MTDiag: A Multi-Turn Diagnostic Dataset Towards Clinically Meaningful LLM Evaluation

Pia Chouayfati[1], Alexander M. Fichtl[1], Miriam Anschütz[1], George Doumat[2], Georg Groh[1]

[1]Technical University of Munich [2]Department of Internal Medicine, UT Southwestern

Abstract

Clinical diagnosis is fundamentally interactive and incremental, yet the dominant paradigm for evaluating Large Language Models (LLMs) in medicine remains static QA benchmarks or template-based dialogues. These benchmarks say little about whether a model can serve as a diagnostic agent in a dynamic clinical encounter, with LLMs showing significant accuracy and reliability degradation in multi-turn settings. To address this issue, we present MTDiag, a large multi-turn diagnostic dialogue dataset constructed from three heterogeneous sources: DDXPlus, MIMIC-IV, and published case reports (AJCR), covering common ED presentations as well as long-tail rare and atypical conditions. All cases are normalized into a canonical `PatientVector` schema anchored in the most comprehensive and widely-adopted medical knowledge bases (UMLS concept identifiers, with ICD-10 diagnosis codes). We release the `PatientVector` schema, a UserLM-8B-based utterance-generation pipeline, and the physician-validated dataset that converts structured clinical evidence into natural-language utterances. Importantly, we introduce and motivate clinical knowledge-grounded metrics for evaluating LLMs as diagnostic agents, beyond diagnostic accuracy, for the task of multi-turn differential diagnosis.

1 Introduction

Large language models (LLMs) are increasingly seen as viable tools to enhance healthcare by assisting physicians in their daily tasks (Vladika et al., 2026), but their suitability for direct patient-facing settings requires extensive testing and evaluation. The dominant approach for evaluating LLMs in medicine has long been the static multiple-choice examination or question-answering (QA) set-up (Jin et al., 2019, 2021; Nentidis et al., 2025). This approach has driven measurable progress, with frontier models routinely achieving near-perfect scores on medical licensing benchmarks. However, a recent systematic review of 39 clinical LLM benchmarks highlights a severe knowledge-practice gap: while models achieve up to 90% accuracy on knowledge-based questions, performance drops to roughly 45% on practice-based diagnostic tasks (Gong et al., 2025). This degradation occurs because clinical diagnosis is fundamentally an interactive and incremental process of information extraction, prioritization, and synthesis under uncertainty. Existing benchmarks present models with complete, pre-structured, and often self-contained information. In reality, a physician must actively elicit information from a patient who may forget key symptoms, use imprecise language, or omit critical context. High performance on a vignette benchmark is therefore not predictive of real-world diagnostic competence (Alaa et al., 2025).

Recent efforts toward conversational diagnostic AI (Tu et al., 2025; Fan et al., 2025; Johri et al., 2024) confirm this, finding clear accuracy and reliability degradation in interactive settings, while millions of users already seek medical guidance from general commercial LLMs. Alarmingly, a recent qualitative analysis by the medical community of patient conversations with commercial LLMs has shown that LLMs provide unsafe answers to patient-posed medical questions (Draelos et al., 2026).

To address this gap, we present **MTDiag**, a large multi-turn diagnostic dialogue dataset constructed from three heterogeneous real and realistically simulated clinical sources and anchored in the UMLS (Fig. 1), which enables clinical reasoning evaluation beyond diagnostic accuracy. Our primary contributions are:

(1) MTDiag Dataset: A multi-turn diagnostic dataset normalizing data from DDXPlus (Fansi Tchango et al., 2022), MIMIC-IV (Johnson et al., 2023), and a curated diagnos-

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 821–840
August 2–5, 2026.

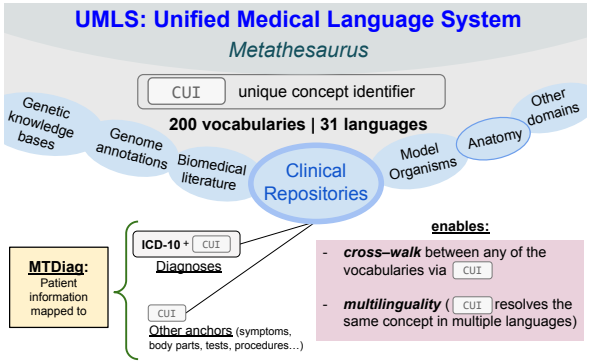

Figure 1: The UMLS (Unified Medical Language System) (Bodenreider, 2004) is a metathesaurus of international biological and medical knowledge bases, ontologies and vocabularies built for healthcare and medical system interoperability. It indexes unique concepts by CUI (unique concept identifier), and instantiates a semantic network (Fig. 2) which links a single concept across all participating source vocabularies.

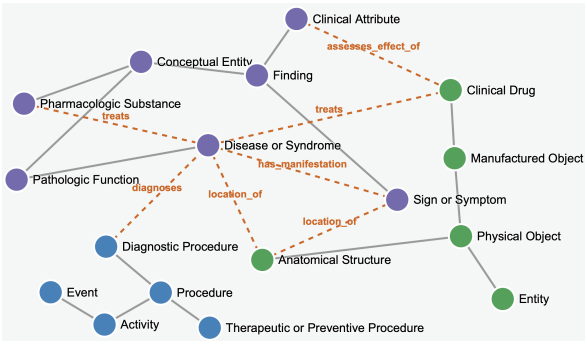

Figure 2: Clinically-relevant subset of the UMLS Semantic Network, showing Semantic Type nodes and typed relation edges. Nodes are color-coded by semantic category:
- **Events** (actions occurring in time), e.g. Procedure
- **Entities** (physical matter and anatomy) e.g. Clinical Drug
- **Conceptual Entities** (abstract classifications and states), e.g. Disease or Syndrome

Solid edges denote hierarchical *is a* relations; dashed edges denote typed associative relations.

tic AJCR case compilation (Hirosawa et al., 2024) into a canonical `PatientVector` schema (Fig. 6), where diagnoses are mapped to ICD-10[1], and symptoms are mapped to UMLS (Fig. 1) `CUI` identifiers.

(2) Utterance Generation Pipeline: A UserLM-8B (Naous et al., 2025) based framework that converts structured clinical evidence into natural, patient-like utterances. It allows case-specific generation of utterances and presentations, providing a set of canonical utterances per `PatientVector`.

(3) Patient Orchestrator: A MedGemma (Sellergren et al., 2025) based runtime agent that selects contextually appropriate responses from the pre-generated utterance pool. This "dialogue runner" simulates a dialogue between a patient and an examiner via the `PatientVector`.

2 Background and Related Work

MTDiag is designed to enable clinically-grounded simulation of the "diagnostic encounter," which we define as: a **patient** willingly seeking a **diagnosis** for a **"chief complaint" (CC)**, via a **dialogue** (examination) with a **medical practitioner** (examiner). The examiner's task is inherently one of **differential diagnosis**: the iterative (and implicit) process of generating, ranking, and progressively narrowing a set of candidate diagnoses by eliciting and weighing clinical evidence turn by turn, until a most probable diagnosis can be committed to (Fansi Tchango et al., 2022; Tu et al., 2025). MTDiag's structure thus indirectly models this process from the examiner side: the ground-truth diagnoses, symptom and history anchors, and dialogue logs serve to evaluate whether a model is correctly reasoning differentially: asking the right questions, in the right order, for the right reasons. In the following sections, we establish why this encounter is structurally different from what current benchmarks capture and survey the prior work that motivates and contextualizes the dataset design.

2.1 The Diagnostic Encounter as a Dialogue Modeling Task

Physicians gather information from patients who may be vague, forgetful, resistant, or unaware of what is clinically relevant. The history of present illness (HPI), a log of the patient's medical past, is never handed over; it is negotiated across a conversation (Redelmeier et al., 2001), with patients routinely exhibiting systematic failure modes in reporting their history, including errors in comprehension, recall, and expression. Vague patient presentations reduce the discriminative power of the symptom and lower the probability of a correct diagnosis (Sonnenberg and Gogel, 2002). Patients will also actively resist or renegotiate diagnostic framings after hearing a physician's assessment (Ijäs-Kallio et al., 2010), a dynamic that static vignettes cannot model. Furthermore, clinical reasoning, whether human or artificial, is vulnerable to cognitive biases that compound these challenges (Vally et al., 2023). In realistic scenarios, LLMs exhibit inconsistent risk-stratification (Heston and Lewis, 2024), fail to follow diagnostic guidelines

[1]ICD-10 (Figure 3): 10th revision of the International Classification of Diseases, maintained by the (WHO) World Health Organization

during incremental information gathering (Hager et al., 2024), and alter factual clinical outputs based on patient linguistic markers (Kearney et al., 2025; Zhou et al., 2025). As a medical direct-to-consumer offering, ChatGPT Health undertriaged up to 52% of emergencies and shifted recommendations when patients minimized their symptoms (Ramaswamy et al., 2026). The questions MTDiag answers are therefore not only *does the model arrive at the right diagnosis?* but more importantly: *can the model conduct a diagnostic conversation well enough to reach the right diagnosis?*

Static Medical QA Benchmarks The standard for assessing the "clinical aptitude" of LLMs (and making the case for them as "medical assistants") has mostly relied on increasingly strong results on static QA benchmarks such as PubMedQA (Jin et al., 2019), MedQA (Jin et al., 2021), and MedMCQA (Pal et al., 2022), which aggregate questions from medical licensing exams and literature. Questions on medical exams are self-contained: with adequate knowledge, everything required for the answer is in the question formulation. This is far from real patient records: Because they test recall under ideal conditions, correctly answering a question when presented with a QA vignette does not predict correct performance on similar real-world cases (Alaa et al., 2025; Gong et al., 2025). The validity of these benchmarks is further undermined by data contamination and leakage (Balloccu et al., 2024; Xu et al., 2024), where test items appearing in pretraining corpora inflate reported scores.

Conversational Diagnostic AI Recent systems have begun probing LLMs in interactive clinical settings. AMIE (Tu et al., 2025) and CRAFT-MD (Johri et al., 2024) demonstrate that models capable of near-perfect performance on static benchmarks degrade substantially when required to elicit information across a multi-turn encounter. AI Hospital (Fan et al., 2025) and the JAMA multi-agent framework (Sangwon et al., 2025) further surface reliability and consistency failures in interactive settings. Arias-Duart et al. (2025) propose automatic evaluation of healthcare LLMs beyond question-answering, and implement a range of sub-tasks, motivating the need for dialogue-level assessment.

Medical Dialogue Datasets Several medical dialogue datasets have been introduced to capture multi-turn dynamics, most recently surveyed by (Gong et al., 2025). We describe the datasets we considered and their suitability in Section 3.2.

