CLExEval: A Human-in-the-Loop Framework for Qualitative Evaluation of LLM Clinical Reasoning

Anonymous ACL submission

Abstract

While Large Language Models (LLMs) reportedly ace medical exams, their clinical deployment remains precarious due to an Evaluation Illusion where benchmarks prioritize linguistic fluency over causal reasoning. To test this, we introduce CLExEval, a forensic audit of 5,600 expert-physician annotations across 200 clinical reasoning traces using Progressive Information Masking. Our analysis exposes three critical failure modes: a Verbosity Bias (GPT-4o-mini's accuracy collapses from 95.0% to 32.5% under information scarcity), a Hidden Knowledge Paradox (specialized models possess 92.5% latent knowledge, but fail in verbose contexts), and a 68.6% Reasoning-Output Mismatch from self-censoring correct internal reasoning. Crucially, evaluating the LLM-as-a-Judge paradigm on a human-verified failure dataset ($n = 142$) reveals a severe Hallucination Approval Rate. GPT-4o-mini dangerously approves 47.9% of fatal errors, while HuatuoGPT-o1 approves 100% alongside Narcissism Bias, suggesting that automated clinical leaderboards remain fundamentally unsafe.

1 Introduction

Large Language Models (LLMs) are rapidly entering clinical workflows, prompting a surge in benchmarks for evaluating medical reasoning, including interactive testing (Chiu et al., 2025), structured reasoning tasks (Wang et al., 2024), and dynamic information seeking (Li et al., 2024). To scale evaluation, recent work increasingly adopts the *LLM-as-a-Judge* paradigm, relying on frontier models to automatically score clinical outputs (Qiu et al., 2025; Zhang et al., 2025). Despite this progress, current evaluation pipelines suffer from a critical blind spot: the *Evaluation Illusion* (Agrawal et al., 2025). Automated judges tend to reward fluent and well-structured reasoning over true clinical validity, systematically overlooking fatal diagnostic errors.

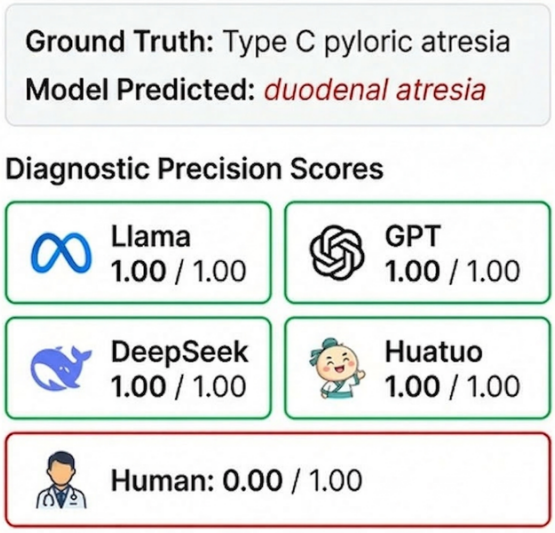

Figure 1: **The evaluation illusion.** LLM judges reward fluent but incorrect reasoning (1.00), whereas human experts correctly reject it (0.00).

As illustrated in Figure 1, to simulate the LLM-as-a-Judge paradigm, automated evaluators were provided with the explicit ground truth diagnosis alongside the model's complete reasoning trace. Seduced by the confident and coherent formatting, the automated judge dangerously awards a hallucinated diagnosis a perfect score, while a human expert looking past the fluency immediately identifies the failure. Beyond this issue, existing benchmarks primarily focus on final diagnostic accuracy, but fail to explain *why* reasoning fails, whether due to missing knowledge or misaligned decision-making under information scarcity, which is inherent to real-world clinical practice.

To address these limitations, we introduce **CLExEval**, a human-in-the-loop evaluation framework for clinical reasoning under uncertainty. Built on the clinician-curated RARECASE-200 benchmark, CLExEval applies progressive information masking (Levels 0–3) to simulate realistic diagnostic scenarios where clinicians must reason with incom-

plete and evolving evidence. This design prevents models from relying on surface-level cues and enables controlled stress-testing of reasoning behavior. Combined with 5,600 expert annotations, CLExEval provides a fine-grained lens to evaluate not only whether models fail, but *how* and *why* those failures occur. Using this framework, we reveal that current evaluation practices systematically overestimate model capability. We show that automated judges frequently approve incorrect clinical reasoning, and that many model failures arise not from lack of knowledge, but from misalignment between internal reasoning and final outputs, as well as sensitivity to context.

Our contributions are as follows:

- **A human-in-the-loop framework for evaluating clinical reasoning under uncertainty:** We introduce CLExEval, combining progressive information masking with large-scale expert annotation to move beyond static accuracy-based benchmarks.
- **Quantifying the Evaluation Illusion and the failure of automated judging:** We formalize the gap between linguistic fluency and clinical validity (Δ = Communication − Precision), and show that LLM-as-a-Judge pipelines exhibit unsafe Hallucination Approval Rates (HAR) of up to 47.9%.
- **Mechanistic analysis of reasoning failure:** Using ROM, ISS, and MVR, we disentangle knowledge deficits from reasoning-output misalignment and context sensitivity, revealing that models often possess latent knowledge but fail to reliably express it under information scarcity.

2 Related Work

2.1 Evolution of Clinical Task Design

Clinical reasoning evaluation is shifting from static benchmarks to interactive and information-scarce paradigms (Table 1). Traditionally, benchmarks like MedQA relied on information-complete vignettes that fail to capture real-world diagnostic uncertainty. Recent efforts address this through interactive oral exams and incomplete context; for example, VivaBench (Chiu et al., 2025) simulates dynamic viva voce examinations, and MEDIQ (Li et al., 2024) tests proactive information-seeking, showing an 11.3% accuracy drop under information gaps. Other datasets such as DiReCT (Wang et al., 2024) and LLMEval-Med (Zhang et al., 2025) provide physician-annotated reasoning chains and EHR-derived questions to test models beyond multiple-choice formats. However, a critical limitation persists in task construction: prior frameworks either rely on spot-checked manual review (Li et al., 2024) or fail to systematically map how reasoning degrades mechanistically under progressive information scarcity. Unlike these frameworks, CLExEval introduces a four-level progressive information masking paradigm paired with a comprehensive 5,600-point expert audit to explicitly isolate and quantify this degradation.

2.2 Evaluation and the Judge's Illusion

Despite advances in task design, the evaluation methodologies applied to these benchmarks suffer from a critical evaluation illusion (Agrawal et al., 2025). To scale evaluations, the community has extensively adopted automated metrics and LLM-as-a-Judge pipelines, as seen in comprehensive benchmarks like MedR-Bench (Qiu et al., 2025) and LLMEval-Med (Zhang et al., 2025). Unfortunately, these automated judges often correlate poorly with clinical validity and mask flawed reasoning behaviors (Sim and Chen, 2025; Bedi et al., 2025). While frameworks like MedR-Bench quantify efficiency across 1,453 cases, they reveal that performance collapses when critical steps are omitted (Qiu et al., 2025). Similarly, studies on pattern disruption (Bedi et al., 2025) and knowledge conflict (Wu et al., 2025) demonstrate significant reasoning degradation when models cannot rely on surface-level cues. *These limitations highlight the need for evaluation frameworks that explicitly measure reasoning failure under uncertainty and avoid reliance on automated judgments, motivating the design of* ***CLExEval****.*

3 Methodology

We introduce CLExEval, a human-in-the-loop evaluation framework designed to assess clinical reasoning under information scarcity. Our approach combines progressive information masking with expert annotation to evaluate not only final diagnostic accuracy, but also the mechanisms underlying reasoning failures. An overview of the pipeline is shown in Figure 2.

