DIAGRARE-BENCH: SEPARATING DISEASE PRIORS FROM PATIENT EVIDENCE IN LANGUAGE MODELS FOR CLINICAL DIAGNOSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Diagnostic benchmarks tell us whether a language model selected the correct disease, but they do not identify what drove that decision. A diagnosis can rank highly because the findings in the case support it, because the disease is common, or because both signals agree. We introduce DIAGRARE-BENCH to separate these effects at the level of the model's diagnostic ranking. Given a ranked differential, we fit a Plackett-Luce choice model with two interpretable features: a disease-prior feature and a case-specific evidence feature. The fitted coefficient γ_{evidence} measures how strongly pairwise diagnostic log odds change with an evidence contrast when the prior contrast is held fixed. The estimator is model agnostic and uses ranked outputs only.

We evaluate 25 open-weight checkpoints, spanning both general-purpose models and models specialized for medicine. The primary evaluation contains 984 diagnostic cases covering 82 diseases and nine organ systems; 21 checkpoints yield identifiable baseline coefficients. Evidence responsiveness is associated with top-1 accuracy ($r = 0.843$), but our main tests ask whether the coefficient predicts behavior when the evidence is changed independently of the rankings used to estimate it. The cross-model ordering transfers to all 3,562 CUPCase reports ($r = 0.662$) and all 22,901 RareArena cases ($r = 0.790$) under a different evidence representation. In a randomized 3×3 experiment that varies evidence strength and stated prevalence independently, baseline γ_{evidence} predicts the change in target choice caused by stronger evidence ($r = 0.689$, $p = 5.5 \times 10^{-4}$). In a sequential diagnosis task, it predicts whether a model revises an initial competing diagnosis after target-specific evidence arrives ($r = 0.860$, $p = 9.8 \times 10^{-6}$). QLoRA adaptation, self-consistency, and test-time reasoning experiments further show that diagnostic accuracy and evidence responsiveness can change by different amounts. These results support evidence responsiveness as a distinct and measurable property of diagnostic decision behavior.

1 INTRODUCTION

Clinical diagnosis combines two kinds of information. Patient findings determine how well a disease explains the case, while disease prevalence affects how plausible that disease is before those findings are considered. These signals often agree, so a correct diagnosis does not reveal which one drove the model's decision. The distinction becomes important when they conflict, as when a rare disease is strongly supported by a characteristic imaging finding, laboratory abnormality, or combination of symptoms.

Most diagnostic evaluations report accuracy, target rank, or reasoning quality. These measures do not show how strongly a model's ranking depends on the evidence in the case. Two models can have the same top-1 accuracy while responding differently to changes in patient evidence, and accuracy can improve through mechanisms such as better candidate recall or more stable decoding. We therefore ask a more specific question: when patient evidence and disease prior disagree, how much does the model's diagnostic ranking move with the evidence?

Recent benchmarks have made diagnostic evaluation more realistic through case reports, structured clinical records, longitudinal interactions, and explicit reasoning tasks (Ma et al., 2024; Jiang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2025; Perets et al., 2025; Chen et al., 2026; Wu et al., 2025; Zhu et al., 2026; Ness et al., 2025). These settings expose failures that examination-style accuracy can miss, but they still do not separate the contribution of disease prior from the contribution of case-specific evidence.

We study this distinction through the ranked differential produced by the model. A language model and its decoding rule induce an ordering over candidate diagnoses. We approximate this ordering with a two-feature Plackett-Luce model, using one feature for disease prior and one for case-specific evidence. The fitted evidence coefficient, γ_{evidence}, has a direct pairwise interpretation: it measures how fitted diagnostic log odds change with an evidence difference when the prior difference is held fixed. This is an output-level description of the decision rule and does not assume that the model explicitly represents these quantities internally.

This formulation gives a direct empirical prediction. If γ_{evidence} measures sensitivity to evidence, models with larger baseline values should change their decisions more when the evidence contrast is strengthened independently. We test this prediction in stages. We first estimate γ_{evidence} on 984 diagnostic cases spanning 82 diseases and nine organ systems. We then test whether the model ordering transfers to CUPCase and RareArena, where both the case distribution and evidence representation change. Next, we vary evidence strength and stated prevalence independently in a randomized factorial experiment. Finally, we present evidence sequentially and measure whether a model revises an earlier competing diagnosis when more specific evidence arrives.

The same pattern appears across these settings. Evidence responsiveness is associated with diagnostic accuracy, transfers across the two case-report datasets, predicts response to randomized evidence, and predicts recovery after an initial competing diagnosis. The corresponding prior coefficient does not show the same interventional relationship, so we use it mainly to separate prevalence from evidence. We also evaluate QLoRA adaptation, self-consistency, and Qwen3 reasoning budgets, and find that diagnostic accuracy and evidence responsiveness can change by different amounts.

Our contributions are:

- **Evidence responsiveness from ranked diagnoses.** We introduce a model-agnostic Plackett-Luce estimator that separates case-specific evidence from disease prior using ranked outputs alone.
- **An evaluation that separates prior from evidence.** We construct 984 cases covering 82 diseases and nine organ systems, together with randomized and sequential evidence experiments.
- **Validation under independent evidence changes.** We show that the cross-model ordering transfers across two case-report datasets and predicts both randomized evidence effects and sequential diagnostic revision.
- **A distinction between correctness and evidence use.** We show that adaptation and inference procedures can change diagnostic accuracy and evidence responsiveness differently.

