

PROVENA-MED: Benchmarking Provenance Faithfulness in Multimodal Clinical Reasoning

J. Ben Tamo[♣], Yining Yuan[♣], Benoit L. Marteau[♣], Micky C. Nnamdi[♣], Yuxing Lu^{♣♥},
and May D. Wang[♣]

[♣]Georgia Institute of Technology [♥]Peking University

Corresponding authors: maywang@gatech.edu

Abstract

Clinical language and vision-language models increasingly generate diagnostic and treatment rationales, but those claims remain difficult to audit. Existing benchmarks usually score final answers over fully revealed records, without testing whether each claim is grounded in valid patient evidence or whether models can seek missing information efficiently. We introduce PROVENA-MED, a benchmark for evaluating provenance-faithful multimodal clinical reasoning over linked EHR tables, clinical notes, chest radiographs, and guideline rules. Models must produce a staged artifact comprising a problem list, differential, workup, therapy, and monitoring, in which each claim cites typed evidence units and may abstain when information is insufficient. PROVENA-MED supports both full-information evaluation and an interactive evidence-seeking track. We further propose MM-AIS, which measures citation validity, objective support under evidence-removal tests, salient-evidence recall, and, for open-weight models, sensitivity to cited evidence. Across ten open-weight models and five cohorts, the best attribution yield is only 0.42, citation hallucination is common, and diagnostic accuracy, safety, and provenance do not reliably co-vary. PROVENA-MED provides an auditable testbed for evaluating grounded clinical reasoning beyond final-answer accuracy.

1 Introduction

Clinical decision-making is inherently multimodal, iterative, and grounded in evidence: clinicians develop problem lists, update differential diagnoses, request targeted diagnostic tests, and ultimately select treatments and follow-up strategies based on structured data, narrative documentation, imaging, and clinical guidelines (Elstein et al., 1978; Croskerry, 2005). Large language models and vision-language models are now being tested on analogous clinical tasks that span structured records, free-text notes, and medical images

(Acosta et al., 2022; Singhal et al., 2023; Tu et al., 2023). Yet the rationales they generate remain difficult to audit. For any diagnostic or therapeutic claim, clinicians must be able to determine whether the referenced patient evidence exists, whether it supports the claim, whether key evidence was omitted, and whether the model should have abstained because essential information was unavailable.

Existing benchmarks only partially address this need. Medical QA datasets test clinical knowledge but are disconnected from patient records (Jin et al., 2021; Pal et al., 2022). EHR-grounded and multimodal benchmarks incorporate tables, notes, or images, but typically evaluate answer retrieval, semantic parsing, visual QA, or outcome prediction rather than claim-level provenance (Bardhan et al., 2022; Lee et al., 2022; Bae et al., 2023; Chen et al., 2023; Kweon et al., 2024). Interactive diagnostic systems evaluate information gathering, but generally do not require each generated claim to be linked to typed patient evidence and guideline rules (Tu et al., 2023; McDuff et al., 2025; Saab et al., 2026). Attribution frameworks such as AIS, FActScore, and ALCE provide tools for checking whether text is supported by cited sources, but they are not designed for temporally filtered, multimodal clinical records or budgeted evidence seeking (Rashkin et al., 2023; Min et al., 2023; Gao et al., 2023).

We introduce **PROVENA-MED**¹, a benchmark for evaluating provenance faithfulness in multimodal clinical reasoning. Each case is a temporally filtered patient bundle containing structured EHR events, clinical note spans, chest-radiograph findings, and guideline-rule metadata. Models produce a staged reasoning artifact comprising a problem list, differential diagnosis, workup, therapy, and monitoring, in which

¹Code: <https://github.com/Wizaaard/PROVENA-MED>.
Dataset: <https://huggingface.co/datasets/provena-med/provena-med>

each claim must cite typed evidence identifiers (TABLE_ROW, NOTE_SPAN, IMAGE_FINDING, or GUIDELINE_RULE) and may abstain when evidence is insufficient. PROVENA-MED supports both full-information evaluation and an interactive evidence-seeking setting, in which models begin with limited information and request additional evidence under a discrete budget.

We construct PROVENA-MED from linked clinical resources, including MIMIC-IV, MIMIC-IV-Note, MIMIC-CXR, MIMIC-III, and eICU, with evaluation targets triangulated from existing expert and objective signals such as radiology reports, diagnosis codes, structured abnormalities, clinician actions, and source-traceable guideline rules (Johnson et al., 2016, 2019, 2023b; Pollard et al., 2018). Its Multimodal Attributable to Identified Sources (MM-AIS) metric separates citation validity, evidence support, salient-evidence recall, and open-weight sensitivity to cited evidence. Evaluating ten open-weight models across five cohorts shows that provenance remains weak: the best attribution yield is 0.42, citation hallucination is common, and diagnostic accuracy, safety, and provenance dissociate. Our contributions are:

- (i) We formalize staged multimodal clinical reasoning as an auditable artifact with typed patient citations, guideline traces, safety status, missing-information fields, and abstention.
- (ii) We construct PROVENA-MED, a temporally filtered multimodal benchmark with both full-information and interactive evidence-seeking settings.
- (iii) We propose MM-AIS, which evaluates citation validity, objective support, salient-evidence recall, and open-weight model sensitivity to cited evidence.
- (iv) We show that accuracy, safety, and provenance are distinct capabilities, with current models frequently producing unsupported, invalid, or unused citations.

2 Related Work

Verifiable clinical reasoning draws on three lines of work that, to our knowledge, have not yet been brought together in a single benchmark: medical foundation models and their evaluation, faithfulness and counterfactual attribution in NLP, and multimodal grounding and interactive diagnosis.

Medical Foundation Models and Benchmarks. Recent medical foundation models demonstrate

substantial clinical knowledge, with systems such as Med-PaLM and Med-PaLM 2 performing strongly on medical QA (Singhal et al., 2023). Concurrently, vision-language models such as Med-PaLM M, Med-Flamingo, LLaVA-Med, and RadFM have extended this capability to radiology and biomedical visuals at scale (Tu et al., 2023). However, evaluating these systems requires moving beyond decontextualized, exam-style QA (Jin et al., 2021; Pal et al., 2022). More realistic EHR-grounded benchmarks, including DrugEHRQA, EHRSQL, EHRNoteQA, EHRXQA, and MC-BEC, incorporate structured tables, notes, and imaging (Bardhan et al., 2022; Lee et al., 2022; Kweon et al., 2024; Bae et al., 2023; Chen et al., 2023), but they still primarily evaluate answer retrieval, semantic parsing, or outcome prediction. In both cases, a correct answer is the unit of evaluation, and the chain from patient evidence to claim is not scored. A model that arrives at the right diagnosis by ignoring the imaging finding that should have driven it is indistinguishable from one that reasoned correctly.

Faithfulness and Counterfactual Attribution.

Outside medicine, evaluating whether generated text is supported by evidence often relies on attribution frameworks such as Attributable to Identified Sources (AIS) (Rashkin et al., 2023), FActScore (Min et al., 2023), and Automatic LLM Citation Evaluation (ALCE) (Gao et al., 2023), with recent medical adaptations testing post-hoc source verification (Wu et al., 2025). Surface-level citation, however, does not guarantee genuine model dependence: a model can cite an evidence span it never used. A stricter line of work studies *causal faithfulness* by intervening on the input or the model to test whether the cited evidence actually mediates the output. Methods include counterfactual text editing (Abraham et al., 2022; Gat et al., 2024), input perturbation (Matton et al., 2025), and mechanistic activation patching (Heimersheim and Nanda, 2024; Geiger et al., 2024; Zaman and Srivastava, 2025). While this literature establishes rigorous intervention-based provenance testing, it remains largely text-only, explanation-centric, or model-internal. Extending it to cross-modal clinical evidence requires an intervention operator that is well-defined for table rows, note spans, and imaging findings simultaneously, as well as a benchmark whose evidence units are stably keyed for that operator to act on.

