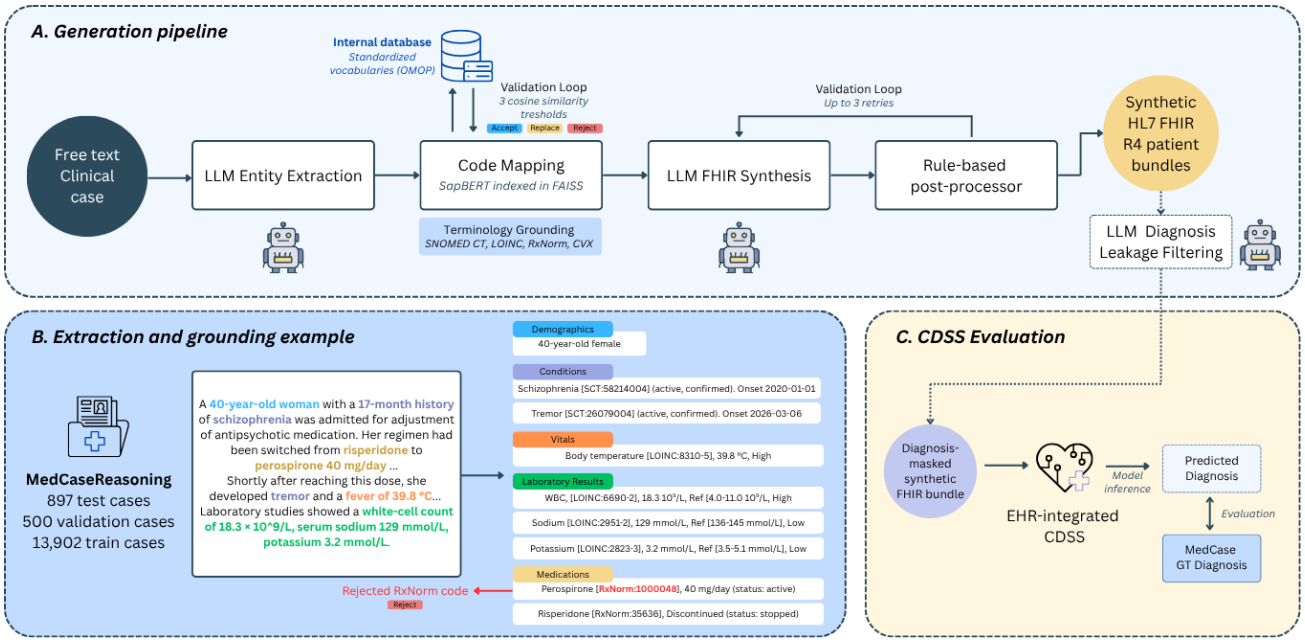

Figure 1. Overview of MedCase-Structured. (A) Free-text cases are converted into terminology-grounded HL7 FHIR R4 bundles. (B) An example MedCaseReasoning (Wu et al., 2025) case shows extraction, grounding, and rejection of an invalid RxNorm code. (C) Diagnosis-masked bundles are used for EHR-congruent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

methods for generating clinically realistic patient data that can systematically evaluate model reasoning under diverse and challenging conditions.

To address this gap, we introduce a pipeline for generating clinically realistic 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). Each case in **MedCase-Structured** is represented as a complete, terminology-validated FHIR R4 patient bundle, preserving the diagnostic complexity of the original clinical narratives while encoding them in a structured, interoperable format. The dataset provides a rich testbed for training and evaluating CDSS over realistic, EHR-style inputs.

[1]Full dataset will be released upon the publication of this paper. To see example data, please refer to Section A.

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 often leads to hallucinated or invalid clinical codes, structural inconsistencies across resources, and leakage of diagnostic information that can bias 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.

Unlike agent-based approaches where the model dynamically decides when to invoke tools (Frei et al., 2026), our pipeline calls the LLM at three fixed stages: clinical information extraction, FHIR synthesis, and semantic leak detection. These stages are supported by deterministic terminology grounding, structural and clinical-consistency validation, and rule-based post-processing. Terminology grounding validates extracted codes against a curated clinical terminology store using SapBERT embeddings (Liu et al., 2021) indexed in FAISS (Johnson et al., 2017). Validation errors are fed back into the synthesis stage through a repair loop, while post-processing handles completeness and normalization. All LLM calls use Anthropic's Claude (claude-sonnet-4-20250514) (Anthropic, 2025) at temperature 0 for reproducibility.

