DEDUCTIVE CONSTRAINT SATISFACTION VS. PREVALENCE PRIORS: BENCHMARKING LLM LOGIC IN CLINICAL DIAGNOSTICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models are increasingly evaluated for complex decision-making, yet their ability to maintain logical invariance when deductive constraints conflict with statistical priors remains poorly characterized. We introduce DiagRare, a structured deductive benchmark of 696 clinical vignettes grounded in an expert-curated 58-disease biomedical ontology. By explicitly controlling present (*modus ponens*) and absent (*modus tollens*) evidence, DiagRare isolates logical constraint satisfaction from associative pattern matching. Evaluating five frontier LLMs reveals a stark 31.6% to 100% Top-1 accuracy spread and exposes three phenomena. First, the weakest model exhibits prevalence-driven mode collapse, funneling 278 of 476 errors into a single common diagnosis spanning 45 unrelated diseases (diagnostic entropy collapsing to 3.33/5.86 bits). Second, mid-tier reasoning models exhibit an inductive downgrade: Claude Sonnet 4.6 achieves 99.7% Top-3 accuracy on rare diseases but only 86.3% Top-1, indicating prevalence priors corrupt the final ranking of correctly deduced candidates. Third, we establish a deductive asymmetry: increasing explicit negative constraints monotonically improves Claude's accuracy from 84.5% to 97.0%, while increasing positive findings degrades it, demonstrating that *modus tollens* binds statistical priors more effectively than *modus ponens*.

Keywords: Large Language Models, Logical Reasoning, Constraint Satisfaction, Prevalence Anchoring, Clinical Diagnostics

1 INTRODUCTION

In structured domains like medical diagnostics, solving a case requires systematic deductive reasoning. An agent must identify a target d^* that explains all observed findings $\mathcal{F}^+$ while strictly avoiding conditions contradicted by absent findings $\mathcal{F}^-$. While LLMs encode vast clinical knowledge (1; 3), their outputs are inherently biased toward the baseline prevalence of tokens in their training corpora. This creates a fundamental tension: when a logical constraint isolates a rare outcome, does the LLM execute the deduction, or default to a statistical shortcut? Recent findings on greedy reasoning (2) and logical invariance (4) suggest this vulnerability is architecturally deep.

We utilize clinical differential diagnosis, where the correct answer may possess a prevalence as low as 2×10^{-5} to stress-test LLM deductive fidelity against base-rate pressure. We make three primary contributions. First, we introduce DiagRare, an expert-curated biomedical ontology of 58 diseases (28 rare, 30 common) yielding 696 difficulty-controlled vignettes with explicitly bounded positive and negative constraints (§2). Second, we formalize the inductive downgrade, proving mid-tier models possess the internal deductive capacity to identify rare targets (99.7% Top-3) but penalize them during final ranking due to prevalence priors (§4.2). Third, we demonstrate a deductive asymmetry: scaling explicit negative constraints (*modus tollens*) monotonically improves logic, while scaling positive evidence degrades it via associative interference (§4.3).

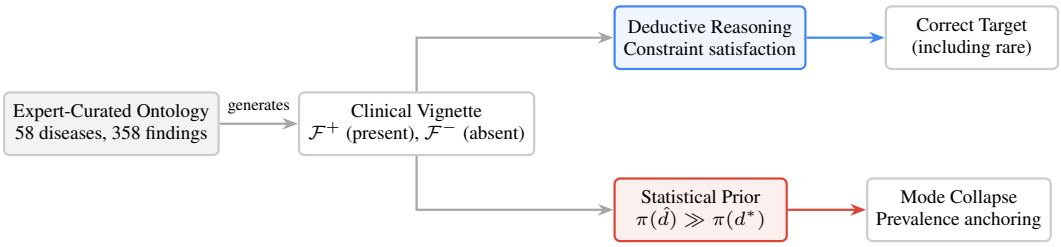

Figure 1: The DiagRare framework. An expert-curated ontology generates vignettes with explicit constraints. Capable models satisfy constraints deductively; weaker models default to prevalence priors.

2 The DiagRare Benchmark

We formalize differential diagnosis as a constraint satisfaction problem over an expert-curated ontology.

Definition 1 (Diagnostic Ontology)**.** *An ontology $\mathcal{O} = (\mathcal{D}, \mathcal{F}, R, E, S, \pi)$ maps diseases $\mathcal{D}$ to required (R), excluding (E), and supporting (S) findings $\mathcal{F}$, alongside a prevalence prior $\pi \in (0, 1)$.*

Definition 2 (Logical Consistency)**.** *A vignette $v = (\mathcal{F}^+, \mathcal{F}^-, d^*)$ is logically consistent if $R(d^*) \cap \mathcal{F}^- = \emptyset$ and $E(d^*) \cap \mathcal{F}^+ = \emptyset$. The first conjunct preserves* modus ponens*; the second enables* modus tollens *elimination.*

Our ontology features 58 diseases (28 rare: $\pi < 0.005$) and 358 clinical findings across 6 organ systems, grounded in Harrison's Principles of Internal Medicine. Prevalences span four orders of magnitude (2×10^{-5} to 0.20). Diseases within each system exhibit high symptom overlap to enforce strict differential logic. We algorithmically generate 696 vignettes (12 per disease) across three difficulty tiers (δ) by modulating $R(d)$ omission and injecting noise from $S(d)$ and $E(d)$. **Easy** (δ_1, $n = 232$) features all required findings, 3–5 exclusions, and no noise. **Medium** (δ_2, $n = 232$) omits up to 1 required finding, adds 2–4 exclusions, and 0–1 noise finding. **Hard** (δ_3, $n = 232$) omits 1–2 required findings, adds 1–3 exclusions, and 1–2 noise findings.