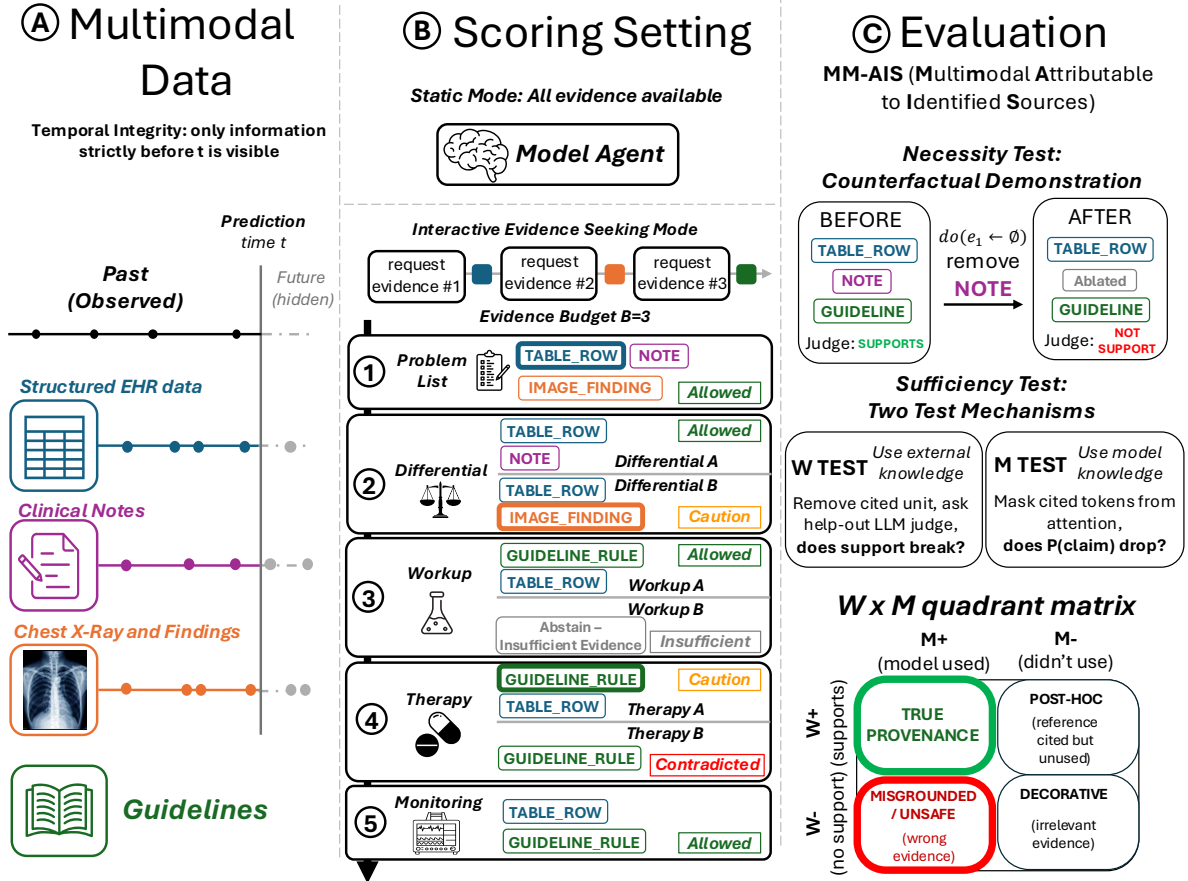


Figure 1: Overview of PROVENA-MED. Given a temporally filtered multimodal patient bundle, an evaluated model may operate either in a full-information setting or as a budget-limited interactive evidence-seeker. The model must produce a staged reasoning artifact in which each claim is linked to typed patient evidence and rule traces, and is then evaluated with counterfactual provenance tests, diagnosis and safety metrics, abstention metrics, and interactive-efficiency metrics.

Multimodal Grounding and Interactive Diagnosis. Real-world clinical reasoning is fundamentally multimodal and sequential. Radiology resources such as CheXpert, REFLACX, MS-CXR, and BenchX provide strong supervision for image-level findings and text alignment (Irvin et al., 2019; Lanfredi et al., 2022; Boecking et al., 2022; Zhou et al., 2024). Concurrently, work on interactive diagnosis treats medicine as an information-gathering dialogue, exemplified by AMIE and its multimodal extensions (Tu et al., 2025; McDuff et al., 2025; Saab et al., 2026), alongside studies on guideline adherence (Fast et al., 2024) and uncertainty recognition (Griot et al., 2025). While these works highlight the importance of multimodal evidence, staged evidence-seeking, and abstention, existing benchmarks do not unify these capabilities. They lack a framework that scores a model’s claim against *all* of them at once: cited to typed cross-modal evidence, causally verified rather than

merely surface-attributed, produced under a budgeted evidence-seeking process, and allowed to abstain.

3 Methodology and Benchmark Construction

To close the gap between multimodal clinical capability and verifiable reasoning, we introduce PROVENA-MED. The benchmark evaluates whether clinical models are *provenance-faithful*, safe, and appropriately uncertain. Unlike prior static benchmarks, it also includes an interactive setting that measures whether models gather evidence *efficiently*.

3.1 Clinical Task Formulation

A patient’s electronic health record is not a static document but a *longitudinal timeline*: vitals result over hours, notes are written, and images are acquired. Clinicians must act at specific moments us-

ing only what has been recorded so far. PROVENA-MED mirrors this setting by slicing each encounter at a fixed decision point and asking the model to reason from that snapshot and, in the interactive setting, to decide what additional evidence to gather.

Formally, for each encounter e , we define a *prediction time* t representing this clinical decision point. To enforce temporal integrity, the model is restricted to a *windowed snapshot*; nothing charted after t is visible. At t , the available evidence is:

$$X = \{X_s, X_n, X_i, X_g\} \quad (1)$$

where X_s are timestamped structured events up to t (labs, vitals, medications), X_n are clinical notes segmented into sentence-level spans, X_i are chest radiographs with their findings, and X_g is a pool of applicable guideline rules.

The model produces a staged artifact spanning five phases: problem list, differential, workup, therapy, and monitoring. Within any stage, each output is represented as a tuple:

$$(c, E, R, s, m) \quad (2)$$

where c is the specific statement, decision, or action the model is generating at that exact stage of its reasoning, $E \subseteq \{\text{TABLE_ROW}, \text{NOTE_SPAN}, \text{IMAGE_FINDING}\}$ are stably-keyed patient evidence citations, $R \subseteq \{\text{GUIDELINE_RULE}\}$ are explicit rule traces, $s \in \{\text{allowed}, \text{caution}, \text{contraindicated}, \text{insufficient}\}$ is a safety status, and m is a missing-information list. A stage may also globally abstain if critical data is lacking.

The same schema is scored in two settings:

- **Full-information** (static): The model receives the entire windowed record X at once.
- **Interactive** (evidence-seeking): The model starts from a minimal initial state s_0 (e.g., chief complaint and initial vitals) and issues $\text{REQUEST}(\tau)$ actions for specific evidence classes τ . The interactive state accumulates the subsets of X revealed so far. Because each request returns the historically recorded result (or “not resulted”), no patient simulator is required. The model operates under a discrete request budget and may COMMIT the staged artifact or ABSTAIN/REFER at any time.

3.2 Dataset Cohorts and Temporal Linkage

To implement this windowed snapshot, all extracted evidence units are strictly timestamped. To prevent label leakage, medications are split: home/pre-admission lists are visible, while inpatient orders are strictly withheld because they implicitly encode the clinician’s treatment decisions.

We assemble a balanced evaluation benchmark of 12,000 encounters across five cohort families (2,400 per cohort; split into 200 dev / 2,200 test cases). These cohorts provide complementary reasoning challenges:

- **Emergency Department:** Provides an information-poor cold start with clinician-approved referral labels, serving as the primary home for the interactive track.²
- **Multimodal ICU & Cardiac:** Links MIMIC-IV (Johnson et al., 2023b,a) structured data and notes to real chest-radiograph pixels from MIMIC-CXR (Johnson et al., 2019). Because cases carry both the radiologist report and the DICOM, this enables direct comparison of text-only and vision–language models.
- **External Validation (eICU & MIMIC-III):** Tests external generalization across a different health system (Pollard et al., 2018) and temporal generalization to an older diagnostic coding era (Johnson et al., 2016).

Each encounter is serialized as a self-contained JSON bundle with stable evidence identifiers and a per-case leakage flag used to monitor and exclude cases in which the primary diagnosis appears verbatim in the narrative.

3.3 Expert-Derived Gold Triangulation

The benchmark derives its gold signals by *repurposing and triangulating* high-quality pre-existing expert annotations and objective facts (Table 1). For example, image findings are extracted directly from the *radiologist’s own report* (FINDINGS/IMPRESSION sections), ensuring that the multimodal evidence reflects clinician-authored ground truth.

To make billed-code diagnoses defensible despite the temporal gap between our 24-hour prediction window and discharge-based coding, we apply a strict triangulation protocol. Each candidate diagnosis is corroborated against three *independent* sources:

²Derived from the MIMIC-IV-Ext clinical-decision-support resource on PhysioNet.

Table 1: **Gold signals** triangulated from expert labels, objective facts, and deterministic procedures.

Construct	Gold source (existing / objective)
Image findings	radiologist report findings (FINDINGS/IMPRESSION of the linked study)
Diagnoses	billed ICD, triangulated (code $\wedge$ narrative $\wedge$ structured; level ≥ 2)
Actions / referrals	clinician-approved referrals, triage acuity, disposition
Evidence salience	objective abnormality (flagged labs, fired rules, positive findings)
Guideline safety	deterministic engine over sourced rules (Phansalkar; Beers; FDA)
Provenance	LLM-judge support + Shapley/-counterfactual tests (§3.4)

- **CODE:** a human-assigned billing ICD label.
- **NARRATIVE:** an explicit mention in the clinical notes, judged by two independent extractors (a BioLORD embedding match and an LLM chart-reviewer). (Remy et al., 2022)
- **STRUCTURED:** an objective lab signature consistent with the diagnosis (e.g., creatinine > 1.5 mg/dL for acute kidney injury).