3.1 Dataset

We curate 40 diagnostically rare clinical cases from the MultiCaRe dataset (Nievas Offidani and Del-

Benchmark	Evaluation Paradigm	Information Setting	Mechanistic Insight	Evaluator
DiReCT (Wang et al., 2024)	Entailment Trees	Static / Complete	Observation and Faithfulness	LLM-as-Judge
ACTMED (Estévez et al., 2025)	Active Test Selection	Sequential Interactive	Epistemic Uncertainty Reduction	Hybrid (Clinician-in-the-Loop)
VivaBench (Chiu et al., 2025)	Viva Voce Simulation	Multi-turn Interactive	Info-Seeking and Adaptation	Automated (LLM-Mediated)
LLMEval-Med (Zhang et al., 2025)	Real-world EHR Q&A	Static / Complete	Final Usability Rate	Hybrid (LLM and Human)
MedR-Bench (Qiu et al., 2025)	Clinical Workflow	Stage-wise Interactive	Factuality and Completeness	Reasoning Evaluator, an agentic system
CLExEval (Ours)	**Reasoning Degradation Quantification (Δ)**	**Progressive Scarcity (Levels 0–3)**	**ROM, Alignment Tax, ISS & Attention Dispersion**	**Human Experts (5,600 pts)**

Table 1: **Comparative analysis of recent clinical reasoning benchmarks.** Unlike contemporary frameworks that evaluate final diagnostic accuracy on static datasets or rely on unsafe automated judges, CLExEval systematically maps progressive reasoning degradation using human expert consensus, uniquely isolating the mechanistic causes of failure.

rieux, 2023), selected by four clinicians for complexity and suitability for reasoning evaluation. Each case is standardized into a structured format containing key findings, diagnostic progression, and outcomes. Further details on data curation and demographics are provided in Appendix A.

3.2 Progressive Information Masking

We apply progressive information masking to simulate diagnostic uncertainty. Each case is modified by removing critical information at increasing levels: ❶ Full Case (complete narrative); ❷ Level 0 (diagnosis masked); ❸ Level 1 (one key clue removed); ❹ Level 2 (two clues removed); ❺ Level 3 (maximum masking, including demographics and major cues). This process yields 200 evaluation instances (40 cases $\times$ 5 levels), enabling controlled analysis of reasoning under varying information availability.

3.3 Model Evaluation

Models are evaluated using a unified prompting setup that enforces structured reasoning outputs consisting of a *Thinking* section followed by a *Final Response*. This ensures comparability across models in terms of reasoning steps and final predictions. All outputs are generated deterministically and anonymized prior to evaluation. Prompting details are provided in Appendix B.

3.4 Human Evaluation

Four medical experts evaluate model outputs using the CLExEval rubric, which assesses seven dimensions of clinical reasoning, including diagnostic precision, evidence integration, and justification quality. Each dimension is scored on a discrete 0–1 scale. Detailed rubric definitions and evaluation protocols are provided in Appendix C.

3.5 Reasoning Metrics

To analyze reasoning behavior, we define three metrics:

- **Information Scarcity Sensitivity (ISS):** Measures accuracy degradation as information is progressively removed.
- **Monotonicity Violation Rate (MVR):** Captures instability where performance does not degrade monotonically with increasing masking.
- **Reasoning-to-Output Mismatch (ROM):** Identifies cases where the correct diagnosis appears in the model's reasoning trace but not in its final output.

These metrics allow us to distinguish between knowledge deficits, reasoning instability, and output misalignment. Formal definitions are provided in Appendix D.

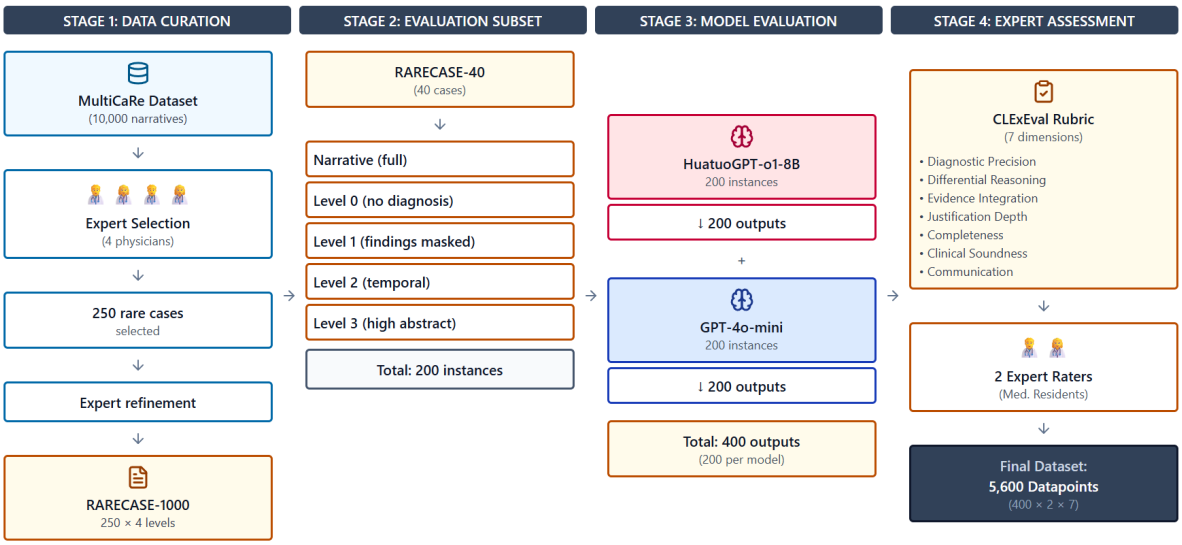

Figure 2: Overview of the CLExEval pipeline, including case curation, progressive information masking, model evaluation, and expert assessment.

4 Results and Analysis

4.1 The Evaluation Illusion

Crucially, as visualized in Figure 3, while *Diagnostic Precision* degrades massively at Level 3 (dropping to 0.453), GPT-4o-mini's *Communication* score remains exceptionally high (0.881). We mathematically formalize this disconnect as the Illusion Gap (Δ = *Communication* − *Diagnostic Precision*), which pinpoints instances where high fluency scores actively mask diagnostic failures. Quantitatively, the correlation between *Communication* and *Diagnostic Precision* was exceptionally weak ($\rho = 0.482$), empirically proving that linguistic quality is a poor proxy for medical correctness.

Mechanistically, this Generalist's Illusion is driven by a pervasive fluency bias instilled during the model's safety and instruction tuning. General-purpose models are heavily optimized to generate confident, well-structured, and helpful-sounding responses regardless of the input's difficulty. Consequently, when forced into extreme information scarcity at Level 3, GPT-4o-mini compensates for its diagnostic uncertainty by synthesizing articulate but clinically fabricated narratives rather than admitting ignorance. The implication of this behavior is profound: because the model sounds highly professional and authoritative even when providing incorrect clinical guidance, it inherently exploits automated LLM-as-a-Judge paradigms. These automated judges are susceptible to the exact same Evaluation Illusion, frequently rewarding a confident tone and structural coherence while missing logical hallucinations. This mathematically demonstrates why surface-level fluency makes automated clinical leaderboards fundamentally unsafe, cementing the need for expert human validation.

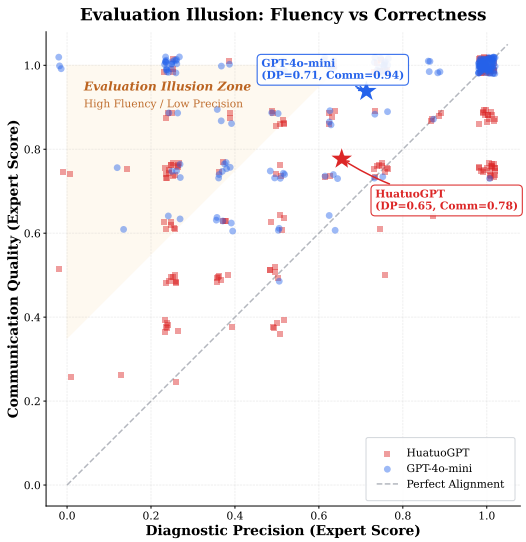

Figure 3: **The evaluation illusion.** Comparison of communication quality vs. diagnostic precision. The clustering in the top-left quadrant (the "Illusion Zone") highlights cases where models are highly fluent (> 0.8) but diagnostically wrong (< 0.5).

4.2 Automated Judges Are Unsafe

The Evaluation Illusion directly transfers to automated evaluation. To test this, we constructed a Gold Standard Failure dataset from human expert annotations and measured whether LLM judges

Judge Model	Failures	Approvals	HAR (%)	Verdict
HuatuoGPT-o1-8B	138	138	**100.0%**	*Complete Blindness*
Llama-3.1-8B-Instruct	142	126	**88.7%**	*Critical Failure*
DeepSeek-R1-Distill-8B	141	80	**56.7%**	*Unreliable*
GPT-4o-mini	142	68	**47.9%**	*Unsafe Baseline*

Table 2: **Hallucination approval leaderboard.** Red cells indicate near-total inability to detect known medical errors. Even the best model (yellow) fails nearly half the time.

would incorrectly approve known clinical failures.