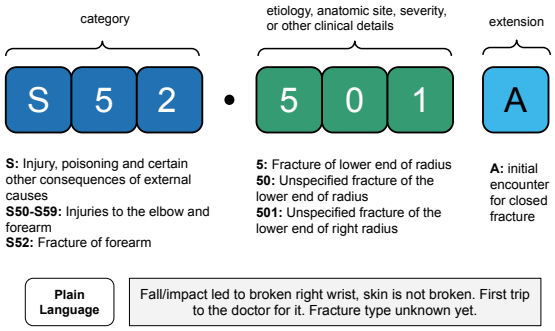

Figure 3: ICD-10-CM code structure illustrated with `S52.501A` (unspecified fracture of the lower end of the right radius, initial encounter for closed fracture). Each position encodes a specific clinical dimension: category, anatomic site/etiology, severity, and encounter extension. The WHO ICD-10 equivalent without the CM extension would be `S52.5`.

3 Ontological Anchoring and Resource Choices

In this section, we discuss the design decisions behind the construction of MTDiag, which were directly informed by practicing medical professionals and by surveying well-established clinical behavioral literature, taking into account a space of logical and practical errors that can occur in a multi-turn differential diagnosis setting. We then introduce and motivate some of the metrics that ontological anchoring enables.

Minimum Anchors We first define three necessary elements (which we refer to as our "Minimum Anchors") for the conduction and evaluation of a diagnostic dialogue:

- The **chief complaint (CC)**: the patient's primary reason to seek care, according to "them." This serves as the opening utterance of the dialogue (turn 0).
- **Symptoms**: the patient's other symptoms.
- The **primary diagnosis**: the ground truth principal finding of the case, as ICD-10 code (Fig. 3). An ICD-10 code match constitutes the minimum evaluation target: did the model reach the right diagnosis?

Dialogue Trace To illustrate why this minimum evaluation target of Diagnosis Accuracy (Diag_{acc}) fails to completely capture clinical competence, we trace two example dialogues, shown in Figure 4. Both LLM examiners receive the same opening patient utterance: *"I've had this headache for about three weeks. It won't go away."* The ground-truth diagnosis is Giant Cell Arteritis (GCA, `M31.6`), a serious inflammatory disease of the arteries. In **Dialogue A**, the examiner probes for scalp tenderness, jaw claudication, and visual changes. After

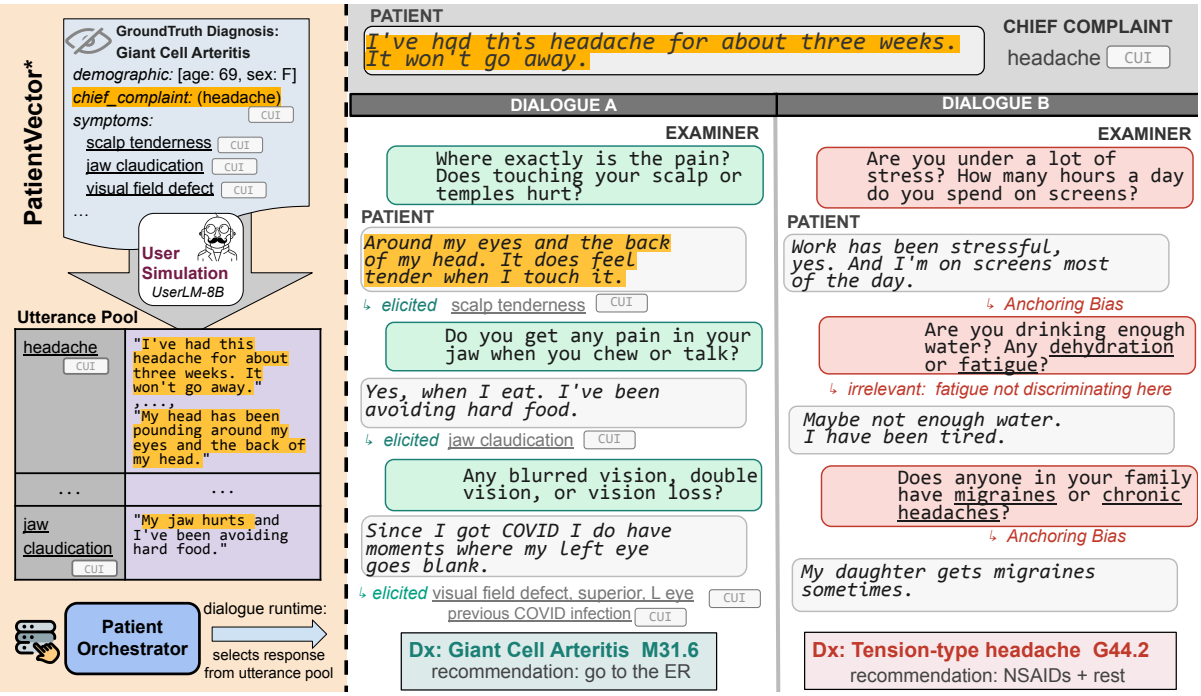

Figure 4: (Condensed) example of ontology-enabled dialogue evaluation - MTDiag AJCR case. Dx: Giant Cell Arteritis (GCA, M31.6) (Szydełko-Paśko et al., 2022). **Left** (of dashed line): the `PatientVector` * for this case, comprising the ground-truth diagnosis (masked from examiner) and UMLS `CUI`-mapped symptoms (subset only). Utterances are pre-generated offline by UserLM-8B for each element of the `PatientVector`, constituting the `CUI`-mapped utterance pool. At dialogue runtime, a Patient Orchestrator LLM answers examiner questions from the utterance pool. **Right:** Dialogue runtime: Two dialogues conducted from the same chief complaint utterance. **Dialogue A**: examiner probes for scalp tenderness, jaw claudication, and visual field loss and reaches the correct diagnosis. **Dialogue B**: examiner anchors prematurely on migraine/tension headache, asks only about stress, screen time, hydration, and family history, and diagnoses tension-type headache. A **Harm Tier 2** error (delay of care) occurs. *simplified

the patient confirms these symptoms, the examiner correctly diagnoses GCA. In **Dialogue B**, the examiner immediately anchors on a migraine hypothesis, and asks only questions that reinforce that initial hypothesis (stress, screen time, hydration, and family history), ultimately diagnosing a Tension-Type Headache (`G44.2`) and completely missing GCA.

Dialogue Metrics In addition to a classification error, which is trivially detected by ICD-10 mismatch, **Dialogue B** has two further issues. The examiner displays ***anchoring bias*** (a line of questioning that serves only to confirm an initial hypothesis rather than distinguish it from alternatives) towards migraine/tension-type headache. `CUI` anchoring makes this detectable by mapping the examiner questions to the symptom cluster for migraine, with no questions mapped to the cluster for GCA. Another more severe issue with Dialogue B is that it ultimately leads to a ***Harm Tier 2*** error (delay of care), where the incorrect diagnosis leads to delay in seeking treatment for GCA, which can cause irreversible bilateral blindness if left untreated. A full discussion of the **Harm Index** and formal definitions of other ontologically-enabled diagnostic dialogue metrics follows in Section 5.

3.1 Patient-Chatbot Realism

Assessing LLMs in user-facing diagnosis means the "patient" can realistically communicate their symptoms via a chat interface, as an alternative to seeking medical care. This rules out patients who are unconscious, intubated, or in agonizing pain. These and more considerations guide an extensive pre-processing of the datasets detailed in Section 4.1. Similarly, the need to realistically simulate a user interacting with an LLM rather than speaking to a doctor led us to employ UserLM-8B (Naous et al., 2025) for utterance generation, detailed in Section 4.3. Unlike standard instruction-tuned, "assistant-style" models, UserLM-8B is trained on user-chats with ChatGPT and optimized for "user-simulation," allowing colloquial, incomplete expression with the natural hedging and imprecision characteristic of lay speech.

3.2 Source Datasets

Three sources were ultimately selected for MTDiag (full discussion in Appendix A) to cover complementary regions of the diagnostic space: a structured synthetic head, a real high-frequency ED

cohort, and a long-tail of rare and atypical case reports.

DDXPlus (Fansi Tchango et al., 2022) is a large-scale synthetic dataset for differential diagnosis comprising 1.3 million patient cases across 49 pathologies and 223 evidences (symptoms and antecedents). Each record includes a differential diagnosis list with associated probabilities and a ground-truth label, as well as associated symptoms for all covered diseases. **MIMIC-IV (Johnson et al., 2023)** is a large, de-identified electronic health record database from the Beth Israel Deaconess Medical Center covering over 200,000 ED (Emergency Department) visits and 65,000 ICU (Intensive Care Unit) stays (2008–2019). It provides real, high-frequency emergency presentations grounded in verified ICD-10 discharge diagnoses, augmented by community extensions supplying narrative HPI text, structured laboratory results, and radiology reports. **AJCR Case Reports (Hirosawa et al., 2024)** is a curated compilation of 392 adult case reports from the *American Journal of Case Reports* (January 2022–March 2023), with clearly stated final diagnoses. Case reports are selected precisely because they document rare, atypical, or diagnostically challenging presentations (the long tail of diagnostic space that DDXPlus and MIMIC-IV do not cover). Malignancies, diverse infections, and vascular diseases account for over 60% of final diagnoses.

4 Methodology

The construction of MTDiag is illustrated in Figure 5. First, (Figure 5**.A**; Section 4.1) the three source datasets are collected and preprocessed. Then, (Figure 5**.B**; Section 4.2) the clinical evidence anchors (symptoms, diagnoses, medications...) are resolved to canonical ontological identifiers via a two-stage NER-to-UMLS pipeline, producing an annotated representation of each case. Each case is then normalized into the `PatientVector` unified schema, then (Figure 5**.C**; Section 4.3) the clinical evidence in the `PatientVector` is converted offline into natural-language patient utterances.

4.1 Data Collection & Preprocessing

The three source datasets introduced in Section 3.2 are structurally heterogeneous but were selected to satisfy the minimum anchor requirements defined in Section 3. **Chief complaint (CC)**, **symptoms** and **primary diagnosis** are present (or extractable), well-defined, and verified. However, the source datasets encode significantly more information than the minimum anchors. **Secondary diagnoses** are additional diagnoses recorded for the same case, capturing comorbidities and incidental findings. These are used to assess diagnostic completeness and to flag cases where the model's diagnosis, while not matching the primary diagnosis exactly, is clinically consistent with the documented picture and may match one or more of the secondary diagnoses. Medical history antecedents, narrative HPI, medications, lab tests, demographic and socioeconomic features, and more are also present with varying sparsities. The `PatientVector` (Fig. 6) schema is the canonical form normalizing all three sources, discussed in Section 4.2.

Source	Type	Scale (cases)
DDXPlus	Synthetic	1.3M
MIMIC-IV	Real EHR	68K
AJCR	Case reports	415

Table 1: Data sources used in MTDiag after filtering and deduplication (detailed in Appendix B.2.2).

Below, we document the key curation and preprocessing decisions for each source, with full pipeline details in Appendix B.

4.1.1 DDXPlus

Synthesized from a proprietary Automatic Disease Diagnosis (ADD) system, DDXPlus encodes structured clinical evidence (symptoms, antecedents) under a closed-world assumption, with each case carrying a differential diagnosis list and ground-truth label. Its value for MTDiag lies in its differential trajectories and structured evidence instantiation. One preprocessing choice worth mentioning here: each patient's `INITIAL_EVIDENCE` field (a randomly selected binary evidence with no inherent clinical salience in the original ADD setup) is retained as the CC for consistency.