A label’s corroboration level is the number of sources that fire. The *triangulated-gold* subset (level ≥ 2) serves as our high-precision evaluation target. Finally, drug safety is evaluated with a deterministic engine over authoritative rules—the ONC high-priority drug–drug interaction list (Phansalkar et al., 2012), the AGS Beers Criteria 2023 (2023 American Geriatrics Society Beers Criteria Update Expert Panel, 2023), and FDA Structured Product Labels (boxed warnings and contraindications)—normalized via WHO ATC / NLM RxNorm so that every flagged violation traces to a specific source row or FDA set_id (full breakdown in App. E).

3.4 Provenance via Causal Attribution

Rather than asking only “is this claim supported?”, we define provenance *causally* through interventions on the discrete, typed evidence units we control (Abraham et al., 2022; Gat et al., 2024). While each modality could in principle admit a bespoke edit (e.g., range-normalizing labs or negating text), such edits risk introducing out-of-distribution artifacts and, for images, generative hallucinations. We therefore adopt a single, artifact-free operator throughout: strict unit *removal* $do(e_i \leftarrow \emptyset)$. Because every unit, including radiologist-report

sentences, is represented as a discrete element, removal simply drops it from the verifier’s context; we perform no pixel masking or in-painting. The separate question of whether models exploit visual pixels rather than textual findings is evaluated through whole-image ablation (§4.3).

(W) Attribution to the world. For a claim c , let $C_c = E \cup R$ be the set of all evidence units the model cited for that claim, including both patient-evidence citations and rule traces from §3.1. An independent held-out LLM judge provides a set-support function

$$S(c, A) \in \{0, 1\}, \quad A \subseteq C_c, \quad (3)$$

where A is any subset (coalition) of the cited units and $S(c, A)=1$ iff A jointly supports claim c . Evaluating S on different subsets of C_c is the removal operator; from it we score each cited unit $e_i \in C_c$ on four properties:

(i) **Sufficiency:** $S(c, \{e_i\})=1$; e_i alone supports the claim.

(ii) **Necessity:** $S(c, C_c)=1$ yet $S(c, C_c \setminus \{e_i\})=0$; removing e_i breaks support (the counterfactual $do(e_i \leftarrow \emptyset)$).

(iii) **Soundness:** a Shapley credit that shares credit fairly when evidence is overdetermined,

$$\phi_i^W = \sum_{T \subseteq C_c \setminus \{e_i\}} w_{|T|} [S(c, T \cup \{e_i\}) - S(c, T)], \quad (4)$$

where T is a coalition of the *other* cited units, the bracket is e_i ’s marginal contribution to support given T , and $w_{|T|} = \frac{|T|! (|C_c| - |T| - 1)!}{|C_c|!}$ is the Shapley weight that averages this marginal over all coalition orderings (Shapley, 1953). A citation is *sound* iff $\phi_i^W > \tau$.

(iv) **Recall:** the fraction of *salient* units (flagged-abnormal labs, fired guideline rules, positive image findings) that the artifact cites at all, exposing evidence the model ignored.

(M) Attribution to the model. The world axis certifies that a citation *is* supportive, not that the model’s computation *used* it. For open-weight systems we localize causal use by attention knock-out (Heimersheim and Nanda, 2024; Geiger et al., 2024): we score the model’s own committed claim, then make a cited unit e_i invisible to attention by masking the key positions of its tokens while holding all sequence positions fixed, and read the drop

$$\Delta_i^M = \log P(c \mid X) - \log P(c \mid X_{\setminus e_i}), \quad (5)$$

Table 2: **Feature comparison** of PROVENA-MED against existing medical evaluation frameworks. While prior benchmarks target isolated capabilities, PROVENA-MED uniquely unifies multimodal inputs, staged interactive diagnosis, and causally-verified claim provenance with first-class abstention.

Benchmark	Family	Multi modal	Staged reasoning	Claim provenance	Abstain	Interactive	Primary Focus
MedQA (Jin et al., 2021)	Text QA	✗	✗	✗	✗	✗	Decontextualized exam knowledge.
DrugEHRQA (Bardhan et al., 2022)	EHR QA	<i>partial</i>	✗	✗	✗	✗	Table and text retrieval.
EHRXQA (Bae et al., 2023)	EHR QA	✓	✗	✗	✗	✗	Multimodal (Table+Img) VQA.
MS-CXR (Boecking et al., 2022)	Vision	<i>partial</i>	✗	<i>partial</i>	✗	✗	Localized radiology grounding.
AMIE (Tu et al., 2025)	Dialogue	✗	✓	✗	✗	✓	Conversational history-taking.
MM-AMIE (Saab et al., 2026)	Dialogue	✓	✓	✗	✗	✓	Multimodal diagnostic chat.
PROVENA-MED (Ours)	Unified	✓	✓	✓	✓	✓	Verifiable, budget-limited active reasoning.

where $X_{\setminus e_i}$ is the input with e_i ’s tokens attention-masked; a citation is *used* iff $\Delta_i^M > \tau_M$. This length-preserving mask enacts the removal counterfactual $do(e_i \leftarrow \emptyset)$ *inside* the network without per-layer surgery, so it transfers across open-weight architectures unchanged.

We apply this intervention to all open-weight models in our evaluation panel. The threshold τ_M is calibrated using knockouts of randomly selected uncited units as a null distribution. Because attention masking may not remove all information carried through residual connections, we additionally validate the measure with a residual-stream intervention that mean-ablates the cited unit’s hidden representation at every layer (Heimersheim and Nanda, 2024; Geiger et al., 2024); validation results are reported in App. E. Closed models, which do not expose the required internal states, are evaluated on W only. Ultimately, reading the W and M axes jointly places every citation into one of four categories: *true provenance* ($W+M+$), *post-hoc rationalization* ($W+M-$, supportive but unused), *misgrounded reliance* ($W-M+$, used but unsupportive—the unsafe case), or *decorative* ($W-M-$). Clinical examples are provided in Table 8 (App. E).

3.5 Metrics and Baselines

Our primary evaluation metric is **MM-AIS**, which aggregates the per-citation world (W) and model (M) scores defined in §3.4, stratified by modality. We report four main quantities: (1) *citation validity*,

the fraction of cited IDs that resolve to a valid evidence unit; (2) *attribution yield*, defined as validity multiplied by citation precision, where precision is the fraction of citations that are sufficient to support the associated claim, $S(c, e_i) = 1$; (3) the distribution of citations across the four $W \times M$ provenance categories; and (4) *salient-evidence recall*, which measures coverage of evidence deemed necessary for the claim.

Our baselines span both general-purpose open-weight models (Gemma-3, Llama-3, and Mistral) and clinically adapted models (Med42 and MedGemma). Vision-capable models are evaluated under both text-only and raw-pixel inputs to isolate the contribution of visual information. We also evaluate the Claude family (Haiku 4.5, Sonnet 5, and Opus 4.8); because these models do not expose the internal states required for the M intervention, they are evaluated on the W axis only. Claude models are accessed through a retention-disabled endpoint consistent with the PhysioNet data-use agreement.

4 Results

We evaluate ten open-weight models across all five cohorts under identical scaffolding. Table 3 reports diagnostic accuracy, attribution, the $W \times M$ faithfulness quadrants, guideline grounding, and unsafe prescribing on one row per model. We rank models by *attribution yield* (citation validity $\times$ judge precision), so that neither sparse citing nor fabricated IDs can inflate the headline score.

Table 3: **PROVENA-MED provenance leaderboard.** Full-information evaluation across five cohorts. Models are ranked by **Yield** (validity $\times$ precision). **Diagnosis**: Hit@1. $W \times M$: supportive and computationally used citations (**True**) vs. unsupportive reliance (**Misgr.**). **Safety**: unsafe-prescribing rate. Best per column in **bold**; $\dagger$ clinical-specialized; n/a = no valid citations. Interactive results are in Table 5.