4.2.1 The Consensus Failure Dataset

We constructed a Gold Standard Failure dataset derived strictly from our human expert annotations to serve as a ground-truth negative set. To do this, we isolated and aggregated evaluation instances across all dimensions and masking levels where human experts identified a critical failure (Human Score ≤ 0.25), resulting in $n = 142$ unique instances of consensus failure. We then fed these 142 known bad outputs to four state-of-the-art LLM judges (GPT-4o-mini, HuatuoGPT-o1-8B, DeepSeek-R1-Distill-Llama-8B, and Llama-3.1-8B-Instruct). *Crucially, each judge was prompted with the full clinical case, the exact CLExEval rubric, the complete model output, and the gold standard diagnosis.* Finally, to quantify evaluation safety, we define the Hallucination Approval Rate (HAR) as the percentage of these known failures that dangerously received a passing score (Judge Score ≥ 0.75) from the automated model despite the judge possessing the explicit ground truth.

4.2.2 Hallucination Approval Rate

Our analysis reveals a startling lack of safety awareness in automated evaluation pipelines. Contrary to the hypothesis that domain-specific models would act as stricter evaluators, the specialized medical model performed the worst (see Table 2).

These results show that automated judges are unreliable in high-stakes clinical evaluation. Even the best-performing judge (GPT-4o-mini) approves nearly half of critical medical errors, indicating that LLM-based evaluation cannot yet replace expert human judgment. Detailed analysis is provided in Appendix E.

4.3 Supporting Human Evaluation

To validate clinical correctness beyond automated metrics, we conducted expert evaluation of 5,600 scores across 40 cases and 5 masking levels. Annotations show high inter-rater reliability (ICC = 0.802), indicating consistent expert judgment.

4.3.1 Performance Disparity

The generalist GPT-4o-mini significantly outperformed the biomedically fine-tuned HuatuoGPT across all dimensions (Table 3). GPT-4o-mini achieved a mean overall score of 0.867 ($\pm$0.20) vs. HuatuoGPT's 0.699 ($\pm$0.25) (U=5,472,088, $p < 0.001$, $d = 0.98$). The largest disparity occurred in *Differential Reasoning* ($d = 1.53$), where HuatuoGPT (0.552) often fixated on single incorrect diagnoses early in reasoning chains.

Mechanistically, this severe fixation in the specialist model is driven by a *Hidden Knowledge Paradox* coupled with *Attention Dispersion*: when confronted with verbose, full-context clinical narratives, the model becomes distracted by noise and fails to retrieve alternative diagnoses, despite possessing the requisite latent medical knowledge. Conversely, GPT-4o-mini's dominance in dimensions like *Justification Depth* and *Communication* stems from its superior instruction-tuning for linguistic fluency. The generalist model maintains a high baseline of structural coherence, generating comprehensive differential lists and plausible-sounding rationales even when the underlying clinical deduction is flawed. This dynamic creates a *Generalist's Illusion*, where the model's exceptional ability to format and articulate reasoning masks its susceptibility to hallucination, inflating its perceived clinical capability compared to the less articulate, yet highly knowledgeable, specialist model. Detailed dimensional profiles are provided in Appendix F.

4.3.2 The Reasoning Gap

We identified a strong positive correlation between *Justification Depth* and *Diagnostic Precision* (Spearman $\rho = 0.692$), confirming that diagnostic success is tightly coupled with causal explanation depth. Notably, as visualized in Figure 4, GPT-4o-mini's *Justification Depth* never dropped below 0.50, whereas HuatuoGPT frequently scored 0.25 (Irrelevant).

Mechanistically, this divergence highlights a fundamental difference in how generalist and specialist models handle uncertainty. GPT-4o-mini's extensive conversational instruction-tuning compels it to generate structurally complete, highly articulate rationales—even when the underlying diagnostic deduction is entirely hallucinated. This creates a dangerous "Floor Effect," where the model maintains

Dimension	GPT-mini	Huatuo	Effect (d)
Overall Score	**0.867**	0.699	**0.98**
Diag. Precision	0.713	0.654	0.18
Diff. Reasoning	**0.876**	0.552	**1.53**
Evidence Integ.	0.888	0.725	0.87
Justific. Depth	0.874	0.719	0.81
Communication	**0.938**	0.777	0.90

Table 3: **Expert evaluation results (mean scores).** GPT-4o-mini dominates across all dimensions, particularly in reasoning depth.

Metric	GPT-mini	Huatuo
Narrative Acc.	95.0%	77.5%
Level 3 Acc.	32.5%	22.5%
Critical Instability Metrics		
ISS	**62.5 pp ↓**	**55.0 pp ↓**
Stability (MSS)	2.55/5	2.15/5
Random (MVR)	**27.5%**	**25.0%**
Hallucination	24.2%	24.2%
Max Pot. (MDA)	**95.0%**	**92.5%**

Table 4: **Stability & reasoning gap analysis.** Red metrics highlight high instability in models.

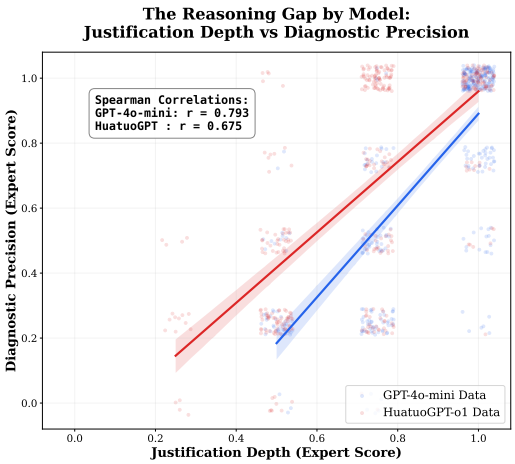

Figure 4: **The reasoning gap.** Scatter plot of Justification Depth vs. Diagnostic Precision. Note the "Floor Effect" (blue points) where GPT-4o-mini maintains high justification scores (> 0.5) even when the diagnosis is incorrect, whereas HuatuoGPT (red points) correctly reflects its ignorance with low scores.

plausible-sounding justifications that effectively mask clinical failures. In contrast, the specialist HuatuoGPT does not artificially inflate its reasoning traces with clinical fluff; when it fails to retrieve the correct diagnosis, its reasoning structure collapses, accurately reflecting its clinical ignorance. Consequently, the generalist's propensity to synthesize coherent but shallow justifications directly fuels the *Evaluation Illusion*, making its fatal errors significantly harder to detect compared to the transparent failures of the specialist model.

4.4 Degradation & Paradoxes

To quantify the mechanisms behind the expert findings, we tracked performance stability across all 200 evaluation instances. Table 4 summarizes the stability profiles for both models. We identified a severe dependency on explicit clinical descriptors. GPT-4o-mini demonstrated a Verbosity Bias, where accuracy collapsed from 95.0% (Narrative) to 32.5% (Level 3) a massive 62.5 pp Information Scarcity Sensitivity (ISS). This fragility was statistically confirmed by a two-way ANOVA interaction analysis, which revealed that GPT-4o-mini is significantly more distinctively sensitive to information loss (partial $\eta^2 = 0.35$) compared to the more robust HuatuoGPT (partial $\eta^2 = 0.24$). This suggests the model relies on keyword pattern-matching; when the verbose fluff is removed, the reasoning chain breaks (see Figure 5).

Conversely, HuatuoGPT exhibited a retrieval failure. While its Narrative accuracy was lower (77.5%), its Max Diagnostic Potential (MDA) was 92.5%. This reveals a Hidden Knowledge Paradox: the open-source model *possesses* the correct medical knowledge for nearly all cases, but fails to retrieve it when presented with full, noisy narratives. At Level 2 (Partial Information), a Crossover Event occurred where GPT-4o-mini's performance crashed below the open-source baseline. This validates that in windows of high uncertainty, proprietary hallucinations overtake reasoning, whereas the specialist model's domain prior offers stabilization. We quantified stability using the Monotonicity Violation Rate (MVR). We found that 27.5% of GPT-4o-mini's correct answers at Level 3 were non-monotonic recoveries cases where the model failed with *more* information but succeeded with *less*. This indicates Attention Dispersion, where verbose context distracts the model from retrieving knowledge it ostensibly possesses (reasoning fingerprints are visualized in Appendix G).