4.1.2 MIMIC-IV

Unlike typical uses of MIMIC for longitudinal time-series applications, our use case requires coherent narrative patient presentations. We draw on four official modules (Hosp, ED, Note) and two community extensions (BHC hospital course narratives; CDM curated abdominal pathology cases).

Population selection The target population is patients who walk into an ED of their own volition and could plausibly interact with an LLM instead. From 425,087 raw ED visits, we sequentially exclude ambulance arrivals, visits with pain score above 8/10 or ESI acuity level 1, cases with no primary diagnosis (etc.), and deal with free-text

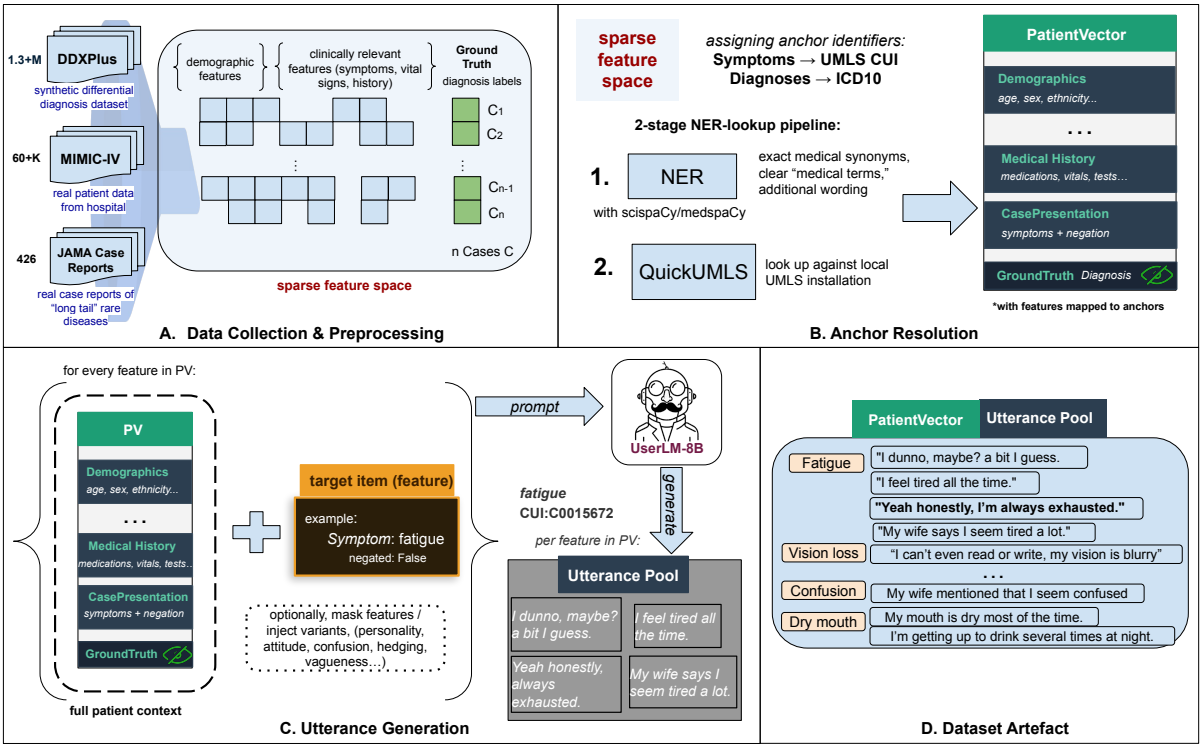

Figure 5: Construction of MTDiag **A. Data Collection & Pre-Processing**: the three source datasets (DDXPlus, MIMIC-IV, AJCR) and their heterogeneous sparse feature spaces, preprocessed to satisfy the minimum anchor requirements and patient realism. **B. Anchor Resolution**: clinical strings from all three sources are segmented by `medspaCy` NER and matched against a local UMLS installation via `QuickUMLS` (Soldaini and Goharian, 2016), yielding a `CUI` and `mapping_confidence` score per span. Diagnoses follow a separate ICD-10 crosswalk path. The pipeline runs fully locally against a fixed UMLS release, producing reproducible, verifiable mappings with no external API calls. **C. Utterance generation:** For each item in `CasePresentation`, UserLM-8B is conditioned on `PatientVector` context and generates candidate utterances spanning expression styles (precise, colloquial, third-party); masking and persona parameters or noise injections can be injected at generation time. Elements of the `PatientVector` context can be selectively masked to produce presentation variants. **D. Dataset Artefact:** The **dataset proper** consists of the `PatientVector`, and the **dataset artefact**, is the result of utterance generation over the `PatientVector`. This artefact is structured, reproducible, and fully decoupled from the dialogue runtime described in Section 4.4.

entries for some numeric values as "extra context" which we preserve. The preprocessing details are in Appendix B.2.2.

Two-track design After filtering, the cohort partitions into two clinically distinct populations. **Track 1 (ED Discharge):** The patient is admitted into the Emergency Department, then discharged. Only ED data is used (chief complaint, vitals, questionnaire); ground truth is the ED diagnosis.
Track 2 (Hospital Admission): The patient is admitted into the ER, evaluated, and then admitted to the hospital for further work-up. This augments triage (ED) data with the History of Present Illness and, where available, granular structured clinical data from the BHC and CDM extensions. The CDM extension alone contributes pre-extracted HPI narratives, physical examination findings, 138,788 laboratory results across 480 unique tests, 4,403 microbiology results, and 5,959 radiology reports (CT, X-ray, ultrasound, MRI) with diagnostic conclusions explicitly stripped from radiology findings and cases where the pathology name appears in the HPI excluded, making these narratives genuinely suitable for diagnostic simulation without leakage.

Deduplication A hierarchical strategy retains one visit per patient (to prevent "frequent flyers" from saturating the baseline), prioritized by data richness (CDM + hospital course > CDM only > hospital course only > earliest visit). The final subset comprises **68,346 unique patients**: 49,440 ED-discharged (Track 1) and 18,906 admitted (Track 2).

Zero Data Retention As MIMIC-IV consists of de-identified yet real patient data, its use is subject to PhysioNet guidelines and processing is only permitted under ZDR (Zero Data Retention). As such, all processing runs on the GWDG KISSKI secure HPC.

4.1.3 AJCR Case Reports

We developed a general-purpose scraping and parsing pipeline for AJCR articles: given a DOI, it

retrieves full-text HTML and extracts structured metadata, abstracts, section-split body text, figures, and tables into a per-case JSON schema, with automatic detection and splitting of case series. We apply it here to the Hirosawa et al. (Hirosawa et al., 2024) compilation precisely because it provides pre-curated adult diagnostic cases with verified final diagnoses, yielding 415 cases from 392 reports (see Table 1). The pipeline is applicable to any AJCR DOI.

4.2 Normalization and Anchor Resolution

4.2.1 `PatientVector` Unified Case Schema

To unify the three heterogeneous sources into a single structure, we define the `PatientVector` (Figure 6), a canonical per-case data schema that all sources are normalized into prior to dialogue generation. The schema comprises four top-level blocks:

Provenance records the origin of each case: a globally unique `mtdiag_id`, the source dataset (`ddxplus`, `mimic`, or `ajcr`), the native identifier in that source, and pipeline metadata. This block enables stratified evaluation and ablation by source.

GroundTruth is populated at ingestion and never exposed to the dialogue generation pipeline or the examiner. It contains the primary `Diagnosis` (ICD-10 code, UMLS `CUI`, and human-readable label), a list of secondary `Diagnosis` entries, and, where available, a ranked differential diagnosis list. Each diagnosis carries a `provenance_tier` field recording the reliability of the ground-truth label (distinguishing, for instance, a verified hospital discharge diagnosis (`mimic_discharge`) from a preliminary ED assessment (`mimic_ed_prelim`)) (additional details in Appendix B.5).

PatientFacts holds all patient information available beyond the presenting complaint: demographics (e.g., age, sex, race/ethnicity, nationality, and socioeconomic proxies like insurance, marital status), vitals (blood pressure, BMI, height), test results (`CUI`-mapped, with value, unit, reference range, and relative date), narrative text (HPI and free-text notes), family history, and current medications (`CUI`, name, and dosage). The sparsity varies by source: MIMIC-IV Track 2 cases with records in the CDM and/or BHC extensions provide data for most fields, while DDXPlus cases carry only some demographics and binary/categorical evidence.

CasePresentation is the snapshot of the patient as they initiate the dialogue. It contains the chief complaint: one or more `SymptomPresentation` entries forming the opening utterance and the full symptom list. Each `SymptomPresentation` pairs a `Symptom` (`CUI` and string) with an `Experience` capturing temporality (onset, frequency, duration), locality (`CUI`-mapped body region), characterization (severity scale and/or free-text descriptor), and aggravating/relieving factors. The full field-level specification of all classes is provided in Appendix B.5.

4.2.2 Anchor Resolution

The core design feature of MTDiag is that all clinical anchors (and additional patient facts) are mapped to established canonical identifiers: the chief complaint, symptoms, and other details where possible (lab tests, medications...) are mapped to UMLS `CUIs` (Fig. 2); diagnoses are mapped to ICD-10 codes (Fig. 3). A UMLS `CUI` situates a concept within a graph of over 200 integrated clinical vocabularies. Rather than asking whether two symptom strings match, one can ask how far apart two concepts are in the UMLS graph (Fig. 2), whether one subsumes the other, or whether they share a common parent, resulting in a graded, clinically meaningful similarity measure. It also enables crosswalk between vocabularies: a symptom expressed in lay language and its formal medical term (in any of the included vocabularies) resolve to the same `CUI`. Mapping the heterogeneous and sparse feature spaces of our source datasets to `CUIs` proved non-trivial; both LLM-based and UMLS API string search-based resolution proved unsuitable for large-scale annotation; we discuss these failure modes in Appendix B.4. Our approach (Figure 5.B) relies on a 2-stage NER-based pipeline. In the first stage, clinical named entity recognition is performed using `medspaCy` (Eyre et al., 2021), a spaCy-based framework for clinical text processing, which segments input text and identifies candidate medical concept spans. In the second stage, each candidate span is passed to `QuickUMLS` (Soldaini and Goharian, 2016), a fast approximate string matching engine built over a local UMLS installation, which returns the best-matching `CUI` along with a similarity score retained as the `mapping_confidence`. All additional non-anchor patient information is mapped back to either `CUI` identifiers, tightly structured enumeration options, or numeric scale representations.