System	Diagnosis $\uparrow$	Attribution to world (W) $\uparrow$				$W \times M$ faithfulness			Safety	
	Hit@1	Valid.	Prec.	Recall	Yield	Sound	Guide-P	True $\uparrow$	Misgr. $\downarrow$	Unsafe $\downarrow$
MedGemma-27B $\dagger$	0.17	0.77	0.55	0.27	0.42	0.52	1.00	0.49	0.22	0.163
gemma-3-27b	0.18	0.89	0.37	0.37	0.34	0.32	0.49	0.31	0.17	0.173
gemma-3-12b	0.16	0.90	0.35	0.43	0.31	0.25	0.56	0.27	0.26	0.133
gemma-3-4b	0.14	0.78	0.23	0.30	0.18	0.25	0.50	0.20	0.30	0.177
Llama-3.1-8B	0.12	0.39	0.44	0.16	0.17	0.50	0.59	0.38	0.13	0.400
Med42-8B $\dagger$	0.16	0.33	0.55	0.10	0.17	0.48	0.48	0.13	0.06	0.247
BioMistral-7B $\dagger$	0.15	0.81	0.18	0.15	0.14	0.48	0.04	0.06	0.06	0.057
MedGemma-4B $\dagger$	0.12	0.26	0.36	0.07	0.09	0.25	0.55	0.24	0.17	0.127
Llama-3.2-3B	0.10	0.39	0.18	0.07	0.06	0.16	0.19	0.12	0.10	0.200
Mistral-7B	0.14	0.13	0.22	0.03	0.03	n/a	0.78	n/a	n/a	0.180

Models: Gemma 3 (Kamath et al., 2025); Llama 3 (Grattafiori et al., 2024); Mistral 7B (Jiang et al., 2023); BioMistral (Labrak et al., 2024); Med42-v2 (Christophe et al., 2024); MedGemma (Sellergren et al., 2025). Judges: Llama-3.3-70B-Instruct (Grattafiori et al., 2024) for the W axis, and gpt-oss-20b (OpenAI, 2025) as the independent cross-check (App. E.1).

4.1 No model is faithful, and the axes dissociate

The best attribution yield is 0.42 (MedGemma-27B), and most models fall below 0.2: even the strongest system grounds well under half of what it cites in supporting evidence.

Citation Hallucination vs. Coverage. The leaderboard exposes two distinct failures. *Citation hallucination* is severe: Med42-8B and MedGemma-4B invent the majority of their citation IDs (validity 0.33 and 0.26). A naive precision-only score would rank Med42-8B near the top (0.55) despite two-thirds of its references being fabricated, which is why we rank by attribution yield. At the same time, salient-evidence *coverage* consistently trails precision (Fig. 3), with imaging and guideline evidence suffering the most.

Clinical Fine-Tuning is Not a Panacea. Within the gemma-3 family, scale helps predictably (yield $0.18 \rightarrow 0.31 \rightarrow 0.34$), but clinical fine-tuning guarantees neither faithfulness nor safety. Excluding MedGemma-27B, specialized clinical models trail the general gemma-3 family on provenance. Clinical specialization also does not correlate with safety: the specialized models include both the safest system (BioMistral-7B, unsafe rate 0.057) and one of the least safe (Med42-8B, 0.247). Unsafe prescribing varies by roughly $7\times$ across the panel, driven heavily by dangerous opioid-benzodiazepine co-prescribing (which Llama-3.1-8B alone commits 88 times; App. E).

Faithfulness, Safety, and Accuracy Dissociate.

The leaderboard columns do not co-vary. The yield leader (MedGemma-27B, 0.42) is *not* the accuracy leader (gemma-3-27b, Hit@1 0.18); Med42-8B is competitive on accuracy (Hit@1 0.16) yet hallucinates two-thirds of its citation IDs (validity 0.33) and posts one of the worst safety rates, making it a canonical *accurate-but-unverifiable* case; and MedGemma-27B itself leads yield while never grounding the guideline rules it relies on. *Faithfulness, safety, and accuracy are separate, non-entailed capabilities.* For that reason, we deliberately avoid collapsing them into a single scalar “trustworthiness” score.

Stability Across Settings. Provenance performance is relatively stable across cohorts (Fig. 2), suggesting that citation-grounding deficits are not specific to a single dataset. Citation validity is highest on eICU, where evidence identifiers are most structured. Diagnostic accuracy is less robust: Hit@1 falls to 0.02–0.07 on eICU, compared with 0.20–0.32 on ED. Thus, cross-system diagnostic transfer degrades substantially even when citation grounding remains comparatively stable.

Frontier capability improves provenance but does not saturate the benchmark.

Because the W axis is model-agnostic (computed from generated text and cited IDs alone), we add a frontier reference across the Claude family (cardiac cohort, $n=40$, using the same held-out judge and cases; Table 4)). Two key observations emerge. First, provenance quality scales consistently with model

capability: Claude Haiku 4.5 hallucinates a majority of its citation IDs (validity 0.23, yield 0.07), behaving similarly to the smaller open-weight models, whereas Sonnet 5 and Opus 4.8 demonstrate stronger attribution (validity 0.98/0.88, yield 0.32/0.36). Second, while the larger models improve citation validity and recall, they do not yet exceed the highest attribution yield observed in our open-weight panel (gemma-3-27b, yield 0.42). Ultimately, frontier capability improves citation coverage and honesty, but it does not fully solve provenance, confirming that the benchmark is not yet saturated.

Table 4: **Frontier reference (Claude family).** The Claude family vs. the best open model on the W axis (cardiac cohort, same $n=40$ cases and held-out judge). Provenance scales with capability (Haiku→Sonnet→Opus), yet even Opus does not exceed the best open model on attribution yield.

System	Valid.	Prec.	Recall	Yield
gemma-3-27b (best open)	0.986	0.474	0.348	0.468
Claude Opus 4.8	0.880	0.410	0.388	0.361
Claude Sonnet 5	0.981	0.329	0.614	0.323
Claude Haiku 4.5	0.226	0.327	0.115	0.074

4.2 What causal and mechanistic attribution reveal

Soundness lags precision: citation is overdetermined. Replacing singleton precision with Shapley credit over evidence coalitions and a removal counterfactual (§3.4) makes provenance strictly harder for every model. On the cardiac cohort, the Shapley-sound fraction sits well below singleton sufficiency (e.g., 0.33 vs. 0.56 for gemma-3-27b). This occurs because much of the cited evidence is *redundant* rather than individually attributable. Correlational precision therefore overstates faithfulness; the causal axis removes that over-credit.

Misgrounded reliance: the unsafe quadrant. The M attribution axis (attention knockout) verifies whether the model’s computation actually routes through a cited unit. We validate this with a control showing that knocking out *uncited* units leaves the claim probability essentially unchanged ($\Delta_{\text{cited}}^M \gg \Delta_{\text{control}}^M$). Crossing the M and W axes exposes a failure mode that standard text metrics miss: *misgrounded reliance* ($W-M+$), where the model demonstrably depends on evidence that does not clinically support its claim. This unsafe behavior is prevalent, reaching a rate of 0.30 for gemma-3-

4b and 0.22 even for the leading MedGemma-27B. Conversely, *true provenance* ($W+M+$) never exceeds 0.49, and weaker models frequently cite purely for show (decorative $W-M-$ up to 0.66). Grounding a citation and actually using it are therefore distinct properties; only their conjunction guarantees faithful reasoning.

The provenance results are robust to threshold, judge, and intervention choices.

The $W \times M$ results remain stable across four robustness checks. **(i) Threshold sensitivity:** τ_M is calibrated to the 85th percentile of the uncited-control null, and sweeping $\tau_M \in [0.01, 0.2]$ leaves the true-provenance model ranking unchanged (Spearman $\rho=1.0$). The Shapley soundness threshold is similarly stable ($\rho=0.97-1.0$). **(ii) Judge validity:** on 700 citation–claim pairs with support labels determined by objective clinical criteria, the held-out judge achieves 0.87 accuracy and 0.91 precision. **(iii) Judge robustness:** paraphrased judge prompts agree with the canonical prompt on 92–94% of citations ($\kappa=0.84-0.86$). An independent non-Llama judge (gpt-oss; 0.89 accuracy on the objective set) also shows substantial agreement ($\kappa=0.75$) and preserves the model ranking ($\rho=0.93$), reducing concern that the results reflect judge-family preference. **(iv) Intervention robustness:** residual-stream activation patching corroborates attention knockout (Pearson $r=0.83$; 87% agreement on $M+$ classifications). Full results are in App. E.1.

4.3 Multimodality and interaction

Seeing the pixels barely helps. For the vision-capable models, we contrast a *text* condition, in which image evidence is provided as the radiologist’s report, with a *pixel* condition, in which the model receives the raw radiograph, while holding the rest of the patient bundle fixed (Table 6). Visual perception is not transformative: showing gemma-3-12b the image improves its IMAGE_FINDING precision (0.25 → 0.36) but hurts its citation validity (0.99 → 0.90). Meanwhile, MedGemma-27B’s overall precision rises slightly (0.53 → 0.59) but its image recall drops substantially. Current “multimodal” clinical models still appear to rely more on textual findings than on native visual reasoning.

Two paradigms of interactivity reveal complementary failure modes. We evaluate Llama-3.1-8B (the same backbone as the leaderboard’s interactive baseline) under two operationally distinct interactivity protocols on all five cohorts (Table 5).

Table 5: **Interactive evidence seeking across 5 cohorts (Llama-3.1-8B, $n=1500$).** Oracle Hit@5 measures diagnostic yield under progressively revealed evidence (s_0 , +history, +all). The cost-aware agent selects evidence under a 5-unit budget; the temporal agent chooses *wait* or *commit* over four reveal ticks. We report Hit@5, evidence cost, commit time, and force-termination rate. Best oracle Hit@5 per cohort is **bolded**.