3.1. Extraction

The first LLM stage extracts a typed intermediate representation of the patient description and clinical findings (patient demographics, symptoms, findings, vitals, labs, medications, procedures, and history) from free-text input, with a verbatim source quote retained for every extracted item. This separation allows for extraction validation, terminology grounding, and completeness checks on a flat structure before any FHIR is produced.

3.2. Terminology Grounding

SNOMED CT, LOINC, RxNorm, and CVX codes produced by the extraction step are validated against our internally curated terminology store, which aggregates OMOP and other interoperable standards. Candidates identified for repair are identified by keyword search and alternative semantic similarity using SapBERT (Liu et al., 2021) embeddings of preferred terms indexed in FAISS (Johnson et al., 2017). We use three cosine-similarity thresholds to accept, replace, or reject each LLM-provided code whose display does not match the input description or synonyms.

Table 1. MedCaseReasoning (Wu et al., 2025) conversion outcomes across dataset splits. Final usable cases correspond to successfully generated MedCase-Structured examples.

Split	Original Total	Imaging Excluded	Code Errors	Other Excluded	Final
Test	897	777	14	11	**95**
Val	500	438	10	2	**50**
Train	13,092	11,568	232	28	**1,263**

3.3. FHIR Synthesis and Validation

The second LLM stage converts the grounded extracted clinical scenario into FHIR resources following HL7 R4 templates. We support generation of `Patient`, `Encounter`, `Condition`, `Observation`, `MedicationRequest`, `Procedure`, `DiagnosticReport`, `FamilyMemberHistory`, `AllergyIntolerance`, and `Immunization`. The prompt defines the mapping between scenario fields and FHIR resource types to maintain structural conformance and clinical consistency. Validation errors are returned to the LLM for repair for up to three attempts. After generation, rule-based post-processors backfill missing resources and normalize units, dates, and status fields.

3.4. Diagnosis Hiding

We enable configurable suppression of diagnostic conclusions in bundle generation. The assembled bundle is filtered according to 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, exhaustive code- and substring-based filtering is followed by a third LLM stage that performs a semantic scan over all narrative fields to identify and redact residual diagnostic context (abbreviations, implied conclusions, synonyms not listed in the synonym list).

4. MedCase-Structured

In this section, we introduce **MedCase-Structured**, a clinically realistic synthetic dataset for diagnostic reasoning.

4.1. Dataset

Our dataset is derived from 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.

Table 2. Failure modes in MedCaseReasoning (Wu et al., 2025) conversion. Counts for terminology and semantic errors are code-level (a single case may contribute multiple errors), while excluded cases are patient-level (one count per excluded case).

Category	Failure Type	N	Example
Terminology errors	Hallucinated LOINC codes	183	"septic workup"; "pharmacological challenge test"
	Hallucinated RxNorm codes	126	Re-hallucinated invalid Stage 1 code
	Non-specific drug classes	103	"oral antibiotics"; "topical corticosteroid paste"
	CVX synonym gaps	12	"Moderna booster"; "fully immunized"
Semantic mapping errors	Overly specific descriptions	32	"loosening of lower teeth requiring dental implants"
	Incorrect SNOMED category	33	Procedure code assigned to finding
Excluded cases	Missing demographics	4	No age in source description
	Multi-patient descriptions	9	Multiple patients in one case
	Non-human cases	25	Veterinary reports

We filter out case prompts that are non-human, involve multiple patients, or references imaging details, as these are not supported by our generator. The remaining cases are processed through our synthetic patient generation pipeline.

Table 1 shows the final statistics of our dataset. After filtering, 1,408 are successfully converted into valid FHIR representations, corresponding to 82.5% of cases processed by the pipeline.

4.2. Pipeline Failure Modes

As shown in Table 2, terminology grounding remains the primary challenge, with most failures arising from hallucinated or unsupported codes, terminology coverage gaps, and semantic mapping cases. Remaining exclusions reflect input inconsistencies such as missing demographics or multi-patient descriptions. These exclusions reflect design choices to ensure each case contains sufficient context for diagnostic evaluation and corresponds to a single patient.

4.3. Evaluation

We evaluate the diagnostic accuracy of popular LLM models on **MedCase-Structured** and compare it to that on the corresponding questions in text format from the original MedCaseReasoning dataset. The detailed setup of the experiment is shown in Section B.

Table 3 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. This indicates that diagnostic reasoning on structured EHR data is a far more challenging task than simple text-based reasoning.