4.3 Utterance Generation

The `PatientVector` schema is designed to support a spectrum of patient variability. To convert the structured symptom representations into natural

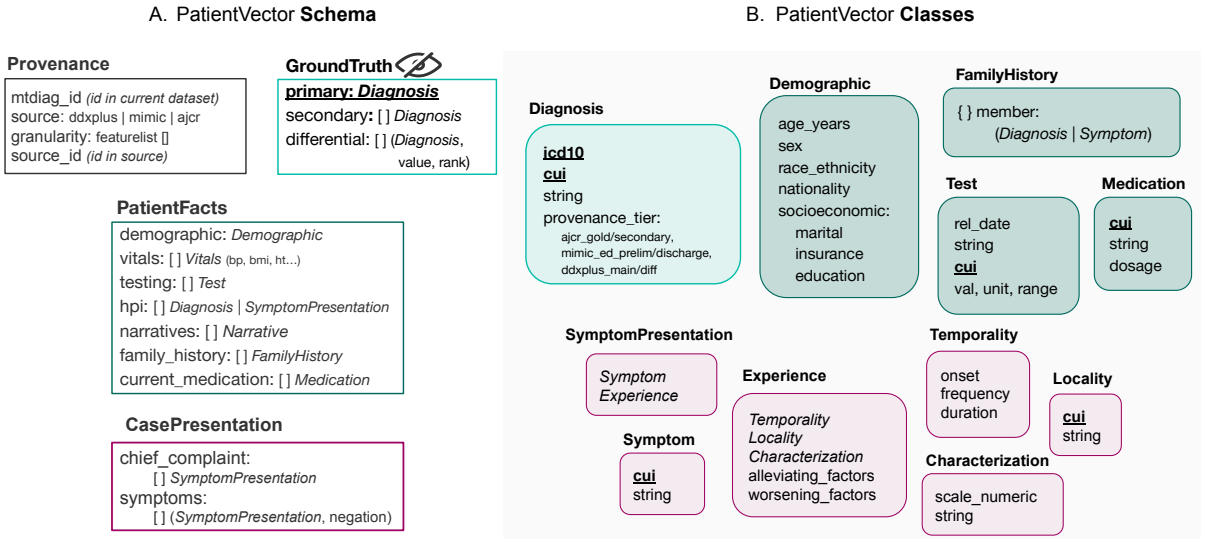

Figure 6: `PatientVector` schema. *Left*: the four top-level blocks: Provenance, GroundTruth (masked), PatientFacts, and CasePresentation, with their fields. *Right*: the full class hierarchy, showing how SymptomPresentation decomposes into Symptom (`CUI` + string) and Experience (Temporality, Locality, Characterization), and how Diagnosis carries ICD-10 code, `CUI`, and `provenance_tier`. All fields are serialized to JSON; sparsity varies by source. *ICD-10 is also indexed in the UMLS, so every ICD-10 code resolves to a `CUI`.

language, we use UserLM-8B (Naous et al., 2025), a model trained to produce human-like *user* utterances (Figure 5.C). We provide a set of canonical utterances per `PatientVector`, which can be edited and re-generated via this pipeline. All utterances are pre-generated, stored, and fully decoupled from live dialogue. The patient's "factual reality" is fixed at dataset creation time, rather than simulated during a dialogue. This "separation of concerns" ensures reproducibility, prevents the patient simulator from hallucinating new symptoms, and allows human review of the utterance pool.

4.4 Dialogue Runtime

At dialogue runtime, two agents interact (illustrated in Fig. 8 in the Appendix). The **Patient Orchestrator**, powered by MedGemma (Sellergren et al., 2025), receives the examiner's most recent question and selects which pre-generated utterance or combination of utterances to respond with. MedGemma was chosen due to being the only medical-specialized and instruction-tuned model available via GWDG ChatAI, our ZDR-compliant inference provider, but it is possible to use any assistant LLM as Patient Orchestrator. The orchestrator draws exclusively from the pre-generated utterance pool, ensuring responses remain grounded in case facts. This design deliberately separates *content* (what the patient says, grounded and `CUI`-anchored) from *selection logic* (which utterance to surface given the conversation), as a middle ground between the rigidity of rule-based response selection and the hallucination risk of free-form patient simulation. The **Examiner** is the model under evaluation. It receives only the chief complaint at Turn 0 and must issue follow-up questions across subsequent turns to gather enough evidence for a diagnosis. The dialogue terminates when the examiner produces a terminal diagnosis or a maximum turn limit is reached. Every turn is logged to a structured JSON record: prompts, raw text, examiner question, patient utterance selected, and `CUIs` extracted from both sides. CUI extraction applied to examiner outputs proceeds using the same two-stage pipeline detailed in Section 4.2, and enables the metrics described in Section 5 over ontological identifiers, rather than surface forms or string matching.

5 Discussion

The primary advantage of MTDiag is its size, variety of sources, and well-defined schema and strict ontological anchoring, which allows editing, masking, noise injection, and generation of different "patient profiles" from a single case. With that, it allows exploration of different areas of the search space.

Demographic Fairness Separating clinical facts from the `Demographic` block in the `PatientVector` allows selective masking or altering of variables (e.g., age, race, insurance) during utterance generation to detect systemic undertriage or biased diagnostic trajectories for otherwise identical clinical

presentations.

Multilingual Assessment Because UMLS `CUI` identifiers are language-agnostic, `PatientVector` strings can be swapped into any of 31 supported languages. Though initially English-focused, the schema natively supports multilingual dataset generation and dialogue simulation (via language-specific UserLMs) while preserving clinical ground truth.

Long-Tail Diagnostics Stratifying by source tests LLM robustness across the frequency spectrum: from routine, high-frequency ED triage (MIMIC-IV) to rare, complex cases (AJCR). Subsetting by specific ground-truth diagnoses further enables targeted, disease-specific "LLM clinical aptitude" assessments.
In addition, because every symptom in the `PatientVector` is mapped to a UMLS `CUI`, and every diagnosis to an ICD-10 code, we move beyond surface-level string matching. This allows us to automatically compute clinically meaningful metrics from dialogue logs and systematically evaluate how LLMs reason, where they fail, and how they interact. Using the `CUI` logs extracted from the examiner's questions at each turn, we propose the following metrics (Figure 4).

Semantic Diagnostic Distance Beyond binary accuracy (Diag_{acc}), ICD-10 anchoring allows us to measure categorical severity. `G44.2` (Chapter G, nervous system) and `M31.6` (Chapter M, musculoskeletal system) sit in entirely different clinical chapters. A model outputting `I21.0` (acute myocardial infarction) when the ground truth is `I21.02` (STEMI) commits a minor underspecification; Dialogue B commits a severe categorical miss.

Symptom Elicitation Score (S_{dise}) This metric assesses whether the model elicited the clinical evidence required for the diagnosis. Let S_d be the set of symptoms associated with disease d (drawn from the Human Phenotype Ontology, HPO), and S_{diag} be the `CUI`-resolved set of symptoms the examiner inquired about. We calculate a frequency-weighted relevance score:

$$f_w(s, S_d) = \omega(s, d) \cdot \mathbf{1}[s \in S_d]$$

where $\omega(s, d) \in \{1.0, 0.75, 0.5, 0.25, 0.1\}$ corresponds to HPO frequency qualifiers (from *obligate* to *very rare*).

Reliability and Bias A model that reaches the correct diagnosis without eliciting enough evidence, or its *sine qua non* (strictly necessary) or pathognomonic (strictly sufficient) symptoms, commits a **diagnostic hallucination**. `CUI` anchoring makes logical errors automatically computable:

- **Reliability Score (R_{score}):** A model is credited only if it reached the correct diagnosis *and* its symptom elicitation score (S_{dise}) exceeds a minimum threshold.
- **Anchoring Bias:** Queried symptoms are exclusively mapped to a single incorrect candidate diagnosis (e.g., Dialogue B asking only migraine-associated questions).
- **Premature Closure:** A model commits to a terminal diagnosis before crossing a minimum threshold of necessary symptom elicitation for the diagnosis.

The Harm Index Diagnostic errors carry varying risk. Harm is clinically categorized into three tiers:

- **Tier 1 (Direct Harm):** Recommending a treatment contraindicated for the ground-truth diagnosis.
- **Tier 2 (Delay of Care):** Diagnosing a low-acuity condition when the ground truth is a time-sensitive emergency (e.g., Dialogue B missing GCA, risking irreversible bilateral blindness).
- **Tier 3 (Instructional Failure):** Omitting critical advice that the case warrants.

Importantly, these metrics are not specific to MTDiag, rather the ontology-grounded metric suite is a dialogue-level evaluation protocol that any system producing CUI-annotated logs can be scored against. More details and additional metrics are discussed in Appendix E.

MTDiag's grounding in clinical data with structured ontological anchoring and decoupled utterance generation means the evaluation substrate is a generative pipeline over structured cases, not a set of static files that can be leaked/memorized. With this work we aim to motivate the design of transparent, reproducible, and knowledge-grounded benchmarks for critical dialogue-based tasks, and continued research into interoperable knowledge bases like the UMLS. Code and dataset available via `github.com/piachouaifaty/MTDiag`.

Limitations

The UMLS as a Living Resource Ontological anchoring is both a significant asset and a source

of experimental overhead. Our development, experiments, and working pipeline are tagged to the 2025AB release of the UMLS. Naturally, rerunning the anchor resolution pipeline on a different UMLS installation may result in different results, or `CUI` assignments, as the UMLS is a regularly updated, living resource. Although this may introduce setup overhead, it is also advantageous in that one may perform a bespoke UMLS installation, and include/exclude certain vocabularies for compatibility with specific health systems.

Data Governance and Release Constraints MIMIC-IV, though formally de-identified, is considered potentially re-identifiable under PhysioNet's Data Use Agreement, which mandates Zero Data Retention (ZDR) and prohibits sharing with third parties. MIMIC-derived portions of MTDiag will be released exclusively via PhysioNet under the standard credentialed access agreement, and all processing and evaluation involving MIMIC data must be conducted in a ZDR-compliant environment.

AJCR Pretraining Contamination Case reports from the AJCR series are widely indexed and freely available online, meaning some portion is likely present in the pretraining corpora of evaluated models. We treat this as a point of inquiry rather than a disqualifier: the structured `PatientVector` generated from each report produces dialogue substantially different from the original prose, and future work can directly test whether models perform better on this subset relative to others.

Utterance Pool Coverage The current release provides a set of canonical utterances per symptom per `PatientVector` entry. This is a conservative starting point: real patients vary enormously in how they describe the same symptom. The pipeline supports full variant expansion across persona parameters, masking policies, health literacy levels, and noise injections (described in Appendix C), but generating and validating this expansion at scale is left to future work.

Multi-Model Evaluation This work introduces the pipeline and evaluation framework, and provides a validated dataset. Systematic multi-model "as an examiner" evaluation is the natural next step and is left to future work.

Patient-Chatbot Realism Revisited As described in Section 3.1, we take steps to ensure the realism of our selected cases and design. However, the necessity of crafting system prompts properly instructing the LLM to act like a medical professional is far removed from the reality of a user prompting an LLM. In real life, a user will simply begin a conversation, rarely setting careful inference parameters, and may have additional, unrelated context, and, potentially, user-defined settings that alter outputs. A completely realistic assessment of LLMs as medical assistants to real users would require a large-scale ontological mapping of highly unstructured, real user chatlogs and the application of similar metrics as the ones we introduce.