Cohort	Oracle Hit@5 trajectory $\uparrow$			Cost-aware (budget=5)		Temporal (passive, $T_{\max} = 3$)		
	s_0 only	+history	+all	Cost/5 $\downarrow$	Hit@5 $\uparrow$	Hit@5 $\uparrow$	$\bar{t}_{\text{commit}}$	forced T_3
ED	0.013	0.397	0.370	2.25	0.227	0.303	3.00	17%
Cardiac_MM	0.103	0.183	0.203	4.26	0.147	0.167	3.00	15%
ICU_MM	0.137	0.180	0.213	1.85	0.163	0.173	3.00	16%
eICU	0.077	0.083	0.087	3.64	0.070	0.057	3.00	5%
MIMIC-III	0.127	0.267	0.247	4.14	0.243	0.217	3.00	13%

Table 6: **Text-vs-pixel contrast** on the ICU cohort: does the model seeing the chest radiograph (vs. the radiologist report as text) change image-finding grounding? $\text{Img-P/Img-R} = \text{IMAGE_FINDING precision/recall}$.

System	Cond.	Valid.	Prec.	Img-P	Img-R
gemma-3-12b	text	0.99	0.33	0.25	0.31
gemma-3-12b	pixels	0.90	0.33	0.36	0.35
MedGemma-27B [†]	text	0.83	0.53	0.58	0.43
MedGemma-27B [†]	pixels	0.68	0.59	0.55	0.33

In the *cost-aware* track, the agent has a typed action menu and a budget of 5 cost units that it can see; in the *temporal* track, evidence accrues passively across four clinical ticks (T_0 admission $\rightarrow T_1$ initial assessment $\rightarrow T_2$ early labs $\leq 12\text{h} \rightarrow T_3$ late workup) and the agent’s only choice each turn is *wait* or *commit*. The two protocols hold the model and the evidence fixed; they differ only in how the cost of waiting is communicated.

Cost-aware evidence seeking depends on the initial information state. Models do not simply over-request evidence; their behavior depends on what is already visible at s_0 . With richer initial states, they often commit prematurely: ICU_MM and ED request no additional evidence in 58% and 28% of cases, respectively. With sparse initial states, as in Cardiac_MM and MIMIC-III, they instead spend nearly the full budget ($\approx 4.2/5$ cost units). Neither extreme is efficient. More evidence is also not uniformly helpful. In ED, oracle Hit@5 falls from 0.397 with history alone to 0.370 when all available evidence is revealed, suggesting that additional labs and home medications can introduce information the model does not reliably integrate. Consequently, evidence acquisition alone does not close the performance gap: the best agent still trails the +history oracle by 0.17 Hit@5 in ED. On eICU, performance remains low under every reveal policy ($\text{Hit@5} \leq 0.087$), indicating a

broader out-of-distribution limitation rather than an evidence-selection failure alone.

Temporal: without explicit cost, models wait for all evidence. When evidence arrives passively and the model can only *wait* or *commit*, the cohort-dependent variation seen in the cost-aware setting largely disappears. Across all five cohorts, the model never voluntarily commits before the final tick, T_3 ; mean commit time is therefore 3.00/3 in every cohort. Even at T_3 , the model continues to choose *wait* in 5–17% of cases and must be force-terminated. This behavior contrasts with the cost-aware setting, where an explicit budget induces substantial variation in when the model stops and how much evidence it acquires. Also, Hit@5 at T_3 remains 3–7 percentage points below the corresponding +all oracle across cohorts. The temporal setting produces little variation in stopping behavior: the model consistently defers commitment to the final tick. This contrast suggests that the form in which evidence cost is presented can substantially influence when the model chooses to commit.

5 Conclusion

We presented PROVENA-MED, a benchmark for provenance-faithful multimodal clinical reasoning. Beyond diagnostic accuracy, it evaluates whether claims are grounded in specific multimodal evidence, bounded by safety guidelines, and appropriately deferred when information is insufficient. Two design choices distinguish it from prior work: an interactive evaluation track that scores models as budget-limited evidence-seekers rather than one-shot readers, and a construction protocol that triangulates gold signals from pre-existing expert labels, objective clinical facts, and counterfactual provenance tests without requiring new clinician annotation. We release all benchmark assets under the same data use agreements as the source.

Acknowledgments

This research was supported by the AI Makerspace (RRID:SCR_028058) and the Partnership for an Advanced Computing Environment (PACE) (RRID:SCR_027619) at the Georgia Institute of Technology. We also gratefully acknowledge funding and fellowships that contributed to this work, including a Wallace H. Coulter Distinguished Faculty Fellowship, a Petit Institute Faculty Fellowship, and research funding from Amazon and Microsoft Research awarded to Professor May D. Wang.

6 Limitations

PROVENA-MED evaluates verifiable, evidence-grounded reasoning using reproducible clinical anchors rather than holistic expert judgment. Its targets are derived from expert and objective signals, including radiology reports, diagnosis codes, structured abnormalities, clinician actions, and traceable guideline rules. Because these signals are not exhaustive, we triangulate independent sources where possible and evaluate provenance at the cited-evidence level rather than relying on a single outcome label. The benchmark has several scope limitations: image provenance is primarily finding-level rather than spatial; safety evaluation covers the implemented guideline rule library rather than all medication risks; and cohorts are limited to hospital settings. The white-box M axis requires model internals and is therefore limited to open-weight models, while the model-agnostic W axis can evaluate closed models when data-use requirements permit. We evaluate the Claude frontier family on W through a retention-disabled, PhysioNet-compliant endpoint; broader closed-model evaluation is constrained by the same requirement. Derived labels, prompts, scripts, and synthetic meta-data are redistributable, while the underlying MIM-IC/eICU records remain credential-gated. These boundaries position PROVENA-MED as an auditable benchmark for evidence-grounded clinical reasoning, not a certification of clinical readiness.

7 Ethical Considerations

PROVENA-MED is built from credentialed, de-identified clinical resources and must be used under the corresponding data-use agreements. The benchmark, prompts, outputs, and evaluation scripts are research artifacts intended to study model behavior under controlled evaluation conditions; they are not clinical decision-support tools, and model outputs

should not be used to guide patient care. Because clinical datasets can reflect demographic, institutional, and practice-pattern biases, results should be interpreted with attention to subgroup performance where demographic fields are available rather than only aggregate scores. High provenance scores indicate auditability with respect to the benchmark's evidence units; they do not establish clinical correctness, fairness, or safety for deployment.

References

- 2023 American Geriatrics Society Beers Criteria Update Expert Panel. 2023. [American geriatrics society 2023 updated AGS Beers Criteria for potentially inappropriate medication use in older adults](#). *Journal of the American Geriatrics Society*, 71(7):2052–2081.
- Eldar David Abraham, Karel D'Oosterlinck, Amir Feder, Yair Gat, Atticus Geiger, Christopher Potts, Roi Reichart, and Zhengxuan Wu. 2022. CEBaB: Estimating the causal effects of real-world concepts on NLP model behavior. In *Advances in Neural Information Processing Systems*.
- Julián N Acosta, Guido J Falcone, Pranav Rajpurkar, and Eric J Topol. 2022. Multimodal biomedical ai. *Nature medicine*, 28(9):1773–1784.
- Seongsu Bae, Daeun Kyung, Jaehee Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kweon, Jeongwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. 2023. EHRXQA: A multi-modal question answering dataset for electronic health records with chest X-ray images. In *Advances in Neural Information Processing Systems*.
- Jayetri Bardhan, Anthony Colas, Kirk Roberts, and Daisy Zhe Wang. 2022. Drugehrqa: A question answering dataset on structured and unstructured electronic health records for medicine related queries. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1083–1097.
- Benedikt Boecking, Naoto Usuyama, Shruthi Bannur, Daniel C Castro, Anton Schwaighofer, Stephanie Hyland, Maria Wetscherek, Tristan Naumann, Aditya Nori, Javier Alvarez-Valle, and 1 others. 2022. Making the most of text semantics to improve biomedical vision–language processing. In *European conference on computer vision*, pages 1–21. Springer.
- Emma Chen, Aman Kansal, Julie Chen, Boyang Tom Jin, Julia Rachel Reisler, David A. Kim, and Pranav Rajpurkar. 2023. Multimodal clinical benchmark for emergency care (MC-BEC): A comprehensive benchmark for evaluating foundation models in emergency medicine. In *Advances in Neural Information Processing Systems*. Datasets and Benchmarks Track.
- Clément Christophe, Praveen K. Kanithi, Tathagata Raha, Shadab Khan, and Marco A. F. Pimentel. 2024.

