Technische Universität München

UT Southwestern Medical Center

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

Pia Chouayfati, Alexander M. Fichtl, Miriam Anschütz, George Doumat, Georg Groh

Ⅰ Motivation & Background

Commercial LLMs are already ubiquitously deployed to the general public and often marketed as "*health assistants*." However, recent qualitative assessments by the medical community have determined that they often give problematic answers to patient-posed questions[1]. LLMs score up to 90% on static QA-style medical knowledge tasks, but performance drops to ~45% in interactive diagnostic tasks[2]. Existing benchmarks present *complete*, *static*, *structured* information, evaluating recall rather than real-world diagnostic competence[3].

Clinical diagnosis is inherently **incremental** and **interactive**; patients frequently use imprecise language, forget symptoms, or omit context. In practice, doctors and diagnosticians are trained in **differential diagnosis[4]:** 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.

The Goal: MTDiag is a dataset to evaluate LLMs on their ability to conduct a dynamic diagnostic conversation and gather evidence under uncertainty. The questions MTDiag answers are 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?*

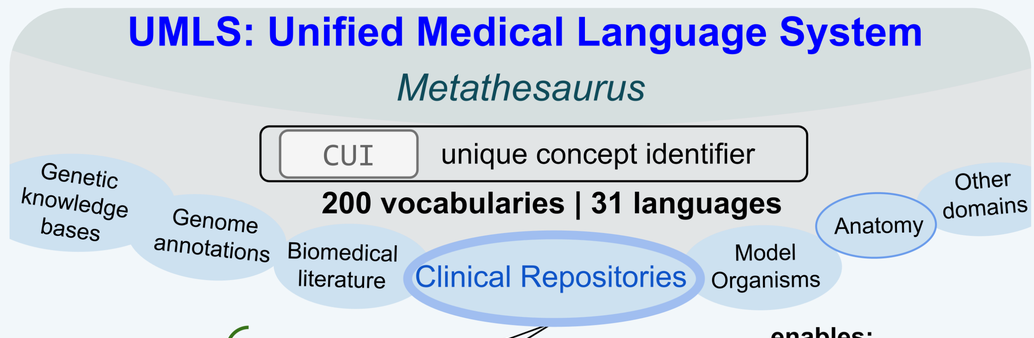
Ⅱ Experimental Setup

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), where the examiner's task is inherently one of **differential diagnosis**.
We define three necessary elements for the conduction of a diagnostic dialogue (**Minimum Anchors**):

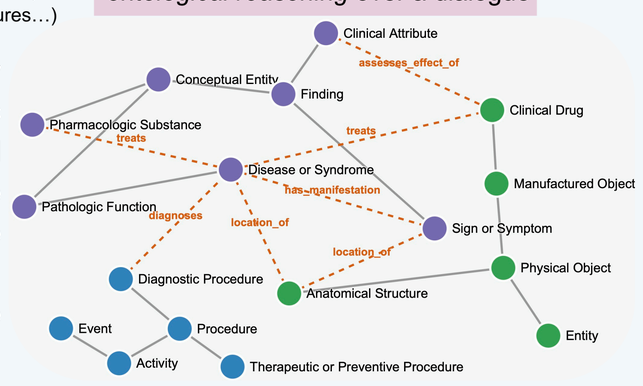
- **Chief Complaint (CC):** the patient's primary reason to seek care, according to "them." This serves as the opening utterance of the dialogue.
- **Symptoms**: the patient's other symptoms.
- **Primary diagnosis*:** the ground truth principal finding of the case.

**match constitutes the minimum evaluation target: did the model reach the right diagnosis?*

The **minimum anchors** (and any available additional patient information) are mapped to the **UMLS[5]**, a metathesaurus of international biological and medical knowledge bases, ontologies and vocabularies, built for healthcare system interoperability.