2 Related Work

Language models evaluated on clinical diagnosis. Evaluation has moved from medical question answering toward settings that more closely resemble diagnostic work. CliBench and MedAgentBench study decisions over structured clinical records and EHR actions (Ma et al., 2024; Jiang et al., 2025). CUPCase and RareArena evaluate uncommon and rare diagnoses from case reports (Perets et al., 2025; Chen et al., 2026). MedCaseReasoning and DiagnosisArena focus more directly on diagnostic reasoning (Wu et al., 2025; Zhu et al., 2026), while MedFuzz shows that aggregate accuracy can hide sensitivity to irrelevant changes in the input (Ness et al., 2025). Our setting is complementary. We do not introduce another notion of diagnostic correctness. Instead, we ask how much a model's ordering over diagnoses changes with clinically relevant evidence after a disease-prior feature is accounted for.

Representations and decision behavior. A large literature studies whether semantic variables can be decoded from hidden states or whether learned concepts have simple geometry in representation space (Park et al., 2025). Our measurement is intentionally at the output level. We fit a low-dimensional utility model to the ordering revealed by the final diagnostic output rather than probing hidden activations. This distinction matters because information can be represented internally without affecting the decision. Conversely, a decision-level coefficient does not identify the internal representation or computation that produced the ranking. We therefore use representation-learning

language only to motivate the separation between information that may be present in the network and information that is expressed in the final prediction.

Choice models and targeted validation. Luce, conditional-logit, and Plackett-Luce models are standard tools for relating choices to observable attributes of the alternatives (Luce, 1959; McFadden, 1974; Plackett, 1975). We use Plackett-Luce as a compact approximation to the diagnostic ordering induced by a language model and its decoder. The coefficient is meaningful only if it predicts behavior outside the rankings used to estimate it. Our main validation therefore changes the case distribution, manipulates evidence directly, and measures sequential revision. This follows a broader direction in model evaluation that separates an aggregate score from the behavior that gives rise to it (Liu et al., 2025).

3 Problem Formulation

3.1 Diagnostic ranking as a choice model

Let x_i denote a diagnostic case and $\mathcal{D}_i$ its candidate diagnoses. A model m, together with its decoding rule, induces an ordering over $\mathcal{D}_i$. We write $s_m(x_i, d)$ for the latent score that would generate this ordering, while observing only the ranked list itself. To describe how the ordering varies with prior and evidence, each candidate receives the feature vector

$$\phi_i(d) = [p(d) \quad e_i(d)]^\top, \tag{1}$$

where $p(d)$ is a disease-prior feature and $e_i(d)$ is a case-specific evidence feature. We approximate the induced ordering with

$$U_{m,i}(d) = \theta_m^\top \phi_i(d) = \gamma_{\text{prior}}^{(m)} p(d) + \gamma_{\text{evidence}}^{(m)} e_i(d). \tag{2}$$

Equation 2 is a statistical approximation to the observed ranking. It gives us a low-dimensional coordinate system for the final decision without making a mechanistic claim about the network. In particular, we do not assume that the model contains two scalar latent variables or computes this utility internally.

For a top-k ranking $r_i = (d_{i1}, \ldots, d_{ik})$, the Plackett-Luce likelihood is

$$P_{\theta_m}(r_i) = \prod_{j=1}^{k} \frac{\exp U_{m,i}(d_{ij})}{\sum_{d \notin \{d_{i1}, \ldots, d_{i,j-1}\}} \exp U_{m,i}(d)}. \tag{3}$$

We estimate $\theta_m = (\gamma_{\text{prior}}^{(m)}, \gamma_{\text{evidence}}^{(m)})$ by regularized maximum likelihood,

$$\widehat{\theta}_m = \arg\max_{\theta \in \mathbb{R}^2} \sum_i \log P_\theta(r_i) - \frac{\lambda}{2} \|\theta\|_2^2, \qquad \lambda = 10^{-3}. \tag{4}$$

3.2 Evidence responsiveness

For two candidates a and b, define $\Delta p_{ab} = p(a) - p(b)$ and $\Delta e_{i,ab} = e_i(a) - e_i(b)$. The fitted pairwise log odds satisfy

$$\log \frac{P(a \succ b \mid x_i)}{P(b \succ a \mid x_i)} = \gamma_{\text{prior}} \Delta p_{ab} + \gamma_{\text{evidence}} \Delta e_{i,ab}. \tag{5}$$

Therefore

$$\frac{\partial}{\partial \Delta e_{i,ab}} \log \frac{P(a \succ b \mid x_i)}{P(b \succ a \mid x_i)} = \gamma_{\text{evidence}}. \tag{6}$$

We refer to γ_{evidence} as evidence responsiveness under the specified protocol. A larger value means that the fitted ordering changes more strongly with an evidence difference after the prior difference is held fixed. The coefficient is therefore comparative rather than absolute: its numerical value depends on the scale of e, the prompt, the candidate set, and the decoding rule. Within an experiment we keep these choices fixed. Across datasets with different evidence representations, we compare the ordering of models rather than the raw coefficient magnitude.

Equation 5 also gives a direct validation target. If an intervention changes the evidence contrast from Δe to $\Delta e + \delta$ while holding Δp fixed, the rank model predicts

$$\Delta \log \mathrm{odds}(a \succ b) = \gamma_{\mathrm{evidence}}\, \delta. \tag{7}$$

Thus, if $\gamma_{\mathrm{evidence}}$ captures response to evidence, models with larger baseline values should show larger changes under an independently assigned evidence contrast.