- Med42-v2: A suite of clinical LLMs. *arXiv preprint arXiv:2408.06142*.
- Pat Croskerry. 2005. The theory and practice of clinical decision-making. *Canadian Journal of Anesthesia*, 52:R1–R8.
- Arthur S Elstein, Lee S Shulman, and Sarah A Sprafka. 1978. *Medical problem solving: An analysis of clinical reasoning*. Harvard University Press.
- Dennis Fast, Lisa C. Adams, Felix Busch, Conor Fallon, Marc Huppertz, Robert Siepmann, Philipp Prucker, Nadine Bayerl, Daniel Truhn, Marcus Makowski, Alexander Löser, and Keno K. Bressem. 2024. [Autonomous medical evaluation for guideline adherence of large language models](#). *npj Digital Medicine*, 7:358.
- Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Yair Gat, Nitay Calderon, Amir Feder, Alexander Chapanin, Amir Sharma, and Roi Reichart. 2024. Faithful explanations of black-box NLP models using LLM-generated counterfactuals. In *International Conference on Learning Representations (ICLR)*.
- Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah D. Goodman, Christopher Potts, and Thomas Icard. 2024. Causal abstraction: A theoretical foundation for mechanistic interpretability. *arXiv preprint arXiv:2301.04709*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Maxime Griot, Coralie Hemptinne, Jean Vanderdonckt, and Demet Yuksel. 2025. [Large language models lack essential metacognition for reliable medical reasoning](#). *Nature Communications*, 16:642.
- Stefan Heimersheim and Neel Nanda. 2024. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silvana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghighi, Robyn Ball, Katie Shpanskaya, and 1 others. 2019. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, and 1 others. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.
- 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.
- Alistair Johnson, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023a. [MIMIC-IV-Note: Deidentified free-text clinical notes](#). *PhysioNet*.
- Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, and 1 others. 2023b. [MIMIC-IV, a freely accessible electronic health record dataset](#). *Scientific Data*, 10(1):1.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. 2019. [MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports](#). *Scientific Data*, 6(1):317.
- Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific Data*, 3(1):160035.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.
- Sunjun Kweon, Jiyoung Kim, Heeyoung Kwak, Dongchul Cha, Hangyul Yoon, Kwanghyun Kim, Jee-won Yang, Seunghyun Won, and Edward Choi. 2024. Ehrnoteqa: An llm benchmark for real-world clinical practice using discharge summaries. *Advances in Neural Information Processing Systems*, 37:124575–124611.
- Yanis Labrak, Adrien Bazoge, Emmanuel Morin, Pierre-Antoine Gourraud, Mickael Rouvier, and Richard Dufour. 2024. [BioMistral: A collection of open-source pretrained large language models for medical domains](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5848–5864, Bangkok, Thailand. Association for Computational Linguistics.
- Ricardo Bigolin Lanfredi, Mingyuan Zhang, William F. Auffermann, Jessica Chan, Phuong-Anh T. Duong, Vivek Srikumar, Trafton Drew, Joyce D. Schroeder, and Tolga Tasdizen. 2022. [REFLACX, a dataset of reports and eye-tracking data for localization of abnormalities in chest x-rays](#). *Scientific Data*, 9:350.
- Gyubok Lee, Hyeonji Hwang, Seongsu Bae, Yeonsu Kwon, Woncheol Shin, Seongjun Yang, Minjoon Seo, Jong-Yeup Kim, and Edward Choi. 2022. EHRSQL: A practical text-to-SQL benchmark for electronic health records. In *Advances in Neural Information Processing Systems*.

- Katie Matton, Robert Osazuwa Ness, John Gutttag, and Emre Kıcıman. 2025. Walk the talk? measuring the faithfulness of large language model explanations. In *International Conference on Learning Representations*.
- Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, and 1 others. 2025. [Towards accurate differential diagnosis with large language models](#). *Nature*, 642:451–457.
- Sewon Min, Kalpesh Krishna, Xinxu Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- OpenAI. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint arXiv:2508.10925*.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikanan Sankarasubbu. 2022. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on Health, Inference, and Learning (CHIL)*.
- Shobha Phansalkar, Amrita A. Desai, Douglas Bell, Eileen Yoshida, John Doole, Melissa Czochanski, Blackford Middleton, and David W. Bates. 2012. [High-priority drug–drug interactions for use in electronic health records](#). *Journal of the American Medical Informatics Association*, 19(5):735–743.
- Tom J. Pollard, Alistair E. W. Johnson, Jesse D. Raffa, Leo A. Celi, Roger G. Mark, and Omar Badawi. 2018. [The eICU collaborative research database, a freely available multi-center database for critical care research](#). *Scientific Data*, 5(1):180178.
- Hannah Rashkin, Vitaly Nikolaev, Matthew Lamm, Lora Aroyo, Michael Collins, Dipanjan Das, Slav Petrov, Gaurav Singh Tomar, Iulia Turc, and David Reitter. 2023. Measuring attribution in natural language generation models. *Computational Linguistics*, 49(4):777–840.
- François Remy, Kris Demuynck, and Thomas De-meester. 2022. [BioLORD: Learning ontological representations from definitions for biomedical concepts and their textual descriptions](#). In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1454–1465, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Khaled Saab, Chunjong Park, Tim Strother, Jan Freyberg, David G. T. Barrett, Yong Cheng, Wei-Hung Weng, David Stutz, Nenad Tomasev, Anil Palepu, and 1 others. 2026. [Advancing conversational diagnostic AI with multimodal reasoning](#). *Nature Medicine*.
- Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, and 1 others. 2025. MedGemma technical report. *arXiv preprint arXiv:2507.05201*.
- Lloyd S. Shapley. 1953. [A value for \$n\$ -person games](#). In Harold W. Kuhn and Albert W. Tucker, editors, *Contributions to the Theory of Games, Volume II*, number 28 in Annals of Mathematics Studies, pages 307–317. Princeton University Press.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, and 1 others. 2023. [Large language models encode clinical knowledge](#). *Nature*, 620:172–180.
- Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Po-Han Chang, Andrew Carroll, Chang Lau, Ryutaro Tanno, Iasonas Ktena, and 1 others. 2023. [Towards generalist biomedical AI](#). *arXiv preprint arXiv:2307.14334*.
- 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*, 642:442–450.
- Kevin Wu, Eric Wu, Kevin Wei, Angela Zhang, Allison Casasola, Teresa Nguyen, Sith Riantawan, Patricia Shi, Daniel Ho, and James Zou. 2025. [An automated framework for assessing how well LLMs cite relevant medical references](#). *Nature Communications*, 16:3615.
- Kerem Zaman and Shashank Srivastava. 2025. A causal lens for evaluating faithfulness metrics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29425–29449.
- Yang Zhou, Tan Li Hui Faith, Yanyu Xu, Sicong Leng, Xinxing Xu, Yong Liu, and Rick Siow Mong Goh. 2024. BenchX: A unified benchmark framework for medical vision-language pretraining on chest X-rays. In *Advances in Neural Information Processing Systems*. Datasets and Benchmarks Track.

A LLM Usage

Large Language Models (LLMs) were used solely as general-purpose assistive tools to help polish the manuscript’s language and to refine instructions within our prompt templates. Specifically, LLMs aided in improving grammar, clarity, and style, and in suggesting alternative phrasings for prompt templates. All scientific ideas, experimental design, and key arguments were conceived and written by the authors, and all factual statements were independently verified

B Dataset Details

This appendix provides a more detailed breakdown of the PROVENA-MED benchmark. While the main text summarizes the overall scale, Table 7 reports the distribution of source encounters across the five clinical cohorts, the composition of multimodal evidence units (structured EHR tables, clinical notes, and chest radiographs), and the reasoning and safety fields exposed to evaluation. These statistics make the benchmark’s scale, diversity, and evidence density explicit and auditable.

Each encounter is serialized as a JSON bundle of typed, stably identified evidence units together with the staged-artifact target, triangulated gold labels, the interactive reveal schema, and per-case flags. Below we show a de-identified multimodal cardiac admission (PhysioNet credentialed; PHI rendered as ____). This case contains 46 NOTE_SPAN, 15 TABLE_ROW, 100 IMAGE_FINDING, and 10 GUIDELINE_RULE units; we list a representative excerpt.

Initial state s_0 . Chief complaint, triage acuity, demographics, and the initial vital signs: the deliberately information-poor cold start used in the interactive track.

Evidence units (excerpt).

- NOTE_SPAN:hpi:0 — “____ is a ____ year old woman with HTN, HLD, DM2, obesity who presents as a transfer...”
- NOTE_SPAN:hpi:12 — “Over the next few days, she could not lay flat at night, she became very short of breath...”
- TABLE_ROW:lab:bicarbonate — “bicarbonate = 27.0”
- TABLE_ROW:lab:chloride — “chloride = 100.0”

Table 7: **PROVENA-MED benchmark composition.**

Source counts are extracted patient encounters across the five cohort families; bundle and modality statistics are aggregated over the leaderboard sample. Every evidence-unit count is auditable directly from the released JSON bundles. Vocabularies represent the union across all cohorts.