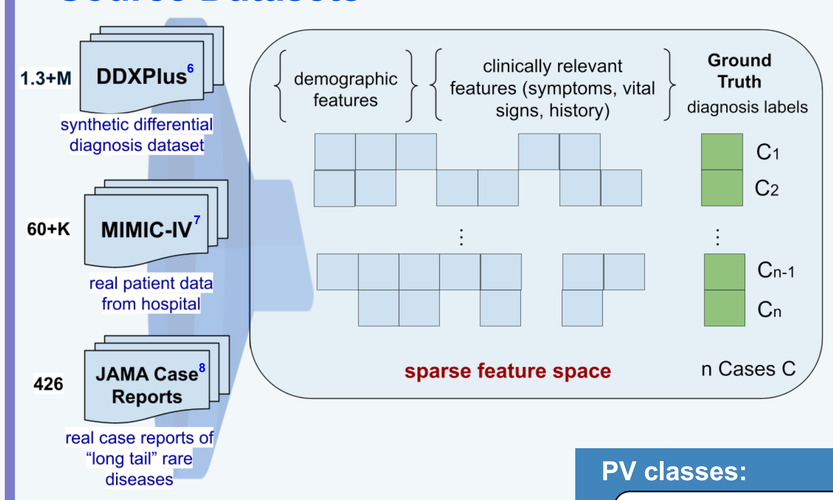


The **UMLS** indexes unique concepts by **CUI**, and instantiates a **semantic network (right).** This allows ontological reasoning over a diagnostic dialogue and the computation of clinical dialogue metrics (beyond diagnostic accuracy), from the **PatientVector** at dialogue runtime.



Ⅲ Dataset Construction

Source Datasets



The three source datasets are heterogeneous and were selected due to their suitability for our experimental set-up.
They were individually pre-processed (the MIMIC-IV subset specifically was constructed from several community extensions from PhysioNet[9]) and normalized into the below **PatientVector** schema, which constitutes the ***deterministic*** part of the dataset.

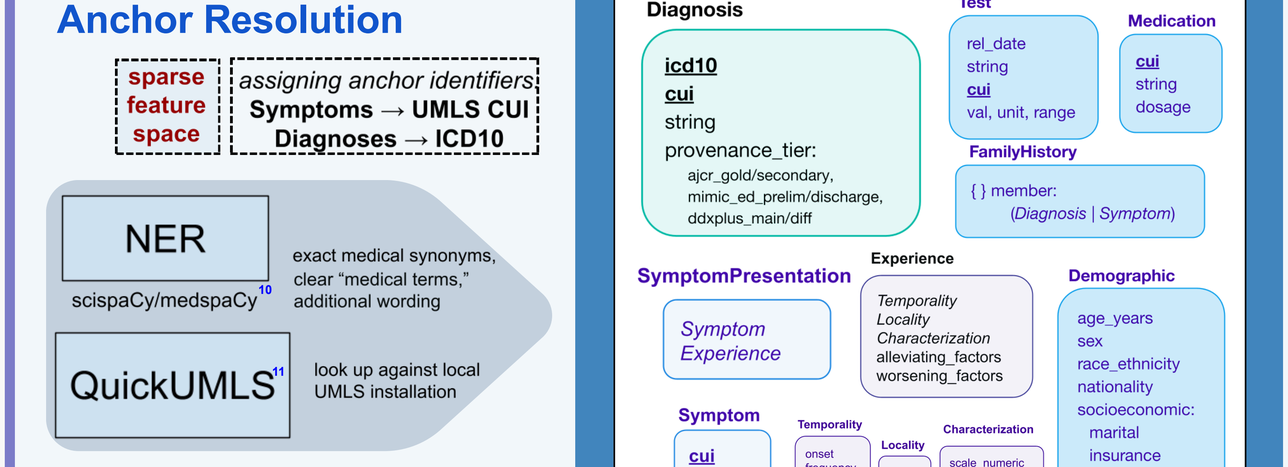
Anchor Resolution

sparse feature space — *assigning anchor identifiers* — **Symptoms → UMLS CUI, Diagnoses → ICD10**

