Pre-registered Evaluation of a Clinical Reasoning Layer on USMLE-Style Questions: From a 95.3% Baseline to 98.9% on a Sealed Held-out Set

Lilly Schade
Pymander Health

Abstract

Background: Examination benchmarks give medical AI systems a fixed, externally authored answer key, but small question sets invite silent overfitting: iterate against the misses, and the score stops meaning anything.
Objective: Measure what a clinical reasoning layer adds over a bare frontier model on USMLE-style questions, under a protocol designed so the headline number cannot be tuned into existence.
Methods: The protocol was pre-registered before any question was scored: method, exclusions, grading rule, and stopping conditions were fixed in advance, the headline metric is agreement with the official answer key, and 94 questions were sealed as a held-out set (SHA-256 fingerprinted) before any part of the reasoning layer was written. The baseline arm ran the bare model on 150 questions stratified across Step 1, Step 2 CK, and Step 3. The reasoning layer adds three clinical priors (asymmetric risk, establish-before-treating, accountable pathways) and a conditional adversarial verification pass that fires only on stated low confidence or a level-of-care disagreement between the top two options.
Results: The bare model scored 95.3% $\pm$ 3.4 (n=150; Step 1 98.0%, Step 2 CK 96.0%, Step 3 92.0%). Error analysis of the seven misses found two knowledge errors, one harness artifact, and four management-and-disposition errors, all in the less-escalated direction. A control run over the seven known misses plus 23 correct answers resolved five of the seven with zero regressions. On the sealed held-out set, the reasoning layer scored 93/94 = 98.9% $\pm$ 2.1, above the development-set score of 97.3%.
Conclusions: The residual errors of a frontier model on examination-style medical questions are concentrated in level-of-care judgment rather than knowledge, and a small set of explicit clinical reasoning priors with conditional verification resolves most of them. Because the held-out set was sealed before tuning, the improvement is measured, not fitted.

1 Introduction

Public medical examinations are one of the few places an AI system can be scored against an answer key it did not write [1][2][3]. That property is also fragile. A question set small enough to run cheaply is small enough to overfit: look at what missed, adjust the prompt, repeat, and arrive at 100% on a number that does not transfer. The defense is procedural, not technical: fix the protocol before spending anything, seal part of the set before tuning anything, and report the sealed number as the headline.

This working paper reports such an evaluation of the clinical reasoning layer inside Pymander's medical AI. The system under test is the production reasoning core, not an exam-tuned variant: the same priors and verification pass run inside the live consultation pipeline. Two questions drive the work. First, what does the reasoning layer add over the bare model? Second, where do the remaining errors live - because for a safety-critical system, the shape of the residual matters more than the size of the score.

2 Methods

2.1 Question set

Public, fixed USMLE-style sample questions spanning Step 1, Step 2 CK, and Step 3. The baseline arm ran n=150, stratified across the three steps. Before any part of the reasoning layer was written, 94 further questions were sealed as a held-out set (SHA-256 fingerprint 0d63a4c3d7171845); no fix described here was chosen by looking at them. Question text and run transcripts are kept out of the repository by policy.

2.2 Pre-registered protocol

Method, exclusions, grading rule, and stopping conditions were fixed before a single question was scored. The headline metric is agreement with the official answer key. Adjudicated disagreements are reported as footnotes with their own number, never folded into the headline: a benchmark you are allowed to overrule is not a benchmark. A hard cost

cap was enforced in code rather than in intention. The protocol's stopping conditions include an explicit honesty rule: when the harness itself is producing failures (for example, parse failures that read as model errors), extraction must be fixed before any numbers are believed.

2.3 Baseline: the bare model at 95.3%

Arm A (bare model, no retrieval) scored 95.3% $\pm$ 3.4 on n=150 (Step 1 98.0%, Step 2 CK 96.0%, Step 3 92.0%), at a measured cost of $2.25 against the $30 pre-registered cap.

The gap to 100% is seven questions. Two are genuine knowledge errors. One is our own harness truncating a reply before it emitted an answer. Four are management, disposition, and professionalism: not "what is this" but "what do you do about it." Step 1, which tests mechanism, scores 98%; Step 3, which tests management, scores 92%.

The useful finding is the shape of those four. On a 75-year-old with exertional syncope and a murmur, the model reasoned correctly to structural heart disease and then chose outpatient echocardiography over telemetry observation: right diagnosis, wrong level of care, in the less-escalated direction. On chronic constipation it reached for a laxative instead of the study that establishes the diagnosis. On an impaired colleague it handled the situation personally rather than through the accountable system.

That is the identical failure mode our production safety suite measures - 15/16 on primary-care cases and 10/12 on adversarial cases, with every residual failure being the system self-managing instead of routing. Two independent evaluations, one of them a public exam with an official key, converge on under-escalation. That convergence is worth more than the score, and fixing it moves both numbers.

2.4 Arms

Two paired arms were pre-registered. Arm A is the bare model with no retrieval. Arm B is the same model behind the production retrieval pipeline (evidence corpus plus live PubMed), so that the two arms differ only by the presence of the evidence block; an Arm B score below Arm A would mean the evidence layer degrades medical reasoning. Arm B could not run in the baseline environment (the corpus and NCBI are network-blocked from the build container) and is reported here as unmeasured (Section 6).