Property	Value
Source Corpus	
Cohort families	5
Source encounters (extracted)	23,446
Unique patients	22,573
Emergency Department	9,137 enc. / 9,137 pts.
Cardiac (MIMIC-IV + CXR)	4,761 enc. / 4,289 pts.
ICU multimodal (MIMIC-IV + CXR)	3,963 enc. / 3,636 pts.
eICU (external)	2,585 enc. / 2,573 pts.
MIMIC-III (external)	3,000 enc. / 2,938 pts.
Avg. encounters per patient	1.04
Frozen evaluation split (test + dev)	12,000
Bundle Composition & Modalities	
Evidence units in sampled bundles	42,306
TABLE_ROW (structured EHR)	14,990 (35.4%)
NOTE_SPAN (clinical notes)	16,980 (40.1%)
IMAGE_FINDING (chest CXR)	7,097 (16.8%)
GUIDELINE_RULE (rules)	3,239 (7.7%)
Reasoning Artifact & Pools	
Reasoning stages per case	5
Patient-citation types per claim	3
Safety statuses	4
Distinct gold diagnoses (union)	2,339
Guideline-rule library (sourced)	49
rules surfaced in eval	41 (83.7%)
Averages per Case	
Avg. evidence units / case	56.4
Avg. TABLE_ROW / case	20.0
Avg. NOTE_SPAN / case	22.6
Avg. IMAGE_FINDING / imaging case	17.8
Avg. GUIDELINE_RULE / case	4.3
Evaluation Scale	
Open-weight models evaluated	10
Vision-language models in pixel cond.	2
Bootstrap iterations (95% CIs)	1,000×

- IMAGE_FINDING:ct:0 — “CHEST WITHOUT CONTRAST. FINDINGS: ... Thyroid is unremarkable.”
- GUIDELINE_RULE:PH8-SSRI-MAOI — “[CONTRAINDICATED] SSRIs/SNRIs + MAO inhibitors: serotonin syndrome.” (Phansalkar 2012, Table 2)
- GUIDELINE_RULE:B5-WARFARIN-X — “[CAUTION] Warfarin + (amiodarone/ciprofloxacin/macrolide/TMP-SMX/SSRI): increased bleeding.” (Beers 2023)

Gold and flags. Triangulated primary diagnosis: *Acute systolic (congestive) heart failure* (corroboration level ≥ 2 : billed code $\wedge$ narrative). Additional gold includes type-2 diabetes and valvular disease. leak_primary = false, meaning the pri-

mary diagnosis does not appear verbatim in the shown narrative.

Reveal schema (interactive). The initial state exposes vitals only. Requestable actions then map deterministically from unit IDs: `take_history`→`NOTE_SPAN:hpi:*` (cost 1), `order_labs`→`TABLE_ROW:lab:*` (cost 2), `order_imaging`→`IMAGE_FINDING:*` (cost 3), and `consult_guidelines`→`GUIDELINE_RULE:*` (cost 0).

C Prompts

Staged Generation Prompt

System: “You are an experienced emergency medicine physician.”

User: the case header, followed by the evidence units ([ID] text, one per line) under “EVIDENCE UNITS (cite these IDs):”, then the instruction: “For EVERY claim, cite one or more evidence IDs from the list above (use the exact IDs; cite only listed IDs). For therapy/safety claims, also cite any applicable GUIDELINE_RULE IDs... If a stage cannot be supported, return an empty list and add a note to missing_information.”

Required JSON Schema

```
{
  "problem_list": [
    { "claim": "evidence": [ID] },
    { "diagnosis": "rank", "evidence": "workup": [
      { "action": "evidence", "therapy": [
        { "action": "evidence", "safety_status": "monitoring": [
          { "action": "evidence", "missing_information": [ ... ] }
        ]
      ]
    }
  ]
}
```

Provenance Judge: Set-Support $S(c, A)$

System: “You are a meticulous clinical evidence auditor. Judge only from the evidence given.”

User: “EVIDENCE: {the units in coalition A } CLAIM: { c } Does the EVIDENCE directly support the CLAIM?... Answer with exactly one word: SUPPORTS or UNSUPPORTED.” This same operator is reused to compute singleton precision, Shapley credit, and the removal counterfactual.

Triangulation Narrative Judge

System: “You are a careful clinical chart reviewer.”

User: “CLINICAL NOTES: {notes} Do these notes indicate the patient has, or is being treated for, “{dx}”? Answer... YES or NO.”

Vision (Pixel) Condition

Identical staged prompt, with the chest-radiograph image prepended to the user turn so the vision-language model sees the pixels alongside the text bundle.

Verifier choice. We score support with the held-out LLM judge above rather than a natural-language-inference (NLI) entailment model. In pilots, a PubMedBERT model fine-tuned on MedNLI over-credited TABLE_ROW citations through lexical overlap, under-credited IMAGE_FINDING citations, and changed model rankings relative to the LLM judge. We therefore use the LLM judge for all reported provenance results.

D Case Studies

Case 1: Faithful, Shapley-Sound (gemma-3-12b, admission 20051301)

For the problem-list claim “*acute systolic heart failure, evidenced by leg swelling, dyspnea on exertion, and orthopnea,*” the model cites NOTE_SPAN:hpi:0 (the HTN/HLD/DM2/obesity presentation) and NOTE_SPAN:hpi:12 (“could not lay flat at night... very short of breath”—orthopnea and dyspnea). Both IDs resolve, the judge returns SUPPORTS, and the removal counterfactual confirms that hpi:12 is *necessary*: dropping it removes the orthopnea/dyspnea support. This is a $W+$ citation and a representative true-provenance case.

Case 2: Citation That Does Not Resolve (Mistral-7B, same admission 20051301)

For a near-identical history claim, the model emits the citation [NOTE_SPAN:hpi:0]—wrapped in literal brackets—which does not match the bundle identifier NOTE_SPAN:hpi:0 and is therefore counted as a *validity* failure. The cited content is plausible, so a precision-only metric computed over *resolvable* citations would silently ignore the error; citation validity catches it. Malformed or invented identifiers of this kind drive Mistral-7B’s low validity (0.13 on this cohort) and illustrate why attribution yield (validity $\times$ precision) is the more honest

E Detailed Per-Axis Results

The consolidated leaderboard (Table 3) reports one headline scalar per evaluation axis. Here we provide the fuller decompositions. We begin with the $W \times M$ quadrants used throughout the paper (Table 8) and then turn to the per-axis breakdowns.

Drug safety. Each model proposes a differential and an initial medication list for the ED evaluation slice; we check the proposed medications with a deterministic rule engine assembled *entirely from authoritative sources*: the ONC high-priority drug–drug interaction list (Phansalkar et al., 2012), the AGS Beers Criteria 2023 (drug–disease, drug–drug, and renal tables, applied at age ≥ 65) (2023 American Geriatrics Society Beers Criteria Update Expert Panel, 2023), and FDA Structured Product Labels (contraindications and boxed warnings). Drug normalization uses WHO ATC classes from NLM RxNorm/RxCSS, and every flagged violation is traceable to a specific source row or FDA set_id. Table 9 decomposes the unsafe-recommendation rate from Table 3 into its two dominant violation types.

E.1 Robustness and Validation of the Provenance Axes

We stress-test the two design elements the provenance axes depend on—the LLM judge behind the W axis and the attention-knockout threshold behind the M axis—on the cardiac cohort (9 models with valid citations; the per-citation W cache reproduces the published $W \times M$ quadrant counts of Table 3 exactly).

τ_M and Shapley-threshold sensitivity. Table 10 sweeps the knockout threshold τ_M . Its operating point (0.05) sits at the 85th percentile of the uncited-control null ($\Delta_{\text{cited}}^M = +0.253$ vs. $\Delta_{\text{control}}^M = -0.011$). While the absolute quadrant fractions vary smoothly, the model *ranking* is invariant: True-provenance ranking Spearman $\rho = 1.000$ across $\tau_M \in [0.01, 0.15]$ (0.98 at 0.2), and the Shapley soundness threshold shows the same stability ($\rho = 0.97\text{--}1.00$ for $\tau \in [0.3, 0.7]$). The headline attribution-*yield* ranking is a function of W alone and is therefore τ_M -independent by construction.

Judge validity against objective clinical facts.

We validate the held-out judge without new clinician annotation by constructing 700 balanced (evidence, claim) pairs whose support is fixed by deterministic lab cutoffs (the same standard signatures used in triangulation): definitional positives (e.g. sodium 128 $\rightarrow$ hyponatremia), normal-value negatives, and cross-condition negatives. Table 11 reports the result: the deployed Llama-3.3-70B judge agrees with the objective labels at accuracy 0.866 (balanced accuracy 0.866, precision 0.908), and an independent, non-panel, non-Llama judge (gpt-oss) agrees at 0.891—so the axis is not an artifact of one judge’s idiosyncrasies. On the hardest cell, cross-condition negatives, agreement is 100%; the residual errors are conservative (the judge declines borderline single-lab positives).

Attention knockout corroborated by activation patching.

Because residual connections can in principle route a masked unit’s information around the attention block, we corroborate the M axis with an *independent* intervention on the residual stream itself: we replace the cited unit’s hidden state with the mean residual over the prompt at every decoder layer, then re-score the committed claim. Table 12 shows the two causal measures are highly correlated and near-identical in magnitude (pooled Pearson $r = 0.83$; up to 0.999; $M+$ classification agreement 86.9%), so attention knockout reflects genuine residual-level causal use rather than a masking artifact. (Med42-8B, a clinical fine-tune, is a partial outlier where residual mean-ablation occasionally raises the claim probability.)