3 Experimental Setup

We evaluate five frontier LLMs zero-shot: Gemini 3 Thinking, DeepSeek-V3.2, Claude Sonnet 4.6, GPT-5.4 Thinking, and GPT-5.3 Instant. These models were selected to span the current capability frontier, including both reasoning-optimized (Thinking) and latency-optimized (Instant) inference modes across four independent providers. Models receive the 696 vignettes as structured text listing positive findings, negative findings, and a fixed list of the 58 disease names, returning Top-3 ranked diagnoses. We report Top-1/Top-3 accuracy with Wilson 95% CIs, quantifying prediction diversity via diagnostic entropy $H(M) = -\sum p_M(d) \log_2 p_M(d)$ (maximum 5.86 bits). We define prevalence anchoring on an error v if the prediction's prior heavily outweighs the target: $\pi(M(v)) > 1.5 \cdot \pi(d^*)$.

4 Results

4.1 Performance Stratification and Mode Collapse

Table 1 reveals a 68.4-point accuracy spread. GPT-5.3 Instant's low entropy ($H = 3.33$) signals massive mode collapse: it predicts just 22 of 58 diagnoses, channeling 278 of 476 errors entirely into Acute MI across 45 unrelated true targets. Overwhelmed by constraints, it abandons reasoning for the highest-prevalence token. Conversely, Gemini 3 Thinking maintains maximal entropy ($H = 5.86$) and perfect accuracy, proving logical invariance is architecturally sustainable. Difficulty modulation ($\delta_1 \rightarrow \delta_3$) produces only a maximum 5.6-point drop for models $> 88\%$ accuracy, proving the primary challenge is fine-grained differential resolution rather than missing evidence.

Table 1: Comprehensive results on DiagRare (696 vignettes). Accuracy is reported across overall, difficulty tiers (δ_1: easy, δ_2: medium, δ_3: hard), rarity, and organ systems. Δ: common minus rare Top-1 accuracy gap. H: diagnostic entropy ($H_{\max} = 5.86$). 95% Wilson CIs for Overall Top-1 are shown below model names.

	Overall (%)		By Difficulty (%)			By Rarity (%)		By Organ System (Top-1 %)						Metrics	
Model	Top-1	Top-3	δ_1	δ_2	δ_3	Comm.	Rare	Neuro	Metab	Immuno	Hemato	Pulmo	Cardio	Δ	H
Gemini 3 Th. [99.5,100]	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	0.0	5.86
DeepSeek-V3.2 [94.9,97.7]	96.6	96.6	96.6	96.6	96.6	100.0	92.9	100.0	100.0	90.9	100.0	83.3	100.0	7.1	5.86
Claude S. 4.6 [86.9,91.4]	89.4	99.6	91.8	90.1	86.2	92.2	86.3	79.2	86.4	96.2	93.8	86.1	97.2	5.9	5.72
GPT-5.4 Th. [85.5,90.3]	88.1	94.4	90.9	87.1	86.2	88.9	87.2	75.7	93.2	90.2	94.4	77.8	97.2	1.7	5.77
GPT-5.3 Inst. [28.3,35.2]	31.6	38.4	33.6	31.0	30.2	51.4	10.4	27.8	16.7	16.7	45.1	45.8	52.8	41.0	3.33

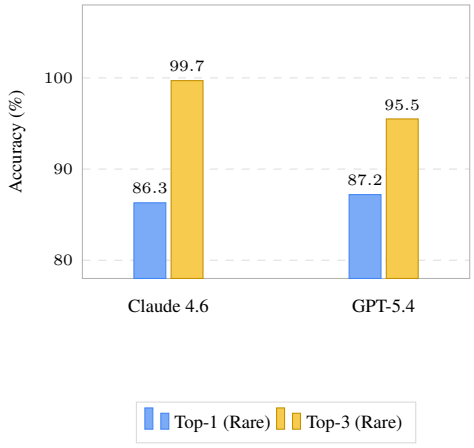

Figure 2: The inductive downgrade. Claude 4.6 deduces rare targets perfectly (99.7% Top-3) but penalizes them in final ranking (86.3% Top-1).