Ethical Considerations

It is important to point out that the hard technical limits of LLM context windows and their "as-assistant" sycophantism combined with the tendency of patients to be imperfect subjects (hedging, minimizing, re-framing) (Redelmeier et al., 2001; Ijäs-Kallio et al., 2010) calls into question the suitability of LLMs for long-range diagnostic tasks. Physicians undergo years of study, specialized training, mentorship, and practical clinical experience, and receive input from other similarly qualified doctors, nurses, and practitioners while considering a diagnosis. Moreover, they are human beings with a quasi-unlimited "context window," equipped with empathy, intuition, and a spirit of discernment. Most importantly, doctors take an oath to do no harm. We therefore present this work as a more rigorous LLM-as-diagnostician evaluation paradigm compared to existing approaches, and highly encourage the use of MTDiag for systematic multi-model evaluation in order to identify LLM failure modes and carefully consider their autonomous deployment in situations that may lead to the harming or loss of human life.

Acknowledgments

The authors gratefully acknowledge the computing time granted by Federal Ministry of Education and Research (BMBF) project AI service center KISSKI (Grant N. 01IS22093A-E). The calculations for this research were conducted with computing resources under the project *Towards a Realistic Multi-turn Medical Dialogue Benchmark for LLMs*. The authors also thank Dr. Maja Miličić Brandt for her ontology expertise, illuminating discussions around the UMLS, and meticulous feedback on the

manuscript.

References

Ahmed Alaa, Thomas Hartvigsen, Niloufar Golchini, Shiladitya Dutta, Frances Dean, Inioluwa Deborah Raji, and Travis Zack. 2025. Position: Medical large language model benchmarks should prioritize construct validity. In *Forty-second International Conference on Machine Learning Position Paper Track*.

Anna Arias-Duart, Pablo Agustin Martin-Torres, Daniel Hinjos, Pablo Bernabeu-Perez, Lucia Urcelay Ganzabal, Marta Gonzalez Mallo, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Sergio Alvarez-Napagao, and Dario Garcia-Gasulla. 2025. Automatic evaluation of healthcare llms beyond question-answering. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 108–130.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Rachel L Draelos, Samina Afreen, Barbara Blasko, Tiffany L Brazile, Natasha Chase, Dimple Patel Desai, Jessica Evert, Heather L Gardner, Lauren Herrmann, Aswathy Vaikom House, and 1 others. 2026. Large language models provide unsafe answers to patient-posed medical questions. *NPJ digital medicine*, 9(1):241.

H. Eyre, A. B. Chapman, K. S. Peterson, J. Shi, P. R. Alba, M. M. Jones, T. L. Box, S. L. DuVall, and O. V. Patterson. 2021. Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu Symp Proc*, 2021:438–447.

Zhihao Fan, Lai Wei, Jialong Tang, Wei Chen, Wang Siyuan, Zhongyu Wei, and Fei Huang. 2025. AI hospital: Benchmarking large language models in a multi-agent medical interaction simulator. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10183–10213, Abu Dhabi, UAE. Association for Computational Linguistics.

Arsene Fansi Tchango, Rishab Goel, Zhi Wen, Julien Martel, and Joumana Ghosn. 2022. Ddxplus: A new dataset for automatic medical diagnosis. *Advances in neural information processing systems*, 35:31306–31318.

Eun Jeong Gong, Chang Seok Bang, Jae Jun Lee, and Gwang Ho Baik. 2025. Knowledge-practice performance gap in clinical large language models: systematic review of 39 benchmarks. *Journal of Medical Internet Research*, 27:e84120.

Paul Hager, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis, and 1 others. 2024. Evaluation and mitigation of the limitations of large language models in clinical decision-making. *Nature medicine*, 30(9):2613–2622.

Thomas F Heston and Lawrence M Lewis. 2024. Chatgpt provides inconsistent risk-stratification of patients with atraumatic chest pain. *PLoS One*, 19(4):e0301854.

Takanobu Hirosawa, Yukinori Harada, Kazuki Tokumasu, Takahiro Ito, Tomoharu Suzuki, and Taro Shimizu. 2024. Comparative study to evaluate the accuracy of differential diagnosis lists generated by gemini advanced, gemini, and bard for a case report series analysis: cross-sectional study. *JMIR Medical Informatics*, 12:e63010.

Taru Ijäs-Kallio, Johanna Ruusuvuori, and Anssi Peräkylä. 2010. Patient resistance towards diagnosis in primary care: Implications for concordance. *Health:*, 14(5):505–522.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2567–2577.

Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Brian Gow, Benjamin Moody, Steven Horng, Leo Anthony Celi, and Roger Mark. 2024. MIMIC-IV. *PhysioNet*. Version 3.1.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1.

Shreya Johri, Jaehwan Jeong, Benjamin A Tran, Daniel I Schlessinger, Shannon Wongvibulsin, Zhuo Ran Cai, Roxana Daneshjou, and Pranav Rajpurkar. 2024. Craft-md: A conversational evaluation framework for comprehensive assessment of clinical llms. In *AAAI 2024 Spring Symposium on Clinical Foundation Models*.

Matthew Kearney, Reuben Binns, and Yarin Gal. 2025. Language models change facts based on the way you talk. *arXiv preprint arXiv:2507.14238*.

Eyal Klang, Idit Tessler, Donald U Apakama, Ethan Abbott, Benjamin S Glicksberg, Monique Arnold, Akini Moses, Ankit Sakhuja, Ali Soroush, Alexander W Charney, and 1 others. 2024. Assessing retrieval-augmented large language model performance in emergency department icd-10-cm coding compared to human coders. *medRxiv*.

Xinzhu Lin, Xiahui He, Qin Chen, Huaixiao Tou, Zhongyu Wei, and Ting Chen. 2019. Enhancing dialogue symptom diagnosis with global attention and symptom graph. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5033–5042, Hong Kong, China. Association for Computational Linguistics.

Wenge Liu, Jianheng Tang, Yi Cheng, Wenjie Li, Yefeng Zheng, and Xiaodan Liang. 2022. MedDG: An entity-centric medical consultation dataset for entity-aware medical dialogue generation. *arXiv preprint arXiv:2010.07497*.

Srija Macherla, Man Luo, Mihir Parmar, and Chitta Baral. 2023. Mddial: A multi-turn differential diagnosis dialogue dataset with reliability evaluation. *arXiv preprint arXiv:2308.08147*.

Tarek Naous, Philippe Laban, Wei Xu, and Jennifer Neville. 2025. Flipping the dialogue: Training and evaluating user language models. *arXiv preprint arXiv:2510.06552*.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Martin Krallinger, Miguel Rodríguez-Ortega, Eduard Rodriguez-López, Natalia Loukachevitch, Andrey Sakhovskiy, Elena Tutubalina, Dimitris Dimitriadis, and 1 others. 2025. Overview of bioasq 2025: The thirteenth bioasq challenge on large-scale biomedical semantic indexing and question answering. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 173–198. Springer.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Ashwin Ramaswamy, Alvira Tyagi, Hannah Hugo, Joy Jiang, Pushkala Jayaraman, Mateen Jangda, Alexis E Te, Steven A Kaplan, Joshua Lampert, Robert Freeman, and 1 others. 2026. Chatgpt health performance in a structured test of triage recommendations. *Nature Medicine*, pages 1–5.

Donald A Redelmeier, Michael J Schull, Janet E Hux, Jack V Tu, and Lorraine E Ferris. 2001. Problems for clinical judgement: 1. eliciting an insightful history of present illness. *Cmaj*, 164(5):647–651.

Karl L Sangwon, Jeff Zhang, Robert Steele, Jaden Stryker, Jin Vivian Lee, Joanne Choi, Krithik Vishwanath, Daniel Alexander Alber, Douglas Kondziolka, Michal Mankowski, and 1 others. 2025. Evaluating large language model diagnostic performance on jama clinical challenges via a multi-agent conversational framework. *medRxiv*, pages 2025–08.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

A Simmons, K Takkavatakarn, M McDougal, B Dilcher, J Pincavitch, L Meadows, J Kauffman, E Klang, R Wig, G Smith, A Soroush, R Freeman, D. J. Apakama, A. W. Charney, R Kohli-Seth, G. N. Nadkarni, and A Sakhuja. 2025. Extracting international classification of diseases codes from clinical documentation using large language models. *Applied Clinical Informatics*, 16(2):337–344.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction.

Amnon Sonnenberg and Howard K Gogel. 2002. Translating vague complaints into precise symptoms: the implications of a poor medical history. *European journal of gastroenterology & hepatology*, 14(3):317–321.

Urszula Szydełko-Paśko, Joanna Przeździecka-Dołyk, Julia Kręcicka, Rafał Małecki, Marta Misiuk-Hojło, and Anna Turno-Kręcicka. 2022. Arteritic anterior ischemic optic neuropathy in the course of giant cell arteritis after covid-19. *The American Journal of Case Reports*, 23:e933471–1.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, and 1 others. 2025. Towards conversational diagnostic artificial intelligence. *Nature*, 642(8067):442–450.

Zunaid Ismail Vally, Razia AG Khammissa, Gal Feller, Raoul Ballyram, Michaela Beetge, and Liviu Feller. 2023. Errors in clinical diagnosis: a narrative review. *Journal of International Medical Research*, 51(8):03000605231162798.

Juraj Vladika, Alexander Fichtl, and Florian Matthes. 2026. Investigating expectations and needs regarding the use of large language models at bavarian university clinics. *Scientific Reports*.

Zhongyu Wei, Qianlong Liu, Baolin Peng, Huaixiao Tou, Ting Chen, Xuanjing Huang, Kam-fai Wong, and Xiangying Dai. 2018. Task-oriented dialogue system for automatic diagnosis. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 201–207, Melbourne, Australia. Association for Computational Linguistics.

Lin Xu, Qixian Zhou, Ke Gong, Xiaodan Liang, Jianheng Tang, and Liang Lin. 2019. End-to-end trainable non-collaborative dialog system for automatic diagnosis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7346–7353.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Guojun Yan, Jiahuan Pei, Pengjie Ren, Zhaochun Ren, Xin Xin, Huasheng Liang, Maarten de Rijke, and Zhumin Chen. 2021. ReMeDi: Resources for multi-domain, multi-service, medical dialogues. *arXiv preprint arXiv:2109.00430*.

Guangtao Zeng, Wenmian Yang, Zeqian Ju, Yue Yang, Sicheng Wang, Ruisi Zhang, Meng Zhou, Jiaqi Zeng, Xiangyu Dong, Ruoyu Zhang, Hongchao Fang, Penghui Zhu, Shu Chen, and Pengtao Xie. 2020. MedDialog: Large-scale medical dialogue datasets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9241–9250, Online. Association for Computational Linguistics.

Yuanzhe Zhang, Zhongtao Jiang, Tao Zhang, Shiwan Liu, Jiarun Cao, Kang Liu, Shengping Liu, and Jun Zhao. 2020. MIE: A medical information extractor towards medical dialogues. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6460–6469, Online. Association for Computational Linguistics.