F Additional Figures

Figure 4 provides a compact panel-level view of the tradeoffs already visible in the leaderboard, while Figures 2–3 unpack those tradeoffs by cohort and modality.

Table 8: **The MM-AIS $W \times M$ quadrants** with illustrative clinical examples. **W** asks whether the cited unit objectively supports the claim; **M** asks whether the model’s computation actually used it. The unsafe quadrant is $W-M+$ (*misgrounded reliance*)—the model demonstrably depends on evidence that does not support its claim.

	Model used it ($M+$)	Model did not use it ($M-$)
Supports ($W+$)	True provenance. Cites the elevated creatinine for an “acute kidney injury” claim; removing it both drops the verifier’s support and lowers the model’s claim probability.	Post-hoc rationalization. Cites an elevated BNP for “heart failure”—which does support it—but the output is unchanged when BNP is patched out: a correct-looking citation the model did not rely on.
Does not support ($W-$)	Misgrounded reliance (unsafe). Cites a “no focal consolidation” span for a “pneumonia” claim; the span does not support pneumonia, yet the output changes when it is removed—the model relied on unsupportive evidence.	Decorative. Cites an unrelated vital (e.g. heart rate) for a “type-2 diabetes” claim; neither objectively supportive nor used by the computation.

Table 9: **Unsafe-recommendation rate** on the ED cohort (full-information), scored against source-traceable rules. “Op+Bz” = opioid+benzodiazepine co-prescriptions (FDA boxed warning); “Allergy” = prescriptions matching a documented allergy. [†]clinical-specialized.

System	Unsafe rate (95% CI) ↓	Op+Bz	Allergy
BioMistral-7B [†]	0.057 ± 0.027	3	6
MedGemma-4B [†]	0.127 ± 0.037	11	13
gemma-3-12b	0.133 ± 0.038	14	12
MedGemma-27B [†]	0.163 ± 0.044	20	16
gemma-3-27b	0.173 ± 0.045	17	11
gemma-3-4b	0.177 ± 0.042	20	14
Mistral-7B	0.180 ± 0.043	17	17
Llama-3.2-3B	0.200 ± 0.045	37	13
Med42-8B [†]	0.247 ± 0.049	37	15
Llama-3.1-8B	0.400 ± 0.053	88	14

Table 10: τ_M **sensitivity** (cardiac, 7,419 cited / 2,590 control citations). “cited > τ ” and control FPR are the fraction of each population exceeding τ_M ; ρ is the Spearman rank correlation of the per-model True-provenance rate vs. the operating point $\tau_M=0.05$.

τ_M	cited > τ	control FPR	ρ (True-prov)
0.01	50.5%	32.4%	1.000
0.03	41.8%	20.1%	1.000
0.05	36.7%	14.8%	1.000
0.10	30.3%	7.3%	1.000
0.15	26.5%	4.3%	1.000
0.20	24.4%	2.8%	0.983

Table 11: **Judge validation** against objective clinical definitions ($n=700$) and **judge robustness** on real cardiac citations. Paraphrase = two reworded support prompts vs. the canonical one; inter-judge = an independent gpt-oss judge vs. Llama-3.3-70B. ρ is per-model rank stability.

Objective validation	acc	bal-acc	prec
Llama-3.3-70B (deployed)	0.866	0.866	0.908
gpt-oss (independent)	0.891	0.891	0.898
Robustness (agreement)	agree	κ	ρ
Prompt paraphrase 1	93.6%	0.862	0.967
Prompt paraphrase 2	92.1%	0.835	0.983
Independent judge (gpt-oss)	87.8%	0.747	0.933

Table 12: **Activation-patching corroboration** of attention knockout (1,723 cardiac citations). r and $M+$ agreement compare per-citation knockout Δ^M against residual mean-ablation.

Model	n	Pearson r	$M+$ agree	$\overline{\Delta}_{\text{attn}}/\overline{\Delta}_{\text{act}}$
Llama-3.1-8B	336	0.913	80.1%	0.52/0.53
Llama-3.2-3B	522	0.999	95.8%	0.15/0.14
Med42-8B [†]	865	0.698	84.2%	0.07/−0.18
Pooled	1723	0.833	86.9%	

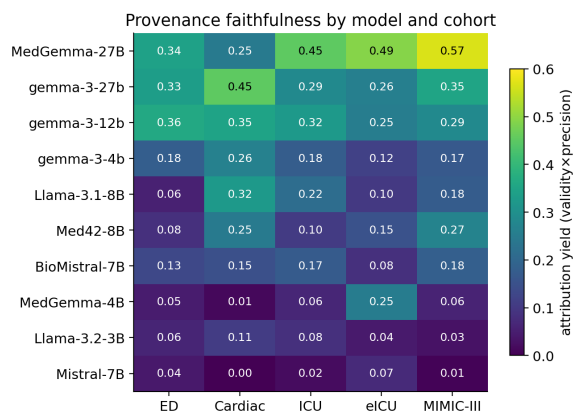


Figure 2: Attribution yield (validity×precision) for every model×cohort. The gemma-3 family and MedGemma-27B lead; external cohorts (eICU, MIMIC-III) are not systematically harder than internal MIMIC-IV cohorts, and per-cohort rankings shift.

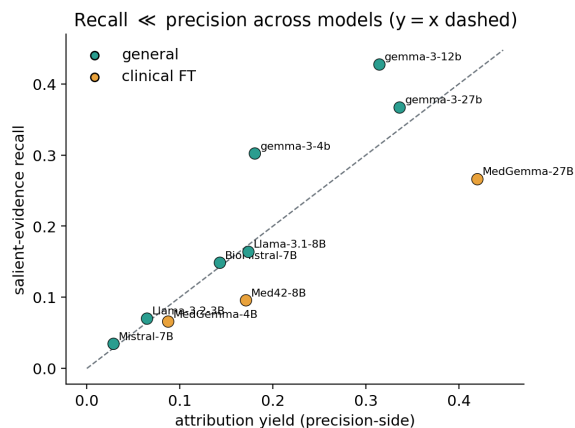


Figure 3: Salient-evidence recall vs. precision-side attribution yield, one point per model. Every model lies below the $y=x$ line: faithfulness on the citations a model makes far exceeds its coverage of the evidence it should cite.

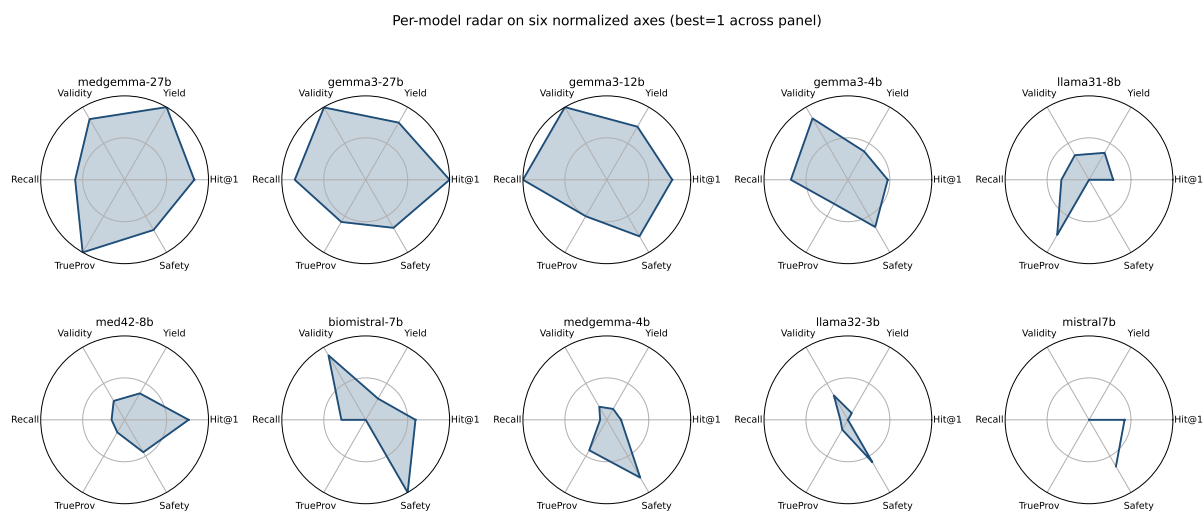


Figure 4: **Panel-normalized tradeoff view across six axes.** Each radar normalizes one model’s scores against the best value observed in the evaluated panel on Hit@1, attribution yield, citation validity, salient-evidence recall, true provenance ($W+M+$), and safety. No model dominates every axis simultaneously: MedGemma-27B is the broadest overall performer, the gemma-3 family is strongest on validity and diagnosis, and several weaker systems show sharply unbalanced profiles. The figure is intended as a qualitative summary of cross-axis dissociation; the underlying values remain the tabled metrics in the main text and appendix.