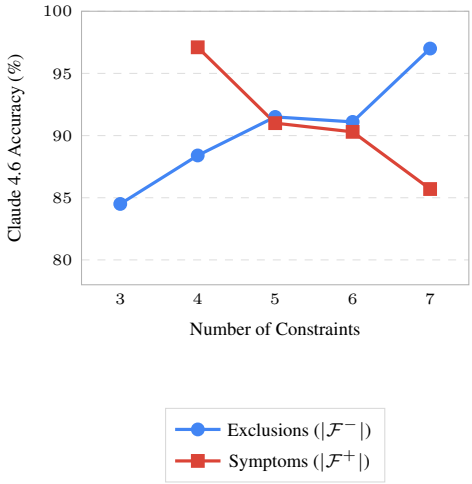

Figure 3: Deductive asymmetry. Exclusions monotonically improve accuracy (+12.5pp); positive findings act as associative noise (-11.4pp).

Across organ systems, Neurological targets proved most challenging for non-perfect models, requiring complex discrimination between six headache-presenting conditions based on laterality, papilledema, and autonomic features.

4.2 The Inductive Downgrade

We observe a stark dissociation between deductive capacity and output selection. Claude 4.6 isolates the correct rare target in its Top-3 candidates 99.7% of the time, yet ranks it first only 86.3% of the time: a 13.4-point downgrade. On common diseases, this gap shrinks to 7.2 points.

Proposition 1 (Inductive Downgrade)**.** *For a model M and class c, the downgrade magnitude is $\Delta_{ID}(M, c) = Acc_{Top\text{-}3}(M, c) - Acc_{Top\text{-}1}(M, c)$. Prevalence-dependent downgrade occurs when $\Delta_{ID}(M, rare) > \Delta_{ID}(M, common)$.*

This asymmetric signature confirms the model successfully executes constraint satisfaction internally, generating a valid candidate pool, but its output layer applies a statistical prior penalty that actively down-ranks rare targets.

4.3 Deductive Asymmetry and Error Orthogonality

If failures stem from priors overriding logic, explicitly manipulating constraint density should force logical compliance. Figure 3 tracks Claude 4.6's accuracy against constraint volume. Scaling explicit negative constraints ($|\mathcal{F}^-|$) from 3 to 7 monotonically improves accuracy from 84.5% to 97.0%. Conversely, scaling positive findings ($|\mathcal{F}^+|$) from 4 to 7 decreases accuracy from 97.1% to 85.7%. While *modus tollens* tightens the deductive bounding box and deterministically suppresses prior-favored hypotheses, *modus ponens* with additive symptoms introduces competing se-

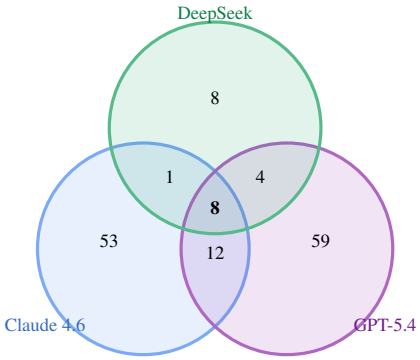

Figure 4: Error orthogonality. Of 148 unique errors across mid-tier models, 123 (83%) are strictly idiosyncratic. Only 8 overlap universally, proving failures stem from model-specific prior biases.

mantic pathways, acting as associative noise. GPT-5.4 Thinking similarly benefits from exclusions (+13.8pp), validating that negative constraints offer a functionally superior deductive signal.

Despite similar overall performance, Claude (74 errors) and GPT-5.4 (83 errors) share only 20 failures. Incorporating DeepSeek (24 errors), only 8 of 148 unique errors overlap universally (Figure 4). This profound orthogonality proves reasoning failures are rarely intrinsic to objective logical complexity; rather, they reflect idiosyncratic prior topographies. Models fail when a target resides in a probability trough within their specific pre-training manifold, dragging predictions into localized attractors.

5 Discussion and Conclusion

DiagRare exposes the architectural conflict between deductive constraint satisfaction and prevalence-driven priors. We demonstrate that mid-tier models frequently deduce correct rare targets internally but suffer an inductive downgrade at generation. Crucially, explicitly stating what a condition is not (*modus tollens*) enforces logical invariance far better than adding positive evidence. Finally, we propose $H(M) < 4.0$ bits as a diagnostic entropy threshold to detect mode collapse prior to deployment. DiagRare is designed as a living benchmark: the ontology can be extended to hundreds of diseases, expanded to natural-language vignettes, and augmented with chain-of-thought and multi-turn evaluation protocols, enabling longitudinal tracking of deductive fidelity as model architectures evolve. We release DiagRare as an open deductive benchmarking methodology.[1]

References

[1] Singhal, K., et al. (2023). Large language models encode clinical knowledge. *Nature*, 620, 172–180.

[2] Saparov, A. & He, H. (2023). Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. In *ICLR*.

[3] Nori, H., et al. (2023). Capabilities of GPT-4 on medical challenge problems. *arXiv:2303.13375*.

[4] Sahoo, S., et al. (2026). The Reasoning Trap: Logical reasoning as a pathway to situational awareness. *arXiv:2603.09200*.

[5] Nguengang Wakap, S., et al. (2020). Estimating cumulative point prevalence of rare diseases. *Eur. J. Hum. Genet.*, 28, 165–173.

[1] Code and data: `[anonymizedforreview]`.