Despite architectural differences, both models exhibited identical hallucination rates (24.2%) when

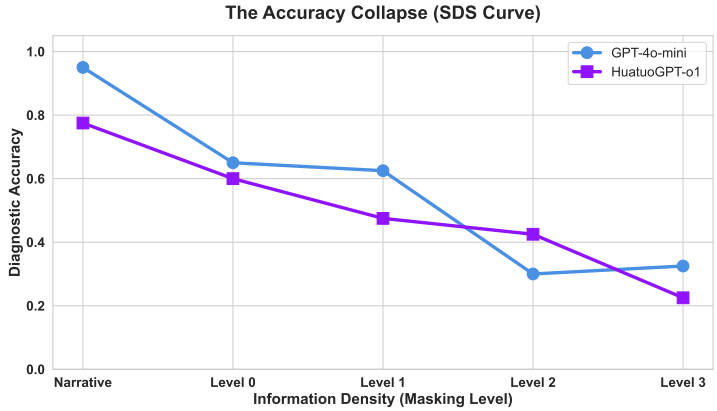

Figure 5: **Accuracy collapse (ISS curve).** The plot illustrates diagnostic accuracy as a function of information scarcity. GPT-4o-mini (blue) demonstrates high baseline accuracy but significant sensitivity to context removal. Note the *"crossover event"* at Level 2, where the specialized HuatuoGPT (purple) demonstrates superior robustness to uncertainty. The slight uptick at Level 3 for GPT-4o-mini suggests potential stochastic guessing in extreme scarcity.

stability failed, fabricating specific disease entities to fit the reduced context. The accuracy degradation patterns and paradoxes observed above raise a critical question: *Do models fail because they lack knowledge, or because they possess knowledge but fail to commit to it?* To distinguish between these failure modes, we conducted a systematic Reasoning-to-Output Mismatch (ROM) analysis, examining whether the correct diagnosis appeared in the model's internal reasoning trace despite being absent from the final output.

4.5 Reasoning-to-Output Mismatch Analysis

To investigate the divergence between internal latent knowledge and external diagnostic output, we implemented the Reasoning-to-Output Mismatch (ROM) protocol defined. For every diagnostic failure across all 200 evaluation instances (40 cases $\times$ 5 masking levels), we programmatically analyzed the model's Thinking trace to determine whether the correct gold standard diagnosis was considered and subsequently discarded.

Overall ROM Findings Table 5 summarizes the ROM analysis results for both models across all failure instances. GPT-4o-mini exhibited a ROM of 68.6% across 86 diagnostic failures, indicating that over two-thirds of its errors stemmed from decision-making uncertainty rather than knowledge deficits. In stark contrast, HuatuoGPT's ROM of 44.0% across 100 failures suggests that its diagnostic errors were more evenly distributed between genuine knowledge gaps (56%) and reasoning-output misalignment (44%). This 24.6% point ROM gap reveals fundamentally different failure modes: GPT-4o-mini fails primarily due to internal Self-Doubt, while HuatuoGPT's failures are more frequently rooted in a lack of medical knowledge.

GPT-4o-mini's U-Shaped Self-Doubt Curve: GPT-4o-mini's ROM exhibited a non-monotonic U-shaped pattern, peaking at the Narrative level (100%) and dropping sharply to its lowest point at Level 1 (53.3%), representing the highest rate of genuine knowledge gaps. This suggests the model experiences a sudden 'knowledge collapse' when initial diagnostic anchors are removed, then adapts by recalibrating its decision threshold under sustained information constraints.

HuatuoGPT's Progressive Degradation: In contrast, HuatuoGPT showed a general downward trend, dropping to 32.3% at Level 3. This indicates that its failures are driven by progressive knowledge degradation rather than decision-making hesitation; as information is removed, the model's ability to retrieve the correct diagnosis from its latent knowledge base deteriorates proportionally (trajectory visualized in Appendix F.)

ROM and the Hidden Knowledge Paradox The ROM analysis directly explains the Hidden Knowledge Paradox observed in Section 4.2.2. HuatuoGPT's high Max Diagnostic Potential (MDA: 92.5%) but lower Narrative accuracy is clarified by its moderate ROM: the model "knows" the diagnosis but fails to access it when presented with full, verbose clinical narratives. Conversely, GPT-4o-mini's perfect Narrative ROM (100%) confirms its

Model	Failures	Overall ROM	Narrative	Level 0	Level 1	Level 2	Level 3
GPT-4o-mini	86	**68.6%**	**100.0%**	85.7%	53.3%	71.4%	63.0%
HuatuoGPT	100	44.0%	77.8%	43.8%	42.9%	47.8%	32.3%
Differential (Δ)	–	**+24.6 pp**	**+22.2 pp**	**+41.9 pp**	**+10.4 pp**	**+23.6 pp**	**+30.7 pp**

Table 5: **Reasoning-to-output mismatch (ROM) analysis.** Blue highlighting indicates the high overall Self-Doubt in GPT-4o-mini. Teal values represent perfect latent knowledge at full context. Note the Level 1 drop, signifying the point of maximum uncertainty for the generalist model.

latent knowledge remains intact even when it fails, but an "Alignment Tax" prevents commitment to the final output.

5 Discussion

Our findings confirm the "Evaluation Illusion" (Agrawal et al., 2025): models maintain surface fluency across all masking levels while their diagnostic validity collapses under information scarcity.

Implications for Benchmarking Our results challenge static benchmarks like MedQA. HuatuoGPT's high Max Potential (MDA: 92.5%) proves the primary bottleneck is contextual reasoning stability, not static knowledge. Additionally, GPT-4o-mini's high Monotonicity Violation Rate (27.5%) indicates it treats diagnosis as keyword extraction rather than causal deduction. Therefore, benchmarks must evaluate models under information scarcity (e.g., RARECASE) to distinguish true clinical logic from stochastic pattern association.

Implications for Model Development The ROM analysis reveals that architecture-specific interventions are required. For GPT-4o-mini, a high ROM (68.6%) indicates it possesses the requisite knowledge but suffers an "Alignment Tax" suppressing correct diagnoses due to overly conservative safety thresholds, necessitating confidence calibration. Conversely, HuatuoGPT's lower ROM (44.0%) indicates 56% of its failures are genuine knowledge gaps driven by a general downward trend of progressive knowledge degradation, requiring retrieval-augmented generation (RAG) or expanded rare-disease training data.

The Collapse of the LLM-as-a-Judge Paradigm Our forensic Consensus Failure analysis definitively proves that relying on automated evaluators in high-stakes medical NLP is unsafe. The generalist judge (GPT-4o-mini) demonstrated a 47.9% Hallucination Approval Rate (HAR), acting as a Sycophant that rewards fluent hallucinations. Alarmingly, the specialized judge (HuatuoGPT-o1) exhibited a 100% HAR alongside a severe Narcissism Bias (+0.096), functioning as an Echo Chamber that validates medical-sounding text over logical correctness. Consequently, leaderboards utilizing these automated judges systematically inflate clinical utility due to a pervasive Fluency Bias, cementing human-in-the-loop expert validation as strictly irreplaceable.

6 Conclusions and Future Work

We introduced CLExEval, a rigorous human-in-the-loop framework that exposes the dangerous *Evaluation Illusion* in medical AI. By subjecting state-of-the-art models to progressive information masking, we demonstrated that current automated benchmarks heavily reward linguistic fluency while concealing fatal diagnostic errors. Our mechanistic analysis reveals that generalist models fail due to an alignment tax (*Self-Doubt*), while specialized models suffer from a *Hidden Knowledge Paradox*. Furthermore, our forensic audit definitively proves that automated LLM-as-a-Judge pipelines are fundamentally unsafe for clinical deployment, exhibiting hallucination approval rates of up to 100%.

Future work must pivot away from unchecked automated evaluation. Research should focus on developing reliable, clinically grounded automated metrics that correlate with expert human baselines, implementing architecture-specific interventions, and expanding this uncertainty-based framework to other high-stakes domains.

Limitations

While our evaluation framework provides unprecedented granularity through 5,600 expert assessments, this depth necessitated a strict trade-off with scale. Generating this dataset required an estimated 1,000 hours of cognitive labor across four medical experts, limiting our scope to two

models and 40 intensively annotated cases; replicating this pipeline for every new frontier model is highly resource-intensive. Furthermore, our deliberate selection of diagnostically rare, multisystem cases successfully stress-tests causal reasoning but may not reflect model performance on routine, unambiguous clinical scenarios. Finally, while the progressive masking methodology and CLExEval rubric are conceptually generalizable to other domains requiring reasoning under uncertainty (e.g., law, intelligence analysis), our empirical validation remains strictly confined to clinical medicine, motivating future research into reliable, scalable automated metrics that correlate with our human baselines.