4 DIAGRARE-BENCH

4.1 PRIMARY DIAGNOSTIC CASES

The primary evaluation contains 984 diagnostic cases covering 82 diseases and nine organ systems. Each disease contributes 12 cases distributed across three difficulty levels. A case specifies findings that are present and findings that are explicitly absent. At higher difficulty, the target shares more of its presentation with a more prevalent disease from the same organ system, while retaining findings that distinguish the target. Rare and common targets are balanced so that prevalence alone cannot solve the task.

For diagnosis d, the prior feature is $p(d) = \log P(d)$, where $P(d)$ is the prevalence recorded in the ontology. The evidence feature is a log-likelihood-ratio score obtained from the findings in the case. The ontology distinguishes required, supporting, excluding, and unrelated findings. The scoring parameters are fixed before the model comparison and varied in the sensitivity analysis. A reference scorer defined from the same ontology reaches 98.1% top-1 and 99.8% top-3 accuracy, which verifies that the intended diagnosis is recoverable from the encoded disease-finding relations.

4.2 EXTERNAL CASE REPORTS

The structured cases make prior and evidence easy to define separately, but that structure could also make the result specific to our ontology. We therefore repeat the rank analysis on all 3,562 CUPCase reports (Perets et al., 2025) and all 22,901 RareArena cases (Chen et al., 2026). These case-report datasets are independently constructed and do not provide our disease-finding ontology, so the evidence representation must change. We set $p(d) = \log(1 + \text{PubMed count})$ and define $e_i(d)$ as TF-IDF cosine similarity between the case text and a description of diagnosis d. This evidence feature selects the correct candidate in 49.6% of CUPCase cases and 35.5% of RareArena cases, above the 25% chance rate of the four-candidate task. Since the feature scales differ from the primary benchmark, we evaluate transfer through cross-model rank correlation and Pearson correlation of model-level coefficients.

4.3 RANDOMIZED EVIDENCE AND STATED-PRIOR EXPERIMENT

Cross-dataset agreement would show that the model ordering is not tied to one benchmark, but it would still be observational. To test the interpretation of $\gamma_{\mathrm{evidence}}$ directly, we construct a separate 3×3 factorial experiment over 27 target and competing-disease pairs. Target-supportive evidence is assigned to weak, medium, or strong. The prompt independently states that the target is rarer than, equally common as, or more common than the confounder. The model chooses between the two diagnoses.

Let $Y_{m,j}(e, p) \in \{0, 1\}$ indicate whether model m chooses the target for pair j under evidence level e and stated-prior condition p. We define

$$\Delta_E^{(m)} = \frac{1}{3J} \sum_{j=1}^{J} \sum_{p} \left[Y_{m,j}(e_{\mathrm{strong}}, p) - Y_{m,j}(e_{\mathrm{weak}}, p) \right], \qquad J = 27. \tag{8}$$

Δ_E is the finite-difference effect of strengthening the evidence in this experiment. Because evidence strength and stated prevalence are assigned independently, this effect has a causal interpretation with respect to the prompt manipulation and the model output. It is not a claim about causal structure inside the network.

4.4 SEQUENTIAL DIAGNOSTIC REVISION

The factorial experiment measures how a model responds before it has committed to a diagnosis. Diagnostic interactions pose a harder problem: new evidence often arrives after an initial assessment has already been made. We therefore construct 82 target and competing-disease pairs with two evidence blocks presented in opposite orders. In the confounder-first order, the first block favors the confounder and the second adds target-specific evidence. Let $Y_{m,j,t}^{C\to T}$ denote target choice at step t. Recovery is

$$R_m = \frac{\sum_j \mathbf{1}[Y_{m,j,1}^{C\to T} = 0] Y_{m,j,2}^{C\to T}}{\sum_j \mathbf{1}[Y_{m,j,1}^{C\to T} = 0]}. \tag{9}$$

The reverse order contains the same final evidence. We therefore also measure

$$O_m = \frac{1}{J} \sum_j \left(Y_{m,j,2}^{C\to T} - Y_{m,j,2}^{T\to C} \right), \tag{10}$$

which captures final-step order dependence. Positive O_m means that target evidence has more influence when it arrives second.

5 EXPERIMENTAL SETUP

Our public panel contains 25 open-weight checkpoints and intentionally mixes general-purpose models with models specialized for medicine. This lets us study evidence responsiveness across training regimes rather than within a single model family or medical specialization. Twenty-one checkpoints provide enough valid ranked outputs for the primary estimate of γ_{evidence}. Twenty overlap with each external case-report analysis, and 17 contribute at least five eligible observations to the sequential-recovery correlation. Because the experiments have different output formats and eligibility conditions, we report the model count for each analysis instead of reducing every comparison to the smallest common subset.

The primary task asks each checkpoint for a three-diagnosis differential from the 82-disease vocabulary. The external tasks rank four supplied candidates. If a response begins with a valid ranking but later departs from the requested format, we retain the longest valid prefix rather than discarding the entire response. Pearson correlation is the primary cross-model statistic, with Spearman correlation reported as a rank-based check; all tests are two-sided. Intervention uncertainty is estimated by bootstrapping disease pairs. The supplement provides the exact prompts, decoding settings, output rules, model coverage, and uncertainty procedures.