- **NER** (scispaCy/medspaCy[10]): exact medical synonyms, clear "medical terms," additional wording
- **QuickUMLS[11]**: look up against local UMLS installation

Utterance Generation

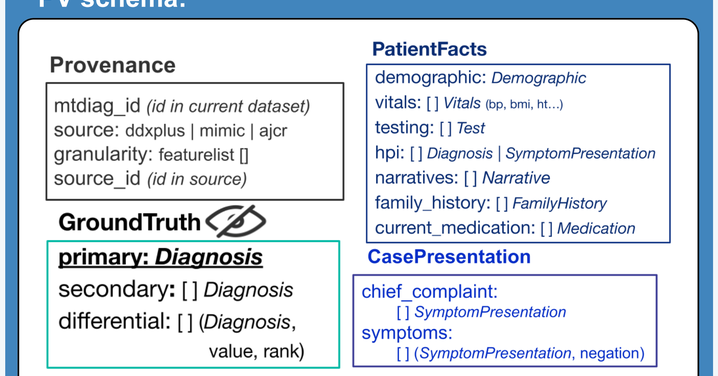
An important design choice is a **separation of concerns** (SoC) between "*patient simulation*" and "*dialogue conduction*." We pre-generate the utterances for each patient feature using UserLM-8B[12], which, unlike "assistant LLMs," is trained on user chats and meant to realistically simulate a user talking to a chatbot. Candidate utterances are generated for each patient feature, constituting the **Utterance Pool**, and at **Dialogue Runtime**, a separate Patient Orchestrator simply serves the appropriate utterance from the pool, given the current conversation.



PatientVector (PV)

PV classes:

PV schema:



In addition to other practical/technical benefits, this decoupling reduces symptom hallucination during live chat and ensures all "patient responses" remain reproducible and grounded in real clinical data, allowing review of the "patient simulation quality" independent of the dialogue. The **Utterance Pool** constitutes the ***generative*** part of the dataset.

Ⅳ Dialogue Runtime

PV

GroundTruth:
C0039483 (Giant Cell Arteritis)
Provenance: ajcr (AJCR933471)
PatientFacts:
demographic: [age: 69, sex: F]
CasePresentation:
chief_complaint: C0018681 (headache)
symptoms:
C0151206 (scalp tenderness)
C0238155 (jaw claudication)
C4303047 (visual field defect)
...

Utterance Pool

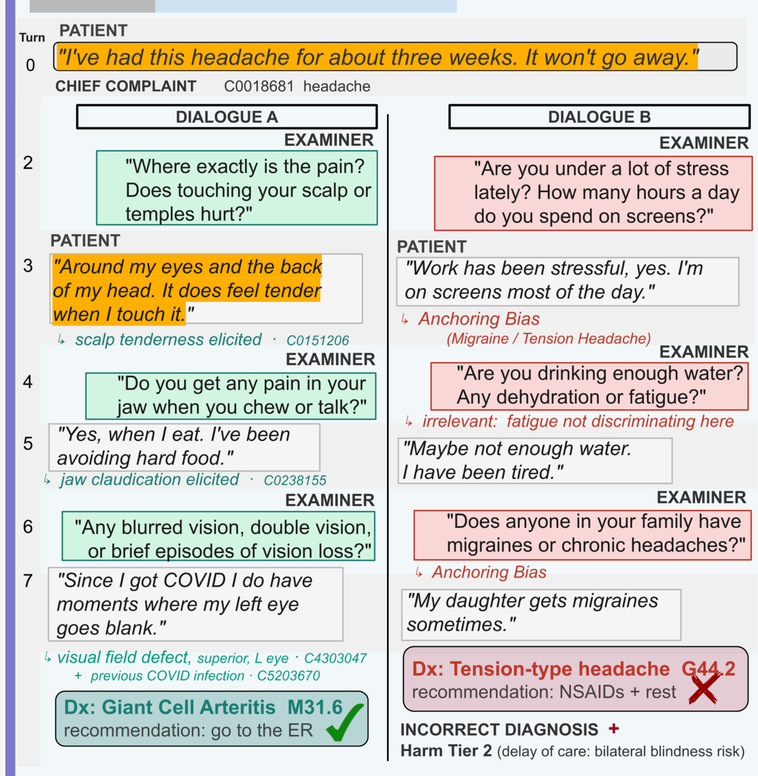
C0018681 headache (**CC**) CUI ...	"I've had this headache for about three weeks. It won't go away." ,...,"My head has been pounding around my eyes and the back of my head."
C0151206 scalp CUI tenderness	"My scalp is very tender. Even combing my hair hurts.", … , "It does feel tender when I touch it."