Yue Zhou, Barbara Di Eugenio, and Lu Cheng. 2025. Unveiling performance challenges of large language models in low-resource healthcare: A demographic fairness perspective. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7266–7278.

A Datasets Considered but Excluded

Several alternative dialogue corpora were evaluated and excluded for the following reasons:

- **Lack of multi-turn diagnostic structure:** MedDialog (Zeng et al., 2020) consists primarily of single-turn, forum-style QA lacking a diagnostic trajectory. MIE (Zhang et al., 2020) is an information-extraction corpus, not a dialogue dataset.
- **Rigid generation:** MDDial (Macherla et al., 2023) relies on template-driven generation, lacking naturalistic variation and sufficient multi-turn depth.
- **Language constraints:** ReMeDi (Yan et al., 2021), MedDG (Liu et al., 2022), and MDD (Wei et al., 2018) are exclusively Chinese-language corpora.
- **Narrow clinical scope:** DX (Xu et al., 2019) and the dataset by Lin et al. (2019) are heavily restricted in scale, covering only five pediatric conditions.

B Source Datasets

B.1 DDXPlus

B.1.1 Evidence Structure

Evidences are typed into three categories.

Binary (`B`) evidences represent yes/no symptom presence.

Categorical (`C`) evidences encode a single value from a fixed set of options (e.g. pain character, onset pattern).

Multi-choice (`M`) evidences are child items parented to a categorical evidence via a `code_question` hierarchy, allowing a patient to affirm multiple sub-features of a single symptom simultaneously (e.g. multiple pain locations).

Evidences are further annotated as `is_antecedent` to distinguish presenting symptoms from medical history items.

B.1.2 Preprocessing and Reinterpretation

The `default_value` of 0 on binary evidences does not encode an explicit patient denial; it reflects the absence of synthesis during dataset generation. Under the closed-world assumption, presence in `EVIDENCES` means affirmed (value = 1); absence means negative (value = 0). Evidence states are therefore reconstructed directly from list membership rather than from default values, preventing misrepresentation of negative findings.

Each patient in DDXPlus carries an `INITIAL_EVIDENCE` field, a randomly selected binary evidence used solely to bootstrap the original rule-based ADD system, with no inherent clinical salience. We retain it as the opening chief complaint, both to preserve the original diagnostic progression structure of the dataset and to ensure consistency across comparisons. The value of DDXPlus for this benchmark lies primarily in its differential diagnosis trajectories and structured evidence instantiation rather than the realism of its opening utterance.

B.2 MIMIC-IV

MIMIC-IV (Johnson et al., 2024) is a large, deidentified electronic health record database sourced from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, MA, covering patients admitted to the emergency department or an intensive care unit between 2008 and 2019. It contains data

for over 65,000 ICU (Intensive Care Unit) stays and over 200,000 ED (Emergency Department) visits, and adopts a modular architecture that facilitates both individual and combined use of disparate data sources. All patient identifiers are replaced with randomized surrogates and dates are shifted by a patient-specific offset, preserving intra-patient temporal consistency while preventing re-identification.

B.2.1 Modules

Unlike traditional uses of MIMIC, involving extraction of dense longitudinal time-series from MIMIC-IV for predictive modeling, our benchmark requires coherent narrative patient presentations. Our dataset makes use of four official MIMIC modules (hosp, ed) and two community extensions (bhc, cdn) (see Table 2).

MIMIC-IV v3.1 (Hosp) (Johnson et al., 2023) is the core hospital module, covering inpatient admissions at BIDMC. We extract five tables: `patients` (demographics and date of death), `admissions` (admission type, insurance, discharge disposition, and in-hospital mortality flag), `diagnoses_icd` (ICD-coded discharge diagnoses with sequence ordering reflecting clinical priority), `d_icd_diagnoses` (the ICD code dictionary, used for crosswalk and label resolution), and `omr` (outpatient measurements including height, weight, and BMI recorded across visits).

MIMIC-IV-ED v2.2 extends MIMIC-IV to the emergency department, covering over 200,000 ED visits. It is the entry point for our entire extraction pipeline. We use `triage` (chief complaint, triage vital signs, pain score, and ESI acuity level), `edstays` (ED stay timeline and crucially the `hadm_id` bridge key that links ED visits to hospital admissions, partitioning our two-track cohort), `diagnosis` (ICD-coded ED-specific diagnoses, assigned independently of the final hospital billing diagnoses and therefore serving as the Track 1 ground truth), `medrecon` (home medications recorded at triage), and `vitalsign` (time-series vital sign measurements taken during the ED stay).

MIMIC-IV-Note v2.2 provides 331,794 deidentified free-text discharge summaries from 145,915 patients, with protected health information replaced by three consecutive underscores following a hybrid rule-based and neural deidentification approach.

BHC extension (Labelled Notes: Hospital Course v1.2.0) is a community-contributed PhysioNet dataset providing 125,123 parsed and labeled hospital course narratives derived from **MIMIC-IV-Note v2.2** discharge summaries, keyed on `hadm_id`. We use the presence and count of parsed hospital course segments as a data richness signal to inform the patient-level deduplication priority scheme for admitted cases, described later.

MIMIC-IV-Ext CDM v1.1 is a curated community extension covering 2,400 admitted cases across four abdominal pathologies: appendicitis (957), cholecystitis (648), pancreatitis (538), and diverticulitis (257). Each case provides pre-extracted, labeled, clinical data including HPI text, physical examination findings, 138,788 laboratory results from 480 unique tests, 4,403 microbiology results, and 5,959 radiology reports across CT, X-ray, ultrasound, and MRI modalities. Crucially, the CDM authors explicitly exclude cases where the pathology name appears in the HPI (indicating pre-diagnosed transfers) and strip diagnostic conclusions from radiology findings sections, making the narratives genuinely suitable for diagnostic simulation. Where a CDM record exists for an admitted case in our pipeline, it takes precedence over the note-derived HPI as the Track 2 narrative input, given its dedicated curation and higher fidelity.

B.2.2 Population Selection and Clinical Framing

The target population is patients who seek care by walking into an emergency department of their own volition, and could plausibly interact with an LLM instead. This framing directly drives the cohort selection criteria. The filtering and selection pipeline is shown in Fig. 7.

The raw MIMIC-IV-ED dataset contains 425,087 ED visits. The first filtering step excludes ambulance arrivals. In the United States, calling an ambulance is expensive and often reserved for acute emergencies; a patient arriving by ambulance is by definition not self-presenting and is unlikely to consult a chatbot as an alternative. This single filter removes a large proportion of the highest-acuity cases by proxy. Remaining high-acuity cases are addressed directly: visits with a triage pain score above 8 out of 10 and visits assigned ESI acuity level 1 (the "immediate/resuscitation" tier of the Emergency Severity Index, a standardized 1–5 severity scale where 1 is most critical) are also excluded. Together these filters operationalize the assumption that the benchmark should reflect cases where a patient is coherent, able to communicate,

Directory	Source	Docs	Key Columns
ed/ *MIMIC-IV-ED v2.2*			
triage.csv.gz		Docs	subject_id, stay_id, chiefcomplaint, temperature, heartrate, resprate, o2sat, sbp, dbp, pain, acuity
edstays.csv.gz		Docs	subject_id, hadm_id, stay_id, intime, outtime, arrival_transport, disposition
diagnosis.csv.gz		Docs	stay_id, seq_num, icd_code, icd_version
medrecon.csv.gz		Docs	stay_id, name, gsn, ndc, etc_rn
vitalsign.csv.gz		Docs	stay_id, charttime, temperature, heartrate, resprate, o2sat, sbp, dbp, pain
hosp/ *MIMIC-IV v3.1*			
patients.csv.gz		Docs	subject_id, gender, anchor_age, dod
admissions.csv.gz		Docs	subject_id, hadm_id, race, admittime, dischtime, admission_type, discharge_location, hospital_expire_flag
diagnoses_icd.csv.gz		Docs	hadm_id, seq_num, icd_code, icd_version
d_icd_diagnoses.csv.gz		Docs	icd_code, icd_version, long_title
omr.csv.gz		Docs	subject_id, chartdate, seq_num, result_name, result_value
note/ *MIMIC-IV-Note v2.2*			
discharge.csv.gz		Docs	subject_id, hadm_id, note_id, text
bhc/ *Labelled Notes: Hospital Course v1.2.0*			
mimic-iv-bhc.csv	PhysioNet	—	hadm_id, text (parsed hospital course)
cdn/ *MIMIC-IV-Ext CDM v1.1* — 2,400 abdominal pathology cases			
history_of_present_illness.csv		—	hadm_id, hpi
physical_examination.csv		—	hadm_id, pe
discharge_diagnosis.csv		—	hadm_id, discharge_diagnosis
laboratory_tests.csv		—	hadm_id, itemid, valuestr, ref_range_lower, ref_range_upper
radiology_reports.csv		—	hadm_id, note_id, modality, region, exam_name, text
microbiology.csv		—	hadm_id, test_itemid, valuestr, spec_itemid
icd_diagnosis.csv		—	hadm_id, icd_diagnosis
icd_procedures.csv		—	hadm_id, icd_code, icd_title, icd_version
discharge_procedures.csv		—	hadm_id, discharge_procedure

Table 2: MIMIC-IV data file organization, sources, and extracted columns.

and could realistically engage in a multi-turn dialogue with a chatbot.

A further quality filter removes cases with no diagnoses codes (ICD-10, or International Classification of Diseases 10th Revision, is the international standard coding system for diagnoses), as ground truth cannot be established for such cases. Finally, pain score entries recorded as unresolvable free-text strings ("unable", "UTA", "critical", and similar) are manually examined and dropped, indicating the patient was unable to communicate with the triage nurse about their pain, and thus falls outside our target population.

To ensure the integrity of the clinical data without reinventing complex data scrubbing rules, we selectively repurposed utility modules from an existing extensive MIMIC-IV processing pipeline. Specifically, we integrated their `outlier_removal.py` functions to enforce physiological bounds on triage vitals (e.g., dropping artifacts/ sensor errors), their `uom_conversion.py` scripts for unit standardization, and their ICD-9 to ICD-10 crosswalk mapping (MIMIC-IV data is recorded across 2 editions of ICD: 9 and 10).

B.2.3 Two-Tracks

After filtering, the remaining cohort naturally partitions into two clinically distinct populations based on whether the ED visit resulted in hospital admission. Of the 425,087 raw visits, 222,071 (52.2%) resulted in discharge without admission and 203,016 (47.8%) in hospital admission. These two groups present fundamentally different diagnostic challenges and map to different real-world scenarios:

Track 1: ED Discharge (telehealth use case): The patient is evaluated and sent home. Feature space is restricted to ED data only (chief complaint, vital signs, ED questionnaire). Ground truth is the preliminary ED diagnosis assigned by the emergency physician at discharge.