6 RESULTS

6.1 EVIDENCE RESPONSIVENESS IS ASSOCIATED WITH DIAGNOSTIC ACCURACY

We begin with a descriptive check on the primary task. Across the 21 models with identifiable baseline estimates, γ_{evidence} is strongly associated with top-1 accuracy (Pearson $r = 0.843$, $p = 1.6 \times 10^{-6}$; Spearman $\rho = 0.860$, $p = 5.7 \times 10^{-7}$; Figure 1). The prior coefficient is not associated with accuracy ($r = -0.120$, $p = 0.61$). Restricting the analysis to the 17 models with at least 50% usable outputs increases the evidence correlation to $r = 0.974$. This association is useful because it shows that evidence responsiveness is not detached from diagnostic performance. It is not, however, a validation of Equation 6: any model-level statistic that tracks overall quality could correlate with accuracy on the same cases.

6.2 THE MODEL ORDERING TRANSFERS TO INDEPENDENT CASE-REPORT DATASETS

The next test asks whether this ordering survives a substantial change in both the data distribution and the construction of the evidence feature. The evidence coefficient estimated on CUPCase correlates with the primary coefficient at $r = 0.662$ ($p = 0.0015$, $n = 20$), and the RareArena coefficient correlates at $r = 0.790$ ($p = 3.4 \times 10^{-5}$, $n = 20$). CUPCase and RareArena coefficients correlate with each other at $r = 0.925$ ($p = 1.1 \times 10^{-7}$, $n = 17$ in the common panel). The ordering is therefore not specific to the ontology-based evidence score used in the primary benchmark. The

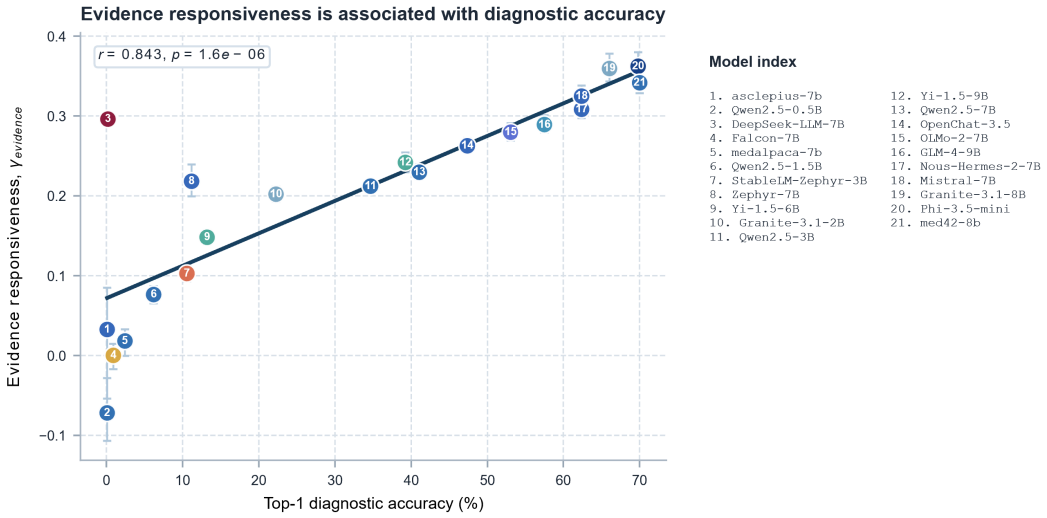

Figure 1: Evidence responsiveness and top-1 diagnostic accuracy. Each point is an open-weight model. The coefficient is estimated from ranked differentials over the same 984 cases. The relationship is descriptive; the following experiments change the data distribution or the evidence itself.

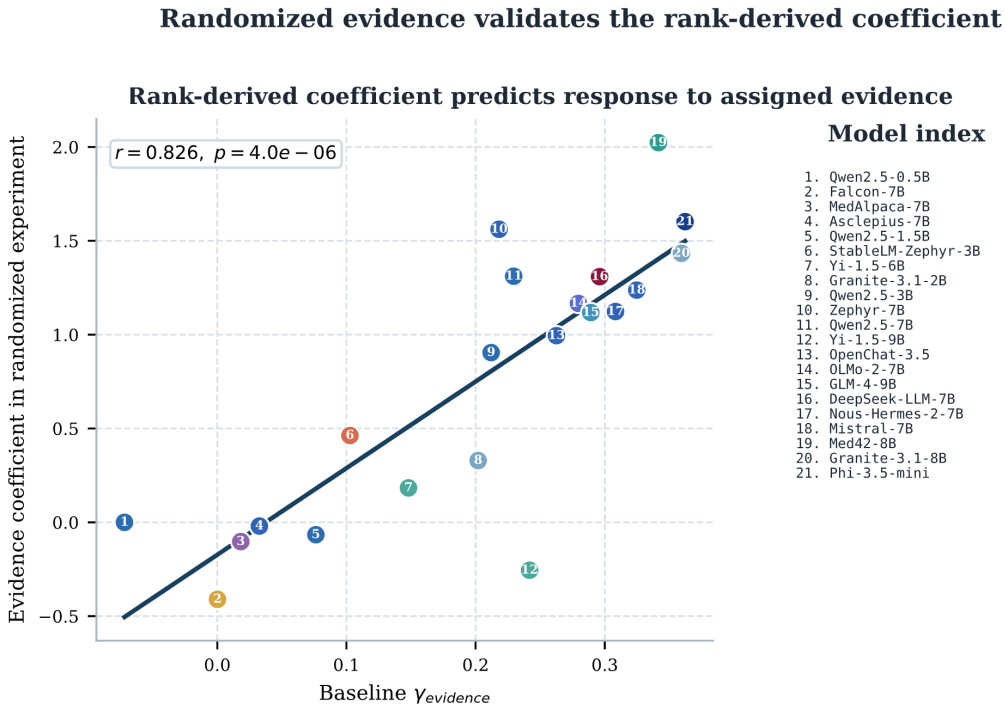

Figure 2: Independent validation with assigned evidence. The horizontal axis is the baseline evidence coefficient estimated from ranked differentials. The vertical axis is the evidence coefficient from the separate randomized evidence and stated-prior experiment.

transfer result makes an ontology-specific explanation less plausible, but it remains correlational. It still does not show that a checkpoint with larger γ_{evidence} reacts more strongly when the evidence itself is manipulated and hence we perform the randomized experiment to test it.