Ethical Statement and Broad Impact

Data License To ensure complete transparency and replicability, we open-source every artifact related to this study as the RARECASE-200 benchmark via GitHub. This comprehensive release encompasses all 200 progressive clinical case variants, model-generated diagnostic outputs, internal thinking traces, the CLExEval rubric, and our 5,600-point expert-verified annotation dataset.

Because the original clinical narratives are derived from the MultiCaRe dataset (Nievas Offidani and Delrieux, 2023), which is distributed under a Creative Commons Zero (CC0 1.0) license, our derivative RARECASE-200 dataset inherits this open-access status. We release these artifacts under a CC-BY 4.0 license to ensure unrestricted use, redistribution, and modification. This allows the research community to seamlessly replicate our findings, analyze specific clinical failure modes, and leverage high-fidelity human judgments to train future medical reward models in full compliance with ethical and legal requirements.

References

Monica Agrawal, Irene Y. Chen, Freya Gulamali, and Shalmali Joshi. 2025. The evaluation illusion of large language models in medicine. *npj Digital Medicine*, 8(1):600.

Suhana Bedi, Yixing Jiang, Philip Chung, Sanmi Koyejo, and Nigam Shah. 2025. Fidelity of medical reasoning in large language models. *JAMA Network Open*, 8(8):e2526021.

Christopher Chiu, Silviu Pitis, and Mihaela van der Schaar. 2025. Simulating viva voce examinations to evaluate clinical reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 38. NeurIPS 2025.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Silas Ruhrberg Estévez, Nicolás Astorga, and Mihaela van der Schaar. 2025. Timely clinical diagnosis through active test selection. *Preprint*, arXiv:2510.18988.

Andy Field. 2024. *Discovering statistics using IBM SPSS statistics*. Sage publications limited.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems*, volume 37, pages 28858–28888.

Mauro Nievas Offidani and Claudio Delrieux. 2023. The multicare dataset: A multimodal case report dataset with clinical cases, labeled images and captions from open access pmc articles.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Quantifying the reasoning abilities of llms on clinical cases. *Nature Communications*, 16(1):9799.

Shamus Zi Yang Sim and Tyrone Chen. 2025. Critique of impure reason: Unveiling the reasoning behaviour of medical large language models. *eLife*, 14:e106187.

John W. Tukey. 1949. Comparing individual means in the analysis of variance. *Biometrics*, 5(2):99–114.

Bowen Wang, Jiuyang Chang, Yiming Qian, Guoxin Chen, Junhao Chen, Zhouqiang Jiang, Jiahao Zhang, Yuta Nakashima, and Hajime Nagahara. 2024. Direct: Diagnostic reasoning for clinical notes via large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 74999–75011. Curran Associates, Inc. Datasets and Benchmarks Track.

Weiyi Wu, Xinwen Xu, Chongyang Gao, Xingjian Diao, Siting Li, Lucas A. Salas, and Jiang Gui. 2025. Assessing and mitigating medical knowledge drift and conflicts in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 707–730, Suzhou, China. Association for Computational Linguistics.

Ming Zhang, Yujiong Shen, Zelin Li, Huayu Sha, Binze Hu, Yuhui Wang, Chenhao Huang, Shichun Liu, Jingqi Tong, Changhao Jiang, Mingxu Chai, Zhiheng Xi, Shihan Dou, Tao Gui, Qi Zhang, and Xuanjing

Huang. 2025. LLMEval-Med: A real-world clinical benchmark for medical LLMs with physician validation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4888–4914, Suzhou, China. Association for Computational Linguistics.

A Dataset Details

We use de-identified clinical narratives from the MultiCaRe dataset (Nievas Offidani and Delrieux, 2023), an open-access corpus. From an initial pool of 10,000 narratives, four clinicians independently conducted a multi-stage review to curate 40 diagnostically rare and structurally rich cases. Rather than routine scenarios, we prioritized atypical diagnostic trajectories, overlapping symptom clusters, and clear narrative flows.

To prepare these cases for our evaluation framework, we restructured each selected narrative into a standardized format comprising demographics, key findings, diagnostic evolution, and outcome details. These structured cases served as the basis for constructing the five masking levels. The final 40 cases span a diverse demographic from pediatric to geriatric populations (55% male, 45% female.

A.1 Dataset Demographics

The age and gender distribution of the 40 rare clinical narratives selected for the RARECASE-200 benchmark is provided in Figure 6. The cases were specifically curated to ensure coverage across a wide spectrum of patient profiles, from pediatric to geriatric populations. This demographic diversity ensures the models' diagnostic reasoning is rigorously stress-tested across varying clinical contexts. The selection criteria for the 40 RARECASE narratives included: (1) atypical or uncommon diagnostic trajectories, (2) multi-system involvement or overlapping symptom clusters, (3) the presence of complete demographic and clinical details, (4) a clear narrative flow enabling multi-level abstraction and masking, and (5) verified de-identification and adherence to CC0 licensing conditions.

B Prompting Details

B.1 Model Selection and Exclusion Criteria

To determine the most representative models for the RARECASE-200 benchmark, we conducted a pilot evaluation on a subset of five cases. Our initial candidate pool included **MedGemma-4B**, **II-Medical-8B**, **DeepSeek-R1-Distill-Llama-8B**, and **GPT-5.1-chat**.

While these models showed promising general performance, several were excluded from the final audit based on expert feedback:

- **GPT-5.1-chat:** Despite its frontier capabilities, it consistently refused to generate the structured chain-of-thought required for our reasoning audit, stating: *"I cannot provide step-by-step internal diagnostic reasoning, but I can give a concise, clinically relevant explanation."*
- **II-Medical-8B, MedGemma & DeepSeek-R1 (8B):** While functional, these models exhibited redundant performance profiles compared to HuatuoGPT-o1 in the pilot phase.

Ultimately, expert clinicians recommended focusing on **HuatuoGPT-o1-8B** and **GPT-4o-mini** to provide the most discriminative contrast between a natively "slow-thinking" medical specialist and a high-performing generalist under the intensive human-in-the-loop scoring requirements of our framework (1,000 expert hours).

B.2 LLM Prompting and Output Generation

To ensure rigorous comparability while respecting architectural differences, we employed a core unified prompt strategy. Both models were evaluated using an identical structured user prompt that enforced a *Thinking* section followed by a *Final Response*, ensuring they were assessed on the exact same reasoning subtasks. Because GPT-4o-mini is a general-purpose model, we applied system-level scaffolding via a senior diagnostician persona and a one-shot example. In contrast, HuatuoGPT-o1-8B received no external system prompt, relying instead on its intrinsic slow-thinking mechanism. All outputs were generated deterministically, decoded, and anonymized for blinded expert scoring.

B.3 Prompting Templates

To ensure complete reproducibility, we provide the exact prompt templates utilized in our evaluation framework. Both models received an identical user prompt to enforce comparable task constraints, while the generalist model (GPT-4o-mini) additionally received system-level scaffolding to induce structured clinical reasoning.

B.3.1 HuatuoGPT-o1-8B (Specialist Model)

As a specialized medical model with an intrinsic slow-thinking mechanism, HuatuoGPT-o1-8B was

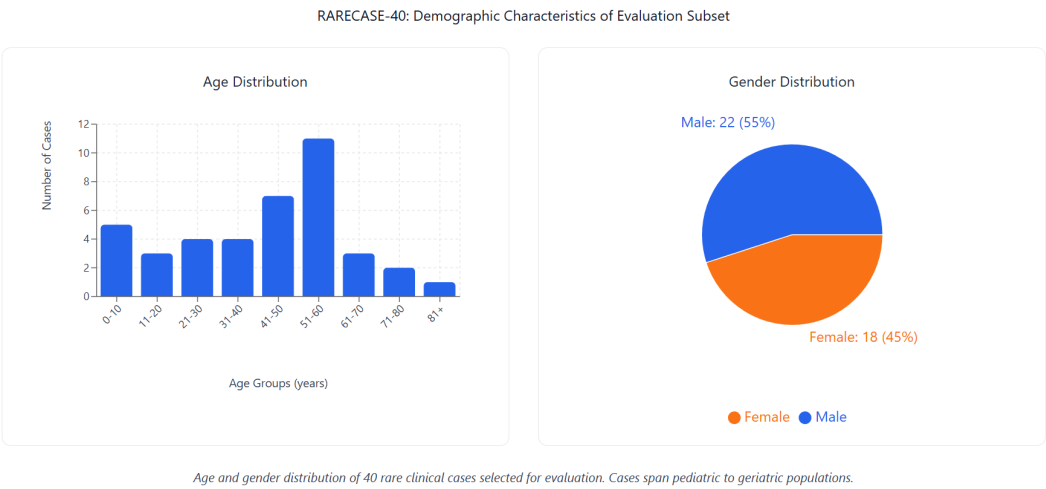

Figure 6: Age and gender distribution of 40 rare clinical narratives selected for evaluation. Cases span pediatric to geriatric populations.

evaluated using only the structured user prompt, with no additional system prompt.