Table 3. 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).

Model	MCR	**MCS**	Δ
GPT-5.4			
w/ zero-shot	65.26	61.05	*–4.21*
w/ 1-shot	74.74	51.58	*–23.16*
w/ 5-shot	74.74	53.68	*–21.06*
Gemini-3.1-Pro			
w/ zero-shot	58.95	52.63	*–6.32*
w/ 1-shot	65.26	53.68	*–11.58*
w/ 5-shot	63.16	57.89	*–5.28*
Claude-Opus-4.6			
w/ zero-shot	68.42	53.63	*–14.79*
w/ 1-shot	69.47	54.74	*–14.73*
w/ 5-shot	66.32	58.95	*–7.37*

5. Conclusion

We introduce **MedCase-Structured**, a clinically realistic synthetic FHIR dataset constructed from clinician-authored case descriptions in MedCaseReasoning (Wu et al., 2025). **MedCase-Structured** enables evaluation of CDSS in structured, EHR-congruent 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 does not fully model longitudinal patient trajectories, instead representing temporal information through repeated, date-aware resources. Terminology grounding also remains a challenge, particularly for hallucinated or unsupported codes, terminology coverage gaps, and clinical descriptions that are too specific or ambiguous to map cleanly to a single standardized concept. Future work should expand resource coverage, improve longitudinal modeling, broaden terminology support, and incorporate stronger context-aware validation to further 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.

References

Anthropic. Claude Sonnet 4, 2025. URL https://claude.ai/.

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.

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/.

HL7 International. FHIR R4 (v4.0.1). 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.

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].

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.

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 sample is available at https://anonymous.4open.science/r/MedCase-Structured/ (anonymized for review).

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


```
{
  "resourceType": "Bundle",
  "type": "collection",
  "entry": [
    {
      "fullUrl": "urn:uuid:5e319753-4f1a-4397-af5e-0efb780ac76e",
      "resource": {
        "resourceType": "Patient",
        "id": "5e319753-4f1a-4397-af5e-0efb780ac76e",
        "name": [
          {
            "use": "official",
            "given": [
              "Synthetic"
            ],
            "family": "Patient"
          }
        ],
        "gender": "female",
        "birthDate": "1975-01-15"
      }
    },
    {
      "fullUrl": "urn:uuid:6e020811-3ce7-44ad-85cc-38348e16e9ad",
      "resource": {
        "resourceType": "Encounter",
        "id": "6e020811-3ce7-44ad-85cc-38348e16e9ad",
        "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/5e319753-4f1a-4397-af5e-0efb780ac76e"
        },
        "period": {
          "start": "2026-04-30",
          "end": "2026-04-30"
        },
        "reasonCode": [
          {
            "coding": [
              {
                "system": "http://snomed.info/sct",
                "code": "271759003",
                "display": "Bullous␣eruption"
              }
            ],
            "text": "bullous␣rash␣on␣her␣left␣arm,␣axilla,␣and␣lateral␣chest␣wall␣accompanied␣by␣subjective␣fever"
          }
        ]
      }
    },
    {
      "fullUrl": "urn:uuid:7a750e34-a26f-41a3-aae6-4f58fb897ebd",
      "resource": {
        "resourceType": "Condition",
        "id": "7a750e34-a26f-41a3-aae6-4f58fb897ebd",
        "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": "271759003",
              "display": "Bullous␣eruption"
            }
          ],
          "text": "bullous␣rash␣on␣her␣left␣arm,␣axilla,␣and␣lateral␣chest␣wall"
        },
        "subject": {
          "reference": "Patient/5e319753-4f1a-4397-af5e-0efb780ac76e"
        },
        "onsetDateTime": "2026-04-28",
        "recordedDate": "2026-04-30"
      }
    },
    {
      "fullUrl": "urn:uuid:d74b1dd6-2e22-4521-87e5-8b2d8c9b931d",
      "resource": {
        "resourceType": "Condition",
        "id": "d74b1dd6-2e22-4521-87e5-8b2d8c9b931d",
        "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": "386661006",
              "display": "Fever"
            }
          ],
          "text": "subjective␣fever"
        },
        "subject": {
          "reference": "Patient/5e319753-4f1a-4397-af5e-0efb780ac76e"
        },
        "onsetDateTime": "2026-04-28",
        "recordedDate": "2026-04-30"
      }
    },
  // ... 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 an OpenAI GPT-5.4 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`}