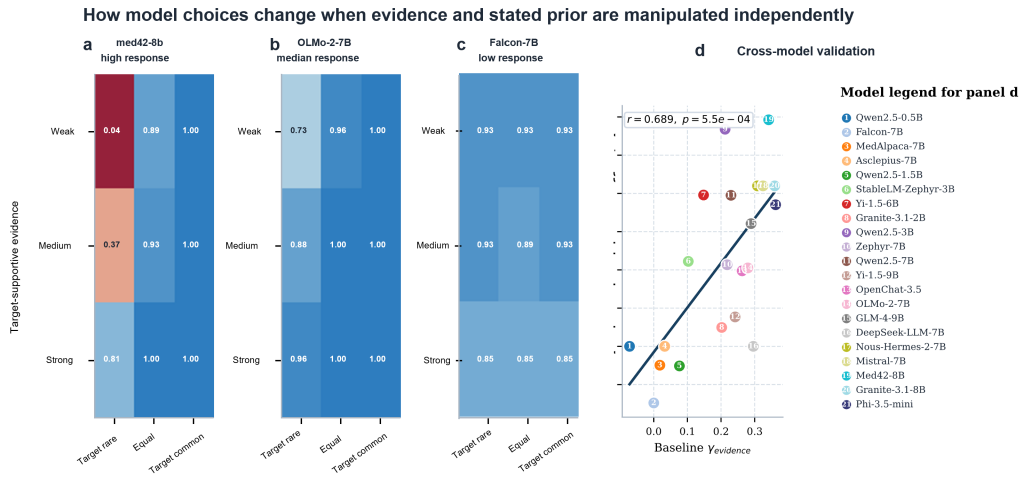

Figure 3: Response to independent changes in evidence strength and stated prevalence. The first three panels show target-choice probabilities for representative models. The last panel compares baseline evidence responsiveness with the finite-difference evidence effect Δ_E.

6.3 Baseline evidence responsiveness predicts the effect of randomized evidence

The randomized experiment gives the most direct test of the rank-model interpretation. Across the model panel, baseline γ_{evidence} predicts the finite-difference effect Δ_E (Pearson $r = 0.689$, $p = 5.5 \times 10^{-4}$; Spearman $\rho = 0.668$, $p = 9.3 \times 10^{-4}$). A related logistic analysis that uses the ordinal evidence levels gives a correlation of $r = 0.826$ ($p = 4.0 \times 10^{-6}$) between the baseline rank coefficient and the intervention coefficient. For the finite-difference result, leave-one-model-out correlations range from $r = 0.639$ to $r = 0.759$.

The evidence manipulation and the stated-prior manipulation have different cross-model structure. Baseline γ_{prior} is not significantly associated with the ordinal stated-prior intervention coefficient (Pearson $r = -0.006$, $p = 0.979$); the corresponding finite-difference analysis using Δ_P is likewise weak and non-significant ($r = -0.143$, $p = 0.535$; Appendix D.2). We therefore treat γ_{prior} as a conditioning coordinate used to account for the specified prior feature in the ranking model, not as a separately validated index of prevalence bias.

6.4 Evidence responsiveness predicts diagnostic revision

The randomized experiment changes evidence before a diagnostic choice is made. The sequential task asks whether the same cross-model ordering predicts a more demanding behavior: revision after the model has already selected the competing diagnosis. Recovery rates vary from 0 to nearly 1 across models. Baseline γ_{evidence} predicts recovery at $r = 0.860$ ($p = 9.8 \times 10^{-6}$, $n = 17$), with Spearman $\rho = 0.811$ ($p = 7.9 \times 10^{-5}$). The relationship remains positive when each model is omitted in turn.

The two presentation orders contain the same evidence at the final step, yet the final diagnosis is not invariant to order. Across models with sufficient coverage, target evidence has greater influence when it appears second. The sequential task therefore measures two properties at once: revision after an earlier decision and path dependence under identical final information.

6.5 Accuracy and evidence responsiveness need not move together

After validating the coefficient under independent evidence changes, we use γ_{evidence} as a second outcome for model interventions. The purpose is not to rank adaptation methods, but to ask whether a change in correctness implies a comparable change in evidence use. QLoRA adaptation of Qwen2.5-7B increases both accuracy and evidence responsiveness. On the original 696-case split, accuracy rises from 40.9% to 63.1% while γ_{evidence} rises from 0.215 to 0.237. On 216 cases from organ