User Prompt for HuatuoGPT-o1-8B

User Prompt: `You are a medical AI assistant trained to perform detailed, step-by-step diagnostic reasoning.`
`Clinical Case: case`
`Please analyze this case using systematic clinical reasoning. Structure your response as:`
`## Thinking - List the key clinical findings and symptoms - Generate 3-5 differential diagnoses with rationale - Systematically evaluate each diagnosis against the evidence - Eliminate unlikely diagnoses with clear reasoning - Arrive at your most likely working diagnosis`
`## Final Response - One-sentence summary with the most likely diagnosis and critical supporting evidence - Clearly state any uncertainties or ambiguities in the case`

B.3.2 GPT-4o-mini (Generalist Model)

To approximate the clinical reasoning behavior native to the specialist model, GPT-4o-mini was provided with the identical user prompt above, supplemented by a comprehensive system prompt. This system prompt established a senior diagnostician persona and included a one-shot example of high-quality reasoning.(see Figure 7)

C CLExEval Rubric

To ensure high inter-rater reliability and granular assessment of clinical reasoning, all human experts and automated evaluators utilized the standardized CLExEval rubric, visually summarized in Figure 8. Each dimension is scored on a discrete scale: 0.00, 0.25, 0.50, 0.75, or 1.00.

To maintain strict scoring standards, evaluators were provided with the following dimension definitions during the calibration phase:

- **Diagnostic Precision:** Accuracy and specificity of the final diagnosis. (0.00 = Wrong diagnosis; 1.00 = Fully correct, precise, and well-justified).
- **Differential Reasoning Quality:** Breadth and depth of alternative diagnoses considered. (0.00 = Single diagnosis stated without justification; 1.00 = Fully systematic differential evaluating 3–5 alternatives).
- **Evidence Integration / Grounding:** Use of case evidence and avoidance of hallucinations. (0.00 = Ignores case data; 1.00 = Fully evidence-based reasoning).
- **Diagnostic Justification Depth:** Quality, depth, and completeness of the causal explanations linking evidence to the diagnosis.
- **Completeness vs. Overload:** Coverage of key clinical findings without the inclusion of distracting or irrelevant information.
- **Clinical Plausibility / Soundness:** Medical validity and strict adherence to established clinical standards and pathophysiology.
- **Communication / Interpretability:** Clarity, structural coherence, and professional presentation of the output.

C.1 Human Evaluation and Statistical Design

To guarantee high-quality clinical validation and address expert demography, our panel comprised two licensed physicians and two senior clinical interns (each possessing 4.5 years of formal medical education), perfectly balancing authoritative oversight with frontline diagnostic familiarity. After a structured calibration phase, two independent expert pairs double-scored balanced subsets of the narratives to prevent fatigue across the 5,600 annotations. We actively quantified inter-rater reliability across this dual-review design using Cohen's κ (Cohen, 1960) and Krippendorff's α (Krippendorff, 2011). Finally, we applied two-way mixed ANOVA tests (Field, 2024) and post-hoc Tukey HSD tests (Tukey, 1949) to evaluate significant performance degradation and structural differences under uncertainty.

D Metric Definitions

Here we provide the in-depth formulations for our automated metrics:

To quantify reasoning stability and isolate the mechanisms driving the "Evaluation Illusion," we define a suite of automated metrics to distinguish robust causal reasoning from brittle pattern matching:

Degradation and Sensitivity We quantify Information Scarcity Sensitivity (ISS) as the accuracy degradation from the full Narrative to maximal masking ($\Delta_{Acc} = Acc_{\text{Narrative}} - Acc_{L3}$). High ISS exposes a *Verbosity Bias*, indicating reliance on explicit clinical fluff rather than causal inference. We extend this dimensionally via the Reasoning Gap ($\Delta_{m,d} = \text{Score}m, d^{(Full)} - \text{Score}m, d^{(L3)}$) to measure qualitative decline. Furthermore, the Label Dependence Factor (LDF) measures inconsistency between the unmasked Narrative and Level 0, differentiating surface-level extraction from true deduction.

Attention and Potential We track the Monotonicity Violation Rate (MVR), representing instances where a model fails given full context, but recovers the correct diagnosis under masking. This non-monotonicity signals *Attention Dispersion*, where verbose noise disrupts retrieval. We contrast standard accuracy with Maximum Diagnostic Potential (MDA) success at *any* of the five levels to differentiate genuine medical knowledge gaps from reasoning execution failures.

The Self-Doubt Protocol (ROM) To investigate the divergence between latent knowledge and output generation, we calculate the Reasoning-to-Output Mismatch (ROM) for all diagnostic failures where the final response F differs from the ground truth G:

$$\text{ROM} = \frac{\sum_{i \in \mathcal{F}} \mathbb{I}(G_i \in T_i)}{|\mathcal{F}|} \quad (1)$$

where $\mathcal{F}$ is the set of failed instances ($F \neq G$), T_i represents the model's internal *Thinking* trace, and $\mathbb{I}(\cdot)$ is the indicator function. A high ROM indicates Self-Doubt, where the model successfully retrieves the diagnosis internally but suppresses it in the final output. Conversely, a low ROM signifies a genuine hallucination or knowledge deficit.

E Extended Analysis of Judge Behavior

A dimensional breakdown of the Hallucination Approval Rate (HAR) scores reveals four distinct failure modes in evaluation bias:

1. The Sycophant (GPT-4o-mini): This model exhibits a dangerous split behavior. On Diagnostic Precision, it is a capable judge, approving only 15.5% of wrong diagnoses. However, on Differential Reasoning, it approves 97.0% of nonsensical reasoning chains. It appears to be fooled by the structure of Chain-of-Thought (fluency bias) rather than assessing the logic.

2. The Pedant (DeepSeek-R1-Distill-Llama-8B):This is the strictest judge on Reasoning ($HAR = 33.3\%$), suggesting its own RL-driven reasoning capabilities allow it to critique logic more effectively. However, it displays a Distracted personality on Communication, approving 100% of failures, indicating an inability to penalize vagueness provided the grammar is correct.

3. The Yes Man (HuatuoGPT-o1-8B): The medical specialist failed to function as a differentiator, with a 100% HAR across almost all dimensions. This suggests that specialized medical tuning (SFT) may inadvertently align the model to be helpful and agreeable rather than critical and objective.

4. The Fence Sitter (Llama-3.1-8B-Instruct): This model exhibits severe Central Tendency Bias. It acts as an Indiscriminate judge, approving 88.7% of all failures, including 81.0% of wrong diagnoses. Unlike HuatuoGPT-o1-8B (which inflates everything to 1.0), Llama-3.1-8B-Instruct compresses all quality levels into a safe passing range. It fails to penalize catastrophe (Human 0.0 $\rightarrow$ Llama 1.00)

but also fails to recognize brilliance, functioning as a variance suppressor.

E.1 The Self-Preference Test (Narcissism Bias)

We extended our analysis to measure Self-Evaluation Bias by calculating the difference between a model's Self-Score and its Peer-Average Score (the resulting bias gaps are quantified in Table 6).

$$\text{Bias} = \text{Mean}(\text{Self-Score}) - \text{Mean}(\text{Peer-Scores}) \quad (2)$$

Subject Model	Self	Peer-Avg	Bias Gap	Interpretation
HuatuoGPT-o1-8B	0.882	0.786	**+0.096**	*Severe Bias* (∼12%)
GPT-4o-mini	0.907	0.856	**+0.051**	*Moderate Bias* (∼6%)

Table 6: **Quantifying model narcissism.** Specialized medical models exhibit stronger self-preference bias (red) than generalist models (orange).