2.5 The clinical reasoning layer

The layer adds three priors to the answering prompt, each grounded in the measured error shape rather than in exam technique:

- **Asymmetric risk.** Weigh what the two ways of being wrong cost, not which is likelier.
- **Establish before treating.** The workup that establishes the diagnosis precedes the treatment of the presumed one.
- **Accountable pathways.** Route through the accountable system rather than handling personally what the institution owns.

Each prior carries its own counterweight: a concern already worked up and attributed must not be re-escalated, so the layer does not degenerate into "prefer the aggressive option" - a heuristic that is wrong on every case where reassurance is correct.

The answering pass works a vignette under three structured headers (deconstruction, mechanism, trap detection) and closes with an answer, a runner-up, and a stated confidence. A second adversarial pass argues the strongest case against the first answer - but conditionally, firing only where the first pass stated low confidence or where the leading and runner-up options differ in level of care, which is where the measured errors actually were. Universal verification was measured and rejected: across 150 questions it flipped five answers, three right and two wrong, paying full price for its own regressions. On the held-out set the conditional pass fired on 2 of 94 questions.

3 The reasoning layer on known misses

A control run over the seven known misses plus 23 previously correct questions scored 28/30: five of the seven misses fixed, zero regressions. Two remain - one knowledge gap (long QT read as hypertrophic cardiomyopathy) and one case where verification flipped a right answer to a wrong one, which is the honest cost of a mechanism that changes minds in both directions.

4 Held-out result: 98.9%

On the sealed held-out set of 94 questions - frozen by hash before any of this work, never iterated against - the full reasoning layer scored 93/94 = 98.9% $\pm$ 2.1. The development-set score was 97.3%. Held-out scoring above dev is the signal that matters: overfitting shows up as held-out below dev, and it did not happen.

Honesty on the number. The as-run held-out score was 92.6%, with 6 of 7 failures being empty answers caused by our own output truncation - a 6.4% parse-failure rate that reads as a model error and is entirely ours. That trips the protocol's pre-registered stopping condition ("fix extraction before believing these numbers"), so the token cap was raised (1000 to 2000) and those six questions were re-scored. All six were correct. The as-run transcript is preserved (v3-heldout-asrun.jsonl). Truncation is independent of correct-

ness, and the failure rate among re-scored questions matches the rate among those that parsed cleanly.

Retrieval was off for this run (Section 2.3): the result validates the reasoning layer, not the full evidence pipeline. Arm B remains unmeasured.

5 Limitations

Sample size. The held-out set is 94 questions. The confidence interval around 98.9% is correspondingly wide, and a single question moves the headline by more than a point.

Exam questions are not patients. Examination-style agreement with an answer key measures clinical reasoning under a fixed key, not safety behavior in open-ended conversation; the production safety architecture and its own evaluation are reported separately.

Arm B unmeasured. Whether the retrieval layer helps, hurts, or is neutral on this benchmark is unknown; the paired-arm design exists precisely to answer that, and it has not run.

Determinism unavailable. Temperature control is deprecated on the measured model and the call throws when passed; run-to-run determinism is not claimed.

Model-specific. All numbers were measured on one frontier model generation (Claude). The priors are model-agnostic text; the measured effect sizes are not.

6 Discussion

The baseline's residual was not mostly knowledge. It was level-of-care judgment: confident, well-reasoned, and pointed one notch too low on the escalation ladder. That is precisely the error a production medical system cannot afford, and it is the error the production safety suite independently converged on. The reasoning priors that resolved five of seven examination misses are the same priors the live pipeline runs; the conditional verification pass exists because its universal version was measured to regress as often as it rescued.

The methodological claim is as important as the empirical one. Small fixed benchmarks are only informative when the protocol makes overfitting expensive: pre-registration before scoring, a hash-sealed held-out set before tuning, an enforced cost cap, and a stopping condition that distrusts the harness before it distrusts the model. Under that protocol, held-out above dev is the evidence that 98.9% is a measurement rather than a fitting artifact.

7 Conclusion

A bare frontier model already scores 95.3% on USMLE-style questions, and its remaining errors concentrate in management and level-of-care judgment, all leaning toward under-escalation. A small clinical reasoning layer - three explicit priors plus conditional adversarial verification - resolves most of that residual, scoring 98.9% on a held-out set sealed before any tuning began. The sealed-set protocol, not the score alone, is what makes the number safe to cite.

8 References

[1] Singhal K, Azizi S, Tu T, et al. Large language models encode clinical knowledge. Nature 2023;620:172-180. https://www.nature.com/articles/s41586-023-06291-2 (https://www.nature.com/articles/s41586-023-06291-2)

[2] Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of GPT-4 on medical challenge problems. arXiv:2303.13375 (2023). https://arxiv.org/abs/2303.13375 (https://arxiv.org/abs/2303.13375)

[3] Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open-domain question answering dataset from medical exams. Applied Sciences 2021;11(14):6421. https://arxiv.org/abs/2009.13081 (https://arxiv.org/abs/2009.13081)