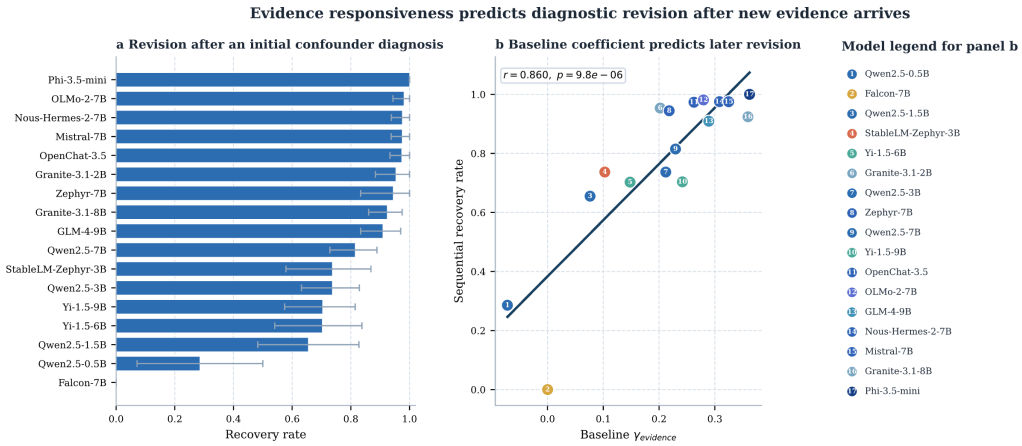

Figure 4: Sequential diagnostic revision. Left: recovery after an initial confounder diagnosis once target-specific evidence is added. Right: recovery rate versus baseline evidence responsiveness. The correlation uses 17 models with at least five eligible recovery cases.

systems excluded from adaptation, accuracy rises from 43.1% to 63.0% and γ_{evidence} from 0.306 to 0.386.

Self-consistency produces a different pattern. Five-sample aggregation increases Qwen2.5-3B accuracy from 34.7% to 41.3%, while γ_{evidence} changes from 0.212 to 0.222. For Qwen2.5-7B, accuracy rises from 41.1% to 47.5%, while γ_{evidence} changes from 0.230 to 0.244.

Qwen3-8B provides a fixed-checkpoint test of additional reasoning tokens. With thinking disabled, the model reaches 74.4% top-1 accuracy, $\Delta_E = 0.309$, and sequential recovery of 0.975. A 512-token thinking budget is exhausted on all 984 baseline ranking prompts, while a 2,048-token budget is exhausted on 89.3% of a 300-case sample. On the shorter binary intervention tasks, Δ_E is 0.309, 0.000, and 0.136 for no-thinking, 512-token, and 2,048-token conditions, respectively; recovery is 0.975, 0.983, and 0.962. Under this protocol, extra reasoning tokens do not improve evidence responsiveness. This result is specific to Qwen3-8B and these budgets.

7 Robustness

We test whether the main findings depend on output parsing, feature construction, or the choice model. With teacher-forced candidate scoring, which removes free-text disease-name parsing, Qwen2.5-1.5B, Mistral-7B-Instruct, and Phi-3.5-mini all retain positive and significant evidence coefficients. Shuffling the evidence feature within each case drives γ_{evidence} toward zero, while shuffling the prior has little effect. A Bradley–Terry estimator also gives a similar model ordering.

The results are stable to the evidence-scoring scheme. Across 16 parameterizations, the highest-scoring diagnosis is unchanged for 99.9% of the 984 cases. Two independently regenerated benchmark variants give Qwen2.5-7B coefficients of 0.226 and 0.228, compared with 0.230 on the primary benchmark. Restricting the candidate set changes coefficient magnitude but preserves the ordering of the tested models.

We also test whether γ_{evidence} simply reflects general prompt sensitivity. Adding a clinically irrelevant sentence produces the opposite pattern: across eight models, γ_{evidence} is negatively correlated with the top-1 flip rate ($r = -0.753$, $p = 0.031$). Models with larger evidence coefficients are therefore more responsive to clinically relevant evidence without being more sensitive to irrelevant text.

8 Discussion

The main result is that a coefficient estimated from ranked diagnoses predicts model behavior under independent changes to the evidence. Its correlation with accuracy is descriptive. The stronger

test comes from the intervention experiments: Equation 5 interprets γ_{evidence} as sensitivity of fitted pairwise log odds to an evidence contrast, the randomized experiment tests the corresponding finite change, and the sequential experiment tests whether the same ordering predicts later diagnostic revision. The transfer results show that this ordering also persists under a different case distribution and evidence representation.

From a representation-learning perspective, γ_{evidence} measures what information is expressed in the final decision, not what can be decoded from hidden states. A model may contain medically relevant information without allowing it to affect the final ranking, while an output-level coefficient does not reveal the internal computation that produced the decision. We therefore treat γ_{evidence} as a property of the observed decision rule under a fixed evaluation protocol.

The sequential experiment also reveals a limitation of static accuracy: models can reach different final diagnoses when the same evidence is presented in a different order. This is important in diagnostic settings where evidence arrives over time. Evidence responsiveness therefore complements accuracy and calibration rather than replacing them. The intervention results support the same conclusion: adaptation, aggregation, or improved candidate generation can raise accuracy without producing the same change in evidence responsiveness.

9 Limitations

The primary benchmark covers 82 diseases and uses simplified disease–finding relations so that prior and evidence can be varied separately. Real clinical evidence is more complex, and the lexical evidence feature used for CUPCase and RareArena should be interpreted as a proxy rather than a calibrated clinical likelihood.

The estimate γ_{evidence} also depends on the prompt, candidate set, feature scale, and decoding rule. We test these choices with coverage restrictions, teacher-forced scoring, alternative estimators, and sensitivity analyses, but the randomized and sequential experiments remain limited to 27 and 82 disease pairs, respectively.