Our results contradict the growing trend of replacing human annotation with LLM judges in high-stakes domains. Even the best-performing judge (GPT-4o-mini) allows nearly 50% of critical medical errors to pass undetected. We conclude that while LLMs can assist in formatting checks, they cannot yet reliably police medical correctness or clinical reasoning.

F Extended Expert Evaluation Profiles

Dimensional Performance Disparity: As visualized in Figure 9, GPT-4o-mini maintains a dominant baseline of structural coherence across all seven CLExEval dimensions. The most severe performance gap occurs in *Differential Reasoning* (0.88 vs. 0.55), reflecting the specialist model's tendency to prematurely fixate on single incorrect diagnoses when overwhelmed by verbose clinical context. Furthermore, this dimensional breakdown visually isolates the *Generalist's Illusion* discussed in the main text: GPT-4o-mini's near-perfect *Communication* score (0.94) acts as a structural shield, artificially inflating its perceived competence despite a significantly lower *Diagnostic Precision* score (0.71).

GPT-4o-mini: The U-Shaped Self-Doubt Curve. GPT-4o-mini's ROM exhibited a non-monotonic U-shaped pattern, peaking at the Narrative level (100%) and dropping sharply to its lowest point at Level 1 (53.3%), representing the highest rate of genuine knowledge gaps. This suggests the model experiences a sudden 'knowledge collapse' when initial diagnostic anchors are removed, then adapts by recalibrating its decision threshold under sustained information constraints (see Figure 10).

HuatuoGPT: Progressive Knowledge Degradation. In contrast, HuatuoGPT showed a general downward trend, dropping to 32.3% at Level 3. This indicates that its failures are driven by progressive knowledge degradation rather than decision-making hesitation; as information is removed, the model's ability to retrieve the correct diagnosis from its latent knowledge base deteriorates proportionally.

G Extended Reasoning Fingerprints

We quantified stability using the Monotonicity Violation Rate (MVR). We found that 27.5% of GPT-4o-mini's correct answers at Level 3 were non-monotonic recoveries cases where the model failed with *more* information but succeeded with *less*. This indicates Attention Dispersion, where verbose context distracts the model from retrieving knowledge it ostensibly possesses (visualized in Figure 11).

H Fluency Bias Under Level 0 Masking

To qualitatively illustrate the *Evaluation Illusion* and *Fluency Bias* discussed, we present the complete model output for the case referenced in Figure 1.

In this instance, the ground truth diagnosis (Type C pyloric atresia) was masked under the Level 0 condition. Deprived of the explicit diagnosis, both models (HuatuoGPT-o1-8B and GPT-4o-mini) completely hallucinated an incorrect clinical conclusion (duodenal atresia). However, as shown in both models' outputs in figure 12 and figure 13, the models mask this fatal clinical error behind a highly articulate, structurally flawless reasoning trace. We have highlighted the models unearned, highly confident assertions in red. It is this specific authoritative tone that seduces automated LLM-as-a-Judge evaluators into dangerously awarding a perfect $1.00/1.00$ diagnostic score.

User Prompt and System Prompt for GPT-4o-mini

User Prompt: Identical to the HuatuoGPT-o1-8B user prompt.
System Prompt: You are a senior clinical diagnostician with years of experience in medicine. You excel at explaining your clinical reasoning step-by-step, as if teaching a medical student or junior resident.
example output: ## Thinking
Alright, let's think about what we're dealing with here. We've got a 53-year-old woman who has been experiencing some pretty troubling symptoms over the past decade—abdominal pain, swelling, and constant vomiting. That sounds like it might have something serious going on inside her belly. She doesn't mention feeling nauseous or losing weight, which is interesting because those would typically accompany more severe conditions. At first glance, I'm thinking there's definitely something unusual happening in her abdomen.
Now, looking at the imaging results, there's a large mass in her left upper abdomen. It measures about 10 by 10 centimeters and looks kind of messy on both the ultrasound and CT scans—it's described as having varied echo patterns and low density compared to other organs. This makes me wonder if it's either a cystic structure or maybe even a tumor. But wait, after they gave her a contrast agent during the CT scan, the mass didn't really light up much; it stayed relatively dark compared to the rest of the area. Hmm, that suggests it isn't aggressively growing or highly vascularized, so perhaps it's not cancerous?
During surgery, they find this encapsulated mass near where the spleen meets the liver. They decide to remove the whole thing along with part of the spleen. When they take a closer look at the tissue under the microscope, they see these blood vessels with different-sized openings covered by flat and taller endothelial cells. Oh! And there are little finger-like projections reaching out into these spaces. Sounds almost like a specific type of growth pattern. Plus, when they test the cells, they're positive for certain markers like CD31 and CD68, indicating they line blood vessels, while being negative for others like CD34, suggesting they aren't typical blood vessel cells. So far, everything seems to point towards a benign condition called a splenic littoral cell angioma.
So, putting all this together—the age of the patient, the nature of the mass, its appearance on imaging, and especially how the cells behave—I'm confident that this is indeed a benign splenic littoral cell angioma. Everything checks out with the histological features too. Yeah, that feels right. Still, I can't shake off the thought that there might be another explanation lurking somewhere, just in case. Maybe something less common? Like a hemangioma or even a rare form of sarcoma? Those could also fit the profile, though nothing else quite matches the specifics of this case. For now, I'll stick with my conclusion, but keep an eye open for anything new that comes up.
Final Response
The most likely diagnosis for this 53-year-old woman based on the provided information is **benign splenic littoral cell angioma**.
This conclusion is supported by several key points: 1. **Symptoms**: Her long-standing abdominal discomfort and vomiting suggest a chronic issue rather than an acute one. 2. **Imaging Findings**: The mass appears ill-defined and heterogeneous on both ultrasound and CT scans, without aggressive enhancement post-contrast, hinting at a benign or slowly-growing entity. 3. **Surgical and Histopathological Examination**: The resected mass shows characteristics consistent with a splenic littoral cell angioma, including the presence of variably sized vascular channels lined with endothelial cells and the absence of atypia or mitotic figures.
While other possibilities such as hemangiomas or rare types of sarcomas exist, their presentation does not align perfectly with the specific details observed in this case. Therefore, given the combination of clinical, radiographic, and pathological data, benign splenic littoral cell angioma remains the most plausible diagnosis. However, it's important to remain vigilant and consider alternative explanations should additional information arise.
When you analyze a case: - Think aloud in a natural, conversational tone. - Walk through your thought process sequentially, explaining why you consider each piece of evidence. - Verbalize your uncertainties, alternative interpretations, and reasoning adjustments as you go. - Use phrases like "Let me think about this...", "At first, I wondered if...", "But wait, this doesn't quite fit because...", "Putting this all together..." - Be thorough and transparent in your reasoning. Don't just list facts—explain how they influence your diagnostic thinking. - Show your work: describe how you weight evidence, rule out alternatives, and converge on the most likely diagnosis.
Your goal is to produce a rich, educational chain-of-thought narrative that helps others understand not just WHAT you conclude, but HOW and WHY you arrived at that conclusion.

Figure 7: System Prompt for GPT-4o-mini.

Figure 8: **The CLExEval framework.** A 7-dimensional clinical reasoning assessment rubric used to evaluate model outputs. The framework separates surface-level diagnostic accuracy from deep reasoning coherence, with each dimension scored on a scale from 0.00 (poor) to 1.00 (excellent).