Track 2: Hospital Admission (specialist/advanced diagnostic use case): The patient is admitted for further workup. Triage data is augmented with a narrative History of Present Illness (HPI), the structured account of symptoms, onset, duration, and context recorded by the admitting physician, and, where available, structured clinical modalities from the BHC and CDM extensions. Ground truth is the final verified discharge diagnosis, established after the full inpatient workup. Note that for this track, we only retain cases where one or more of the hospital discharge diagnoses (after the patient is admitted and leaves) matches one or more of the preliminary ED diagnoses (the initial diagnos(e)s made by the ED physician which led to hospitalization). Since we use the patient's ED presentation for "chief complaint" and "initial patient state" as simulated "point of contact" with the examiner LLM, it would be unfair to expect the LLM to reach a diagnosis informed by a professional work-up after hospital admission, if that diagnosis was not sufficiently clear to the ED physicians upon presentation. Essentially, giving the chatbot a fair chance at the risk of introducing bias and skewing the data in its favor.

Track 1 tests whether a model can triage from

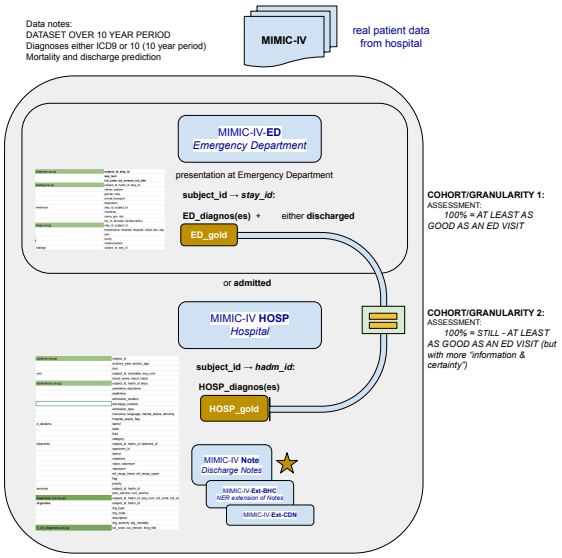

Figure 7: MIMIC-IV: Two Tracks/Cohorts

sparse initial information, while Track 2 tests deeper diagnostic reasoning against a richer clinical picture. It also defines "confidence" in ground truth.

B.2.4 Data Extraction and Linkage

Raw MIMIC-IV tables are preprocessed into a unified 424,995-row candidate dataframe covering 40 (sparse) features per visit. The pipeline initialises from the ED module, merging `triage` and `edstays` via `stay_id` and `subject_id` to form the base cohort. A `LEFT JOIN` (retaining all rows from (ED) table regardless of whether a matching record exists in the right (hospital) table) is then executed to the hospital `admissions` and `patients` tables on `hadm_id` (Hospital Admission ID, the unique key linking an ED visit to an inpatient admission). This join spine preserves non-admitted patients, who receive a null `hadm_id` and route to Track 1, while fully enriching admitted patients (Track 2) with inpatient data. Subsequent steps pack time-series vital signs (`vitalsign`), medication history (`medrecon`), and outpatient measurements (`omr`) into per-visit aggregates; map both ED and hospital diagnoses to ICD-10 with ICD-9 crosswalk applied where needed; and join parsed hospital course narratives from the BHC extension (125,123 records). Triage vital sign outliers and unit inconsistencies are resolved using deterministic cleaning functions from an existing MIMIC-IV pipeline.

For Track 2 cases where a CDM record exists, pre-extracted HPI text, physical examination findings, laboratory results, radiology reports, and microbiology data are merged in via `hadm_id`, taking precedence over note-derived HPI given the CDM's dedicated curation and pre-applied leakage controls. All downstream clinical sections ("Hospital Course", "Assessment and Plan", and similar) are masked throughout the pipeline to prevent ground truth from entering the dialogue generation stage.

B.2.5 Cohort Filtering Steps

The sequential filtering pipeline, applied after extraction, is summarised below. Numbers reflect the state of each cohort at entry to each step.

Pain score normalisation. Free-text pain entries that cannot be resolved to a numeric value are dropped, retained, or edited via a custom semi-manual process. ED-only: 222,006 → 218,902 (−3,104). Admitted: 60,847 → 56,189 (−4,658).

Ambulance arrivals and high pain. Ambulance arrivals and pain scores above 8/10 are excluded. ED-only: 218,902 → 137,896 (−81,006). Admitted: 56,189 → 23,245 (−32,944).

Triage acuity level 1. Immediate/resuscitation cases (ESI level 1) are excluded. ED-only: 137,896 → 135,623 (−2,273). Admitted: 23,245 → 22,059 (−1,186).

Missing ICD-10 codes. Cases missing both primary and all secondary ICD-10 diagnosis codes are dropped, as ground truth cannot be established. ED-only: 135,623 → 74,761 (−60,862). Admitted: 22,059 → 22,059 (−0).

B.2.6 Patient-Level Deduplication

MIMIC-IV contains multiple visits per patient. To prevent "frequent flyers" (patients who return to the hospital recurrently) from dominating the benchmark, and to ensure an i.i.d benchmark, a hierarchical deduplication strategy is applied. For patients with multiple visits, each visit is scored by data richness:

- **Priority 3** (1,460 visits): CDM record present *and* parsed hospital course available - the richest possible entry
- **Priority 2** (185 visits): CDM record present, no hospital course
- **Priority 1** (8,473 visits): hospital course available, no CDM record
- **Priority 0** (8,940 visits): fallback - earliest chronological visit

The highest-priority visit per patient is retained.

B.3 AJCR Case Reports

The original ((Hirosawa et al., 2024)) dataset is a spreadsheet of 392 rows, each containing a PMID, a DOI, up to three gold-standard diagnosis fields (`Final diagnosis1`, `Final diagnosis2`, `final diagnosis34`), and ranked 10-item differential diagnosis lists generated by three LLMs at the time of the original study (Gemini Advanced, Gemini, and Bard), (unused by us). We transform this into a rich structured dataset through a two-stage pipeline. In the first stage, a custom scraping module normalizes each DOI (e.g. `10.12659/AJCR.937787`) into a consistent identifier (`AJCR937787`), constructs the corresponding URL via `doi.org`, and retrieves the full-text HTML of the article, saving it to a per-case directory. Of the 392 cases, 391 were retrieved successfully; 1 failed due to an unavailable page and is excluded. In the second stage, a specialised parser built on `BeautifulSoup` maps the HTML DOM into a standardized nested JSON schema. The parser extracts: granular metadata (title, authors, DOI, journal citation, publication date, and article type) from HTML meta tags and structured page elements; the abstract segmented into *Background*, *Case Report*, and *Conclusions* fields via regex, with a fallback to the page's `meta description` tag where structured paragraph IDs are absent; keywords from the `meta keywords` tag; the full body text split by section (Background, Discussion, Conclusions); figures catalogued by their DOM image anchors with label and caption; and tables detected via their DOM anchor IDs.

Tables in AJCR articles exist in two formats. Where the table is rendered as a native HTML `<table>` element, it is parsed using `pandas` and stored as both a structured record list and a Markdown string representation. Where the table is rendered as an image (a common occurrence in AJCR, where authors submit tables as figures), it is stored with its source URL and caption. In both cases the table is stored uniformly as a string representation within the JSON schema, ensuring downstream access by all LLMs (and not just multi-modal ones for the to tables-as-images).

A key parsing challenge is the disambiguation of case series, in which a single publication describes two or more distinct patients under one DOI. Simple title-string matching for "case series" is insufficiently robust; instead, the parser detects case series by identifying numbered section headers (e.g. "Case 1", "Case 2", "CASE REPORT 1:") within the case report body using a regex pattern, splitting the narrative at these boundaries and assigning paragraphs, figures, and tables to their respective patient. Figures and tables are resolved through named mention in the original HTML markdown. Figures and tables can be referenced by multiple cases and are duplicated accordingly. Assets not referenced by any individual case (e.g. figures appearing only in the Discussion section) are collected in a `general` block. Each individual patient is emitted as a distinct case entry with its own `case_id` (e.g. `AJCR937787_1`, `AJCR937787_2` for a case series, `AJCR937787` for a single-case article), and is treated as an independent case vector downstream.

The final output is a directory of 391 per-article JSON files. Each file contains a `metadata` block, a structured `abstract`, a `cases` list with one entry per patient holding the isolated `main_text` and the patient's associated figures and tables, and a `general` block for unclaimed assets. A processing summary CSV records parsed status and per-article case count across all 392 entries.

B.3.1 Output Structure

The final output is a dataset of JSON files where each entry contains a global `metadata` object, a structured `abstract`, and a list of `cases`. Each case object contains the isolated patient narrative (`main_text`) and references to the specific data assets (tables/figures) relevant to that patient.

B.4 Anchor Resolution: `CUI` and ICD-10 Mapping

Mapping raw clinical strings to canonical UMLS `CUIs` and ICD-10 codes proved non-trivial; we briefly document the failure of some approaches that motivated the eventual 2-stage pipeline design.

LLMs and APIs as ontology resolvers: counterproductive Initial attempts to use MedGemma (Sellergren et al., 2025) as a semantic resolver failed entirely; the model confidently hallucinated non-existent `CUIs` and fabricated terms. This is unsurprising, as LLMs encode token distributions rather than versioned database mappings. This aligns with prior findings that LLMs struggle with ICD-10 labeling (26% accuracy) (Simmons et al., 2025) and exhibit high hallucination rates (35%) even in specialized RAG setups (Klang et al., 2024), making current LLMs highly unrealistic for reliable ontology resolution.

The UMLS REST API proved similarly unsuitable for large-scale annotation. Exact-string

matching yielded poor recall, while fuzzy searches caused unacceptable precision losses by highly ranking superficial string overlaps. Since tuning search parameters per query is unscalable, we abandoned both approaches. Instead, we implemented a tried-and-true deterministic NER pipeline using `medspaCy` and `QuickUMLS` locally against a fixed UMLS release.

Working solution: medspaCy → QuickUMLS. In the first stage, clinical named entity recognition is performed using `medspaCy` (Eyre et al., 2021), a spaCy-based framework for clinical text processing, which segments input text and identifies candidate medical concept spans. In the second stage, each candidate span is passed to `QuickUMLS` (Soldaini and Goharian, 2016), a fast approximate string matching engine built over a local UMLS installation, which returns the best-matching `CUI` along with a similarity score retained as the `mapping_confidence`. All additional non-anchor patient information, across all three sources, is mapped back to either `CUI` identifiers or, where not possible, tightly structured enumeration options or numeric scale representations.

B.5 PatientVector: Supplementary Details

The structural overview of the `PatientVector` schema is provided in Section 4.2.1. Below are specific technical implementations and schema attributes excluded from the main text for brevity:

- **GroundTruth (`provenance_tier`):** In addition to the MIMIC-IV tiers, reliability labels include `ddxplus_gold` (synthetic ground truth) and `ajcr_gold` (peer-reviewed case report ground truth).
- **PatientFacts Extensions:** The `Demographic` block also captures education level. `FamilyHistory` is strictly structured as a dictionary mapping specific family members to their associated `CUI`-mapped diagnoses or symptoms.
- **Explicit Symptom Negation:** Within `CasePresentation`, symptoms carry an explicit negation flag. This structurally distinguishes symptoms the patient actively denies from symptoms that are simply not mentioned (`not_known`). This distinction is critical for evaluating whether an examiner model makes unsafe closed-world assumptions during clinical reasoning.

C Utterance Generation Details

To generate natural language patient utterances from the structured symptom representations in each case presentation variant, we make use of UserLM-8B (Naous et al., 2025), a model released by Microsoft Research that is trained to produce human-like user utterances rather than assistant-style responses. Unlike standard instruction-tuned models, UserLM-8B is optimised to express information the way a person would (colloquially, incompletely, and with the natural hedging and imprecision of lay speech) making it well-suited for converting structured clinical evidence entries into plausible patient utterances.

For each symptom or evidence entry in the visible presentation layer of a variant $CP_{xi}V_j$, UserLM-8B is used to generate a set of utterance candidates spanning a range of expression styles: from relatively precise phrasing ("I have been experiencing blurred vision") to lay and colloquial formulations ("I can't even read or write, my vision is blurry"), and from direct self-report to third-party reporting ("my wife says I seem confused"). The persona parameters of the variant condition the generation to ensure that utterances are consistent with the patient profile. The result is a pool of grounded, `CUI`-anchored utterance candidates per symptom per patient variant per `PatientVector`, which together constitute a structured collection of pre-generated patient expressions, reproducible and decoupled from live "utterance generation" inference.

D Patient Orchestrator

At dialogue runtime, the Patient agent acts as an **orchestrator**: given the Examiner's most recent question or clarification, it selects which pre-generated utterance or combination of utterances best responds to that question, subject to the reveal conditions defined in the variant config (illustrated in Fig. 8). This is not free generation, the Patient draws from the pre-generated utterance pool for the current variant, ensuring that responses remain grounded in the case facts, and are not subject to the "orchestrator" LLM generation. MedGemma (Sellergren et al., 2025) is used as the orchestration model, responsible for interpreting the Examiner's question in context and selecting the appropriate patient response from the candidate set. This approach deliberately separates the *content* of what the patient says (grounded in real clinical data, pre-generated) from the *selection logic* (contextually driven by the dialogue history), avoiding both the

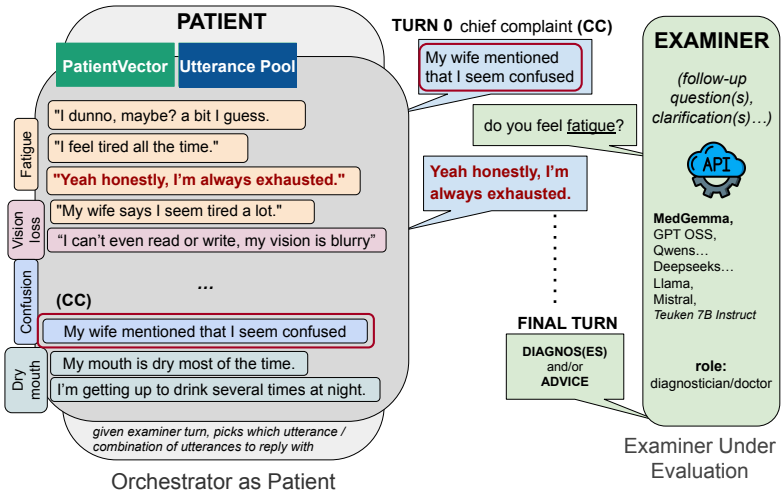

Figure 8: **Dialogue runtime**: The Patient Orchestrator receives the Examiner's turn and selects which pre-generated utterances to surface, drawing exclusively from the utterance pool for the current case. Content (what the patient says) is fixed at dataset creation time; selection logic (which utterance to return) is handled at runtime. The Examiner LM receives only the chief complaint at Turn 0 and issues follow-up questions until a terminal diagnosis or turn limit is reached.

rigidity of purely rule-based response selection and the uncontrolled hallucination risk of fully free-form generation by User-LM.

The Examiner agent receives only the chief complaint utterance at Turn 0 and issues follow-up questions and clarification requests across subsequent turns. The dialogue continues until the Examiner produces a terminal diagnosis or a maximum turn limit is reached. Every turn role, raw text, Examiner question, Patient utterance selected, and `CUIs` extracted is logged to a structured JSON log that serves as the input to the evaluation pipeline.

E Clinical Reasoning Metrics

The evaluation framework MTDiag enables goes beyond diagnostic accuracy. Because every symptom and diagnosis in the dataset is anchored to a `CUI` or ICD-10 code, evaluation can operate over ontological identifiers rather than surface strings, allowing graded, clinically interpretable scoring at both the dialogue and utterance level. We describe the metrics below, organized from the most direct (diagnostic outcome) to the most structurally demanding (reasoning fidelity and harm).

E.1 Diagnostic Outcome Metrics

Diagnosis Accuracy (Diag_{acc})**.** The trivial binary baseline: did the examiner model produce a diagnosis matching the ground-truth ICD-10 code?

$$\text{Diag}_{\text{acc}} = \begin{cases} 1 & \text{if predicted code matches} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Semantic Diagnostic Distance. Because ICD-10 is hierarchical, a binary match/miss discards clinically meaningful information. A model outputting `I21.0` (acute myocardial infarction) when the ground truth is `I21.02` (STEMI of the LAD artery) has committed an under specification, not categorical error (difference in 3-character code). Secondary diagnoses can be used as a near-miss buffer: a predicted code matching any secondary diagnosis is flagged as clinically consistent rather than wrong.

E.2 Symptom Elicitation Metrics

The following metrics assess not whether the model reached the right diagnosis, but whether it asked the right questions to get there. They require the `CUI` extraction step applied to examiner outputs at each turn (Section 4.4).

Disease-wise Symptom Score (S_{dise})**.** Adapted from Macherla et al. (2023), this metric evaluates how many of the symptoms the examiner queried are relevant to the diagnosed disease, adjusted by a turn-efficiency penalty. Let S_d be the set of symptoms associated with disease d (drawn from the HPO disease-phenotype prevalence table or curated lookup tables), and let S_{diag} be the `CUI`-resolved set of symptoms the examiner asked about.

$$f(s, S_d) = \mathbf{1}[s \in S_d] \quad (2)$$

$$C = \frac{\min(N_{\text{gold}},\, N_{\text{pred}})}{\max(N_{\text{gold}},\, N_{\text{pred}})} \quad (3)$$

$$S_{\text{dise}} = \frac{C}{N_{\text{pred}}} \sum_{s_i \in S_{\text{diag}}} f(s_i,\, S_d) \quad (4)$$

where N_{gold} is the number of turns in a reference "ideal" dialogue and N_{pred} is the examiner's turn count. The "ideal" dialogue is only well-defined for DDXPlus instances in MTDiag. For the other datasets, we can consider N_{gold} as the number of `CUI`-mapped symptoms in S_d for the ground-truth diagnosis that is, the size of the relevant symptom set a thorough examiner would ideally cover and N_{pred} as the number of turns the examiner took. High S_{dise} indicates that the model's symptom inquiries were clinically coherent with the diagnosis it ultimately reached.

MTDiag extension: weighted relevance The binary $f(s, S_d)$ treats all associated symptoms as equally important, which is clinically unrealistic. We replace it with a frequency-weighted version drawn from the HPO `phenotype.hpoa` annotations:

$$f_w(s, S_d) = \omega(s, d) \cdot \mathbf{1}[s \in S_d] \tag{5}$$

where $\omega(s, d) \in \{1.0, 0.75, 0.5, 0.25, 0.1\}$ (suggestion) corresponds to HPO frequency qualifiers *obligate*, *very frequent*, *frequent*, *occasional*, and *very rare* respectively. This rewards the prioritization of high-yield symptoms early in the dialogue.

Symptom Precision and Recall

$$\text{Prec} = \frac{|S_{\text{diag}} \cap S_d|}{|S_{\text{diag}}|} \tag{6}$$

$$\text{Rec} = \frac{|S_{\text{diag}} \cap S_d|}{|S_d|} \tag{7}$$

Precision penalizes irrelevant or off-topic questioning; recall captures whether the model elicited a sufficient fraction of the relevant symptom space.

E.3 Reliability and Reasoning Fidelity

Reliability Score (R_{score}). Also adapted from Macherla et al. (2023), this metric unifies diagnostic and symptom performance: a model is credited only if it reached the correct diagnosis *and* its symptom inquiries were sufficiently relevant.

$$R_{\text{score}} = \begin{cases} 1 & \text{if } S_{\text{dise}} \geq t \text{ and } \text{Diag}_{\text{acc}} = 1 \\ 0 & \text{otherwise} \end{cases} \tag{8}$$

The threshold $t \in (0, 1)$ controls strictness. A model that guesses the correct diagnosis while asking clinically incoherent questions receives $R_{\text{score}} = 0$, distinguishing systematic reasoning from lucky outcomes.

Pathognomonic and Sine Qua Non Errors. Some symptoms carry special logical status for a given diagnosis. *Pathognomonic* symptoms are sufficient to confirm a diagnosis when present; *sine qua non* symptoms are necessary: their absence invalidates the diagnosis. A model that reaches a correct diagnosis without ever eliciting a sine qua non symptom has committed a logical failure: the conclusion is unsupported by the evidence gathered. Such cases can be flagged **diagnostic hallucinations**: correct outputs reached via clinically incoherent pathways.

Pathognomonic errors measure whether the model failed to inquire about or recognize a uniquely identifying symptom for its own stated diagnosis. Sine qua non errors capture omission of necessary symptoms. Both can be integrated into R_{score} as hard veto conditions: a diagnosis unsupported by its sine qua non evidence is treated as invalid regardless of string match.

Anchoring Bias A model may pursue a single working hypothesis throughout the dialogue, ignoring evidence that would refocus the differential. Using disease-phenotype associations, this can be detected as the degree to which the examiner's queried symptoms S_{diag} are contained within S_d for a single candidate diagnosis d, with no exploration of symptoms associated with plausible alternatives. A model that asks only cardiovascular-related questions while the case is a pulmonary embolism is anchoring. Anchoring could lead to a correct diagnosis only when the diagnosis matches the model's opening hypothesis and constitutes an error in all other cases.

Premature Closure. If a model commits to a diagnosis before eliciting a sufficient proportion of the presenting symptoms by symptom-frequency weight, the diagnosis is flagged as premature. No established clinical threshold exists for this criterion; it is left as a configurable parameter chosen to represent a minimal evidentiary bar before diagnostic commitment.Sensitivity analysis across different cutoff values is left to future work.

Turn Efficiency. The cost penalty C in S_{dise} captures efficiency implicitly, but turn count can also be reported directly as a standalone metric. Pathologically short dialogues (fewer turns than there are sine qua non symptoms) and pathologically long ones (exceeding the maximum turn limit) are flagged.