with a **PatientVector (PV) + Utterance Pool**:

a *diagnostic dialogue* is conducted between a **PatientOrchestrator*** and an **Examiner**
**MedGemma[13] in our setup* — *LLM under evaluation*

Patient Orchestrator at dialogue runtime, selects response from utterance pool *with the patient's chief complaint (**CC**) as **Turn 0***

We trace 2 representative dialogues (A & B) from the same PV[14], that illustrate the important differential diagnosis metrics which MTDiag's ontological anchoring renders computable. The patient's ground truth diagnosis is Giant Cell Arteritis (GCA), a rare disease inflammatory disease which can lead to irreversible blindness if left untreated.



Clinically-Motivated Metrics

Symptom Elicitation Score S_{dise} did the examiner ask about clinically relevant, discriminating symptoms for the diagnosis reached? **A** scores high: every question (scalp tenderness, jaw claudication, vision loss) maps onto GCA's symptom cluster.

Reliability Score: credits a diagnosis only if S_{dise} clears a threshold too, catching lucky "hallucination" guesses.

Anchoring Bias: when questions map to only one candidate diagnosis, never exploring alternatives. **B** asks only migraine/tension-headache questions (stress, screens, migraine): GCA is never even considered.

Premature Closure: committing to a diagnosis before eliciting enough evidence. **B**

Harm Index (Tiers 1–3):
direct harm | delay of care | instructional failure

$$f_w(s, S_d) = \omega(s,d) \cdot \mathbb{1}[s \in S_d] \qquad C = \frac{\min(N_{gold}, N_{pred})}{\max(N_{gold}, N_{pred})} \qquad S_{dise} = \frac{C}{N_{pred}} \sum_{s_i \in S_{diag}} f_w(s_i, S_d)$$

$\mathbb{1}$ if s is a known symptom of disease d

$\omega \in \{1.0, \ldots, 0.1\}$: HPO frequency weight (obligate → very rare)

turn-efficiency penalty (ideal vs. actual # turns)

S_{diag}: symptoms examiner asked about, relevance-weighted

additional metrics and a more complete discussion of computability can be found in the full text

I: [1](Draelos et al., 2026), [2](Gong et al., 2025), [3](Alaa et al., 2025), [4]([6]Fansi Tchango et al., 2022; Tu et al., 2025); **II:** [5](Bodenreider, 2004); **III:** [7](Johnson et al., 2023), [8](Hirosawa et al., 2024), [9](Goldberger et al., 2000), [10](Honnibal et al., 2020), [11](Soldaini et al., 2016), [12](Naous et al., 2025); **IV:** [13](Sellergren et al., 2025), [14](Szydełko-Paśko et al., 2022)

The calculations for this research were conducted with computing resources under the project **Towards a Realistic Multi-turn Medical Dialogue Benchmark for LLMs**, granted by Federal Ministry of Education and Research (**BMBF**) project AI service center KISSKI (**Grant N. 01IS22093A-E**).