The causal claim is restricted to the randomized prompt intervention: it identifies the effect of changing presented evidence on model output, not causal structure inside the model or human-like medical reasoning. The Qwen3 analysis is similarly limited to one checkpoint and a fixed set of reasoning budgets.

Ethics Statement

The benchmark does not contain patient records or personally identifiable information. CUPCase and RareArena are previously released case-report datasets, which we use under their original terms and with their original attribution. Our experiments are intended to study how language models use diagnostic evidence. They do not evaluate clinical safety or establish that any model is suitable for patient care. In particular, a high evidence-responsiveness score should not be interpreted as evidence that a model is safe or reliable for clinical deployment.

Reproducibility and AI Use Statement

The benchmark data and analysis code are provided with the supplementary material and will be released publicly upon acceptance.

A language model was used for language editing and manuscript organization. The authors reviewed the final manuscript and verified the experimental design, code, data, numerical results, figures, claims, and citations against the underlying analysis outputs.

References

H. Chen, Z. Zhao, S. Zhou, et al. RareArena: a comprehensive benchmark dataset unveiling the potential of large language models in rare disease diagnosis. *The Lancet Digital Health*, 8(2): 100953, 2026. doi: 10.1016/j.landig.2025.100953.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. MedAgentBench: A realistic virtual EHR environment to benchmark medical LLM agents. *arXiv preprint arXiv:2501.14654*, 2025.

Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. RM-Bench: Benchmarking reward models of language models with subtlety and style. In *International Conference on Learning Representations*, 2025.

R. Duncan Luce. *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

Mingyu Derek Ma, Chenchen Ye, Yu Yan, Xiaoxuan Wang, Peipei Ping, Timothy S. Chang, and Wei Wang. CliBench: A multifaceted and multigranular evaluation of large language models for clinical decision making. *arXiv preprint arXiv:2406.09923*, 2024.

Daniel McFadden. Conditional logit analysis of qualitative choice behavior. In Paul Zarembka, editor, *Frontiers in Econometrics*, pages 105–142. Academic Press, 1974.

Robert Osazuwa Ness, Katie Matton, Hayden Helm, Sheng Zhang, Junaid Bajwa, Carey E. Priebe, and Eric Horvitz. MedFuzz: Exploring the robustness of large language models in medical question answering. In *International Conference on Learning Representations*, 2025.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, 2025.

Oriel Perets, Ofir Ben Shoham, Nir Grinberg, and Nadav Rappoport. CUPCase: Clinically uncommon patient cases and diagnoses dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):28293–28301, 2025.

Robin L. Plackett. The analysis of permutations. *Applied Statistics*, 24(2):193–202, 1975.

Kevin Wu, Eric Wu, Rahul Thapa, Kevin Wei, Angela Zhang, Arvind Suresh, Jacqueline J. Tao, Min Woo Sun, Alejandro Lozano, and James Zou. MedCaseReasoning: Evaluating and learning diagnostic reasoning from clinical case reports. *arXiv preprint arXiv:2505.11733*, 2025.

Y. Zhu, Z. Huang, L. Mu, Y. Huang, W. Nie, J. Liu, S. Zhang, P. Liu, and X. Zhang. DiagnosisArena: Benchmarking diagnostic reasoning for large language models. In *Findings of the Association for Computational Linguistics*, pages 3074–3098, 2026.

A Supplementary Overview

The supplement gives the details needed to interpret and reproduce the measurements in the main paper. Section B describes the primary diagnostic cases and the two external case-report datasets. Section C expands the rank model and uncertainty procedure. Sections D and E define the randomized and sequential experiments. Section F records checkpoint coverage and inclusion rules. Section G tests alternative estimators, feature definitions, candidate sets, and negative controls. Sections H and I report model-level comparisons and interventions. Section K specifies the prompting protocol, and Section L describes the public release.

B Structured and External Evaluation Data

B.1 Primary diagnostic cases

The primary evaluation contains 984 cases covering 82 diagnoses and nine organ systems. Each diagnosis is represented by a disease profile containing required findings, supporting findings, findings whose absence is informative, and a disease-frequency attribute. The cases are generated from these profiles rather than from patient records. The evaluation is intentionally narrower than unrestricted diagnosis: the disease vocabulary is fixed, and the data specify which findings are present and absent. This structure is what allows the evidence and prior features to be defined independently.

For case x and candidate diagnosis d, the evidence score $e(d, x)$ summarizes agreement between the observed findings and the disease profile. Required findings receive larger weight than supporting findings, while explicit contradictions reduce the score. The primary scoring rule is fixed before comparing models. To test whether the conclusions depend on the particular weights, we recompute the evidence score over 16 parameter settings for required, supporting, and exclusionary findings. The diagnosis with the largest evidence score is unchanged in 99.9% of the 984 cases across this grid.

The disease-prior feature is derived from prevalence metadata. The main analysis uses a continuous log-prevalence value. Section G.4 shows that replacing this representation with prevalence rank or with three rarity bins leaves the fitted evidence coefficient nearly unchanged for representative models.

B.2 External case-report evaluations

We evaluate all 3,562 released CUPCase reports and all 22,901 RareArena cases. These datasets do not contain the structured disease-finding profiles used in the primary evaluation, so we construct an evidence feature directly from text. For each case, candidate diagnoses are scored by TF-IDF cosine similarity between the case presentation and the candidate diagnosis description. The prior feature is the log of one plus the PubMed article count for the diagnosis name. The text is truncated to 4,000 characters before both model inference and TF-IDF scoring so that the feature and the model receive the same case prefix.

The correct diagnosis has the largest lexical evidence score in 49.6% of CUPCase reports and 35.5% of RareArena cases. Both values exceed the 25% chance level of the four-candidate task. The external analysis therefore uses an evidence signal that is weaker than the structured feature but still informative. Since the numerical scales of the structured and lexical features are different, the paper compares model ordering across datasets rather than raw values of γ_{evidence}.

C Rank Model and Estimation

C.1 Plackett-Luce likelihood

For model m, case x, and diagnosis d, define

$$\phi(d, x) = [p(d) \quad e(d, x)]^\top, \qquad U_m(d, x) = \theta_m^\top \phi(d, x), \tag{11}$$

with $\theta_m = (\gamma_{\text{prior}}^{(m)}, \gamma_{\text{evidence}}^{(m)})^\top$. If the model returns the ordered list $d_1 \succ d_2 \succ \cdots \succ d_K$ from candidate set $\mathcal{D}$, the likelihood is

$$P(d_1, \ldots, d_K \mid x, \theta_m) = \prod_{k=1}^{K} \frac{\exp U_m(d_k, x)}{\sum_{d \in \mathcal{D} \setminus \{d_1, \ldots, d_{k-1}\}} \exp U_m(d, x)}. \tag{12}$$

We estimate θ_m separately for each checkpoint with an L_2 penalty of 10^{-3}. For candidates a and b,

$$\log \frac{P(a \succ b \mid x)}{P(b \succ a \mid x)} = \gamma_{\text{prior}}\big[p(a) - p(b)\big] + \gamma_{\text{evidence}}\big[e(a, x) - e(b, x)\big]. \tag{13}$$

The evidence coefficient is therefore the slope of fitted pairwise log odds with respect to an evidence difference. This interpretation is conditional on the feature scale and candidate set.

The rank model is not assumed to reproduce the network computation. It is a two-feature approximation to the output ordering of the model and decoder. Its usefulness depends on whether the fitted coefficient predicts behavior under independent evidence changes, which is why the randomized and sequential experiments are central to the paper.

C.2 Identifiability and protocol dependence

The estimator requires variation in both candidate features and enough valid ranked outputs to determine the two coefficients. A coefficient is not treated as an intrinsic property of a checkpoint. It can change with the prompt, candidate set, feature scale, and decoding rule. All within-benchmark model comparisons therefore use the same protocol. When the feature definition changes across datasets, we compare relative model ordering.

The method does not require a normalized probability distribution over diagnoses. It only requires an ordered list. This permits estimation from ordinary free-text differential diagnoses. Section G repeats the analysis with teacher-forced candidate scores for three models to verify that free-text parsing is not necessary for the main pattern.

C.3 Uncertainty

For the primary coefficient estimates, we bootstrap at the disease level so that all cases belonging to a resampled disease remain together. This choice treats disease identity as the principal semantic sampling unit. For representative models we also bootstrap at the organ-system level and compute leave-one-disease-out estimates. For Qwen2.5-7B, the disease-clustered 95% interval for γ_{evidence} is $[0.211, 0.251]$, the organ-clustered interval is $[0.211, 0.263]$, and the leave-one-disease-out range is $[0.228, 0.234]$. For Phi-3.5-mini, the corresponding intervals are $[0.328, 0.406]$ and $[0.324, 0.414]$, with a leave-one-disease-out range of $[0.360, 0.372]$.

D Randomized Evidence and Stated-Prior Experiment

The rank coefficient is estimated from naturally varying cases. To test its interpretation more directly, we assign evidence strength and stated prevalence independently.

D.1 Disease pairs and assignment

We use 27 target and confounder disease pairs, with up to three pairs from each of the nine organ systems. Each pair is evaluated under three evidence levels and three stated-prior conditions, producing $27 \times 3 \times 3 = 243$ prompts per model. The stated-prior manipulation is intentionally large so that evidence and prior can be placed in clear conflict. In the target-rare condition, the prompt gives target and confounder rates of 1 and 10,000 per 100,000. The equal condition gives 1,000 and 1,000. The target-common condition reverses the first pair.

Evidence strength is generated from the target disease profile. Weak evidence contains one required target finding. Medium evidence contains approximately half of the required findings and one supporting finding. Strong evidence contains all required findings and up to two supporting findings; where applicable, it also states the absence of findings that define the confounder. These prompts are not part of the data used to fit baseline γ_{evidence}.

D.2 FINITE-DIFFERENCE ESTIMANDS

Let $Y_{m,j}(e,p)$ denote target choice for model m, pair j, evidence level e, and stated-prior condition p. The model-level evidence effect is

$$\Delta_E^{(m)} = \frac{1}{3J}\sum_{j=1}^{J}\sum_{p}\left[Y_{m,j}(e_{\text{strong}},p) - Y_{m,j}(e_{\text{weak}},p)\right]. \tag{14}$$

The stated-prior effect is

$$\Delta_P^{(m)} = \frac{1}{3J}\sum_{j=1}^{J}\sum_{e}\left[Y_{m,j}(e,p_{\text{target common}}) - Y_{m,j}(e,p_{\text{target rare}})\right]. \tag{15}$$

We also compute the evidence-by-prior interaction and the strong-evidence response under the unfavorable target-rare condition. Confidence intervals resample disease pairs.

Across the 21 models with baseline estimates, γ_{evidence} predicts Δ_E with Pearson $r = 0.689$ ($p = 5.5 \times