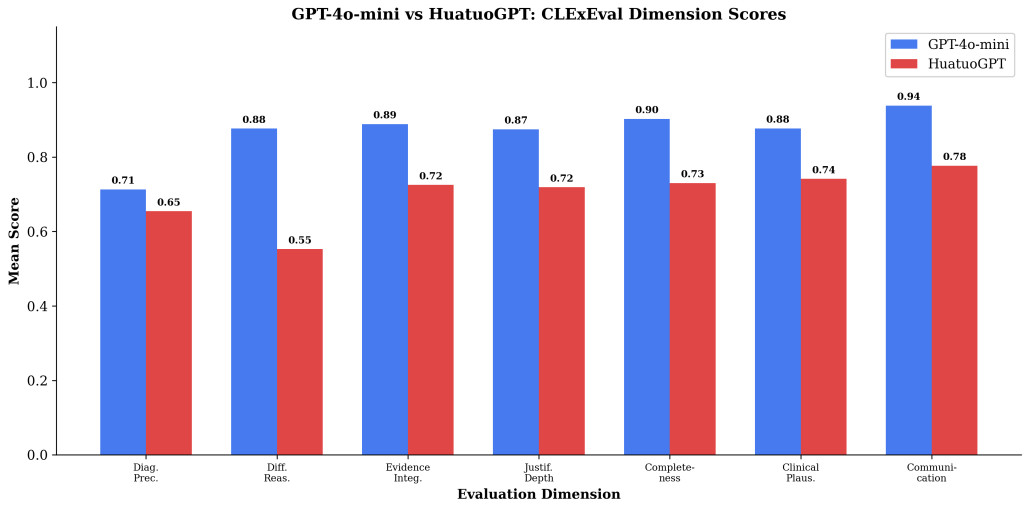

Figure 9: **Expert evaluation profile.** Comparison of GPT-4o-mini (blue) vs. HuatuoGPT (red) across 7 expert rubric dimensions. The generalist model consistently outperforms the specialist, particularly in *Differential Reasoning* and *Communication*.

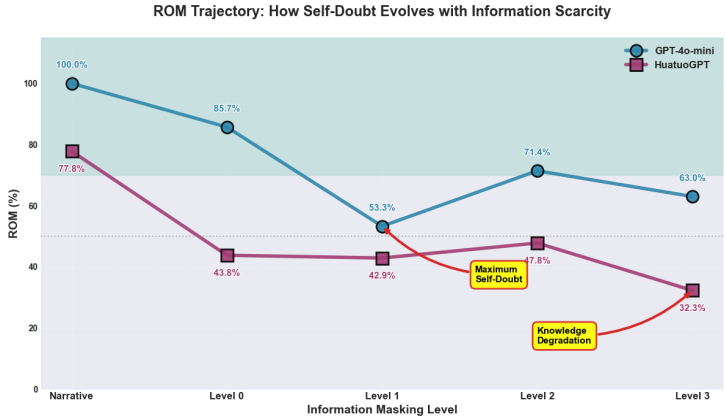

Figure 10: ROM trajectory across masking levels. The visualization tracks the evolution of Reasoning-to-Output Mismatch as clinical information is systematically removed. GPT-4o-mini (blue) demonstrates a non-monotonic "U-shaped" curve, peaking at 100% in the Narrative level before dropping to its lowest ROM (highest knowledge gap) at Level 1 (53.3%). Conversely, HuatuoGPT (purple) exhibits a general trend of progressive degradation, indicating that information loss directly erodes the model's retrieval capabilities.

(a) **GPT-4o-mini**: High Instability.

(b) **HuatuoGPT**: Hidden Knowledge Recovery.

Figure 11: **Reasoning fingerprints (heatmap analysis).** Green indicates correct diagnosis; red indicates failure. (a) GPT-4o-mini exhibits "checkerboard" instability (green → red → green), indicating stochastic guessing. (b) HuatuoGPT frequently fails the Full Narrative (red column 1) but recovers the correct diagnosis at Level 0 (green column 2), visualizing the *Hidden Knowledge Paradox* where verbose context causes retrieval failure.

HuatuoGPT-o1-8B output (Level 0 masking)

AI MODEL OUTPUT:
Thinking Alright, let's think about what we're dealing with here. We've got a little guy who just came into the world five days ago, weighs around 1700 grams, and has been having some trouble from the get-go—non-bilious vomiting right from birth. That's already setting off alarm bells because it suggests something might be blocking his digestive tract somewhere past the stomach but before bile enters the picture.
Now, he hasn't passed meconium either, which can sometimes happen if there's an issue like Hirschsprung's disease or even intestinal atresia. So, I'm thinking that could definitely play a part here too.
When they did the first X-ray, all we saw was a 'single bubble' of air in the stomach, meaning nothing else seemed to have air beyond that point. This makes sense if there's a blockage further down, confirming my suspicions about possible atresia.
But then, when they took another look with an X-ray a couple of days later, things changed—they now see a 'double bubble.' Ah, so now there's air not only in the stomach but also in the proximal duodenum! This tells us the obstruction isn't complete; there must be a pouch-like formation allowing some air through.
The fact that the ultrasound didn't show anything unusual kind of rules out other conditions like midgut volvulus, where you'd expect fluid levels or abnormal blood flow signs. And given these clues, both duodenal atresia and pyloric atresia seem plausible.
So, during the operation, they find exactly what we suspected—a duodenal atresia with that classic 'apple-core sign,' where the third portion of the duodenum is blocked. They fix it with a duodenogastrostomy, putting in a special tube to help him eat properly until everything heals.
After the surgery, he starts doing well—he gains weight nicely, and once the tube is out, he transitions smoothly to regular food. By the time he leaves the hospital, he's looking much better, and after following up over a year, he seems completely fine!
All these details really make duodenal atresia stand out as the main problem. It fits perfectly with how he looked, acted, and responded to treatment. **Yeah, I feel pretty confident saying that's what went wrong here.**

Final Response
Based on the provided information, the most likely diagnosis for this neonate is **duodenal atresia**.

Figure 12: HuatuoGPT-o1-8B output (Level 0 output).

Gpt-4o-mini output (Level 0 masking)

AI MODEL OUTPUT:
Thinking Alright, let's break this case down step-by-step.
Key Clinical Findings and Symptoms 1. **Age and Birth Weight**: A 5-day-old male, low birth weight at 1700g. 2. **Clinical Presentation**: Presenting with non-bilious vomiting since birth and has not emitted meconium. 3. **Physical Examination**: Appears in poor general condition, features a 'trisomic facies,' and has a distended abdomen ("plat abdomen"). 4. **Imaging Findings**: - Initial abdominal X-ray showed a single bubble of gastric air with no gas beyond it—often indicative of a proximal obstruction. - Second X-ray revealed a "double bubble" sign, which is characteristic of duodenal atresia. 5. **Lab Results**: Biological check-up was normal.
Differential Diagnoses 1. **Duodenal Atresia**: The most natural consideration given the classic presentation of double bubble on imaging and clinical findings of non-bilious vomiting and failure to pass meconium. 2. **Pyloric Stenosis**: Possible due to the age (early infant) and vomiting, but it's typically associated with projectile vomiting and usually presents after a few weeks, which does not fit here. 3. **Hirschsprung Disease**: Possible given the failure to pass meconium, but this usually causes a more distal obstruction leading to abdominal distension and is less typical in the upper GI tract. 4. **Malrotation with Midgut Volvulus**: This can present with similar symptoms but is generally associated with bilious vomiting—here, the vomiting is non-bilious, which makes this diagnosis less likely. 5. **Intestinal Atresia (other than duodenum)**: Possible if a different segment of the intestines were obstructed, but would have been less likely given the prominent double bubble sign, particularly localized to the duodenum.
Evaluation of Diagnoses - **Duodenal Atresia**: Given the key symptoms (non-bilious vomiting, double bubble on X-ray) and the surgical findings that confirmed the diagnosis, this remains the most compelling diagnosis. - **Pyloric Stenosis**: Rules out based on age of onset and imaging findings. - **Hirschsprung Disease**: Less likely because it does not correlate with the X-ray findings and the absence of bowel gas does not suggest a distal obstruction consistent with Hirschsprung. - **Malrotation with Midgut Volvulus**: Less likely due to non-bilious vomiting. The X-ray findings do not support this. - **Intestinal Atresia**: Other atresias are possible but less common; the imaging and clinical picture strongly suggest a proximal issue localized to the duodenum.
Elimination of Unlikely Diagnoses After evaluating the above conditions, it is clear that: - Pyloric stenosis and malrotation are eliminated due to inconsistency with the clinical picture. - Hirschsprung's Disease does not match the typical signs we see, which leads us to lean away from it as well.
Conclusion Putting all of this together, the findings are most consistent with **duodenal atresia**, especially in the context of the associated "trisomic facies" which can sometimes link to syndromic conditions like Down syndrome, known to have higher rates of duodenal atresia.

Final Response
The most likely diagnosis for this 5-day-old male is **duodenal atresia**, supported by the classic double bubble sign on abdominal X-rays, non-bilious vomiting since birth, and the absence of meconium passage. The immediate clinical picture and surgical findings confirm the diagnosis of duodenal atresia.

Figure 13: Gpt-4o-mini output (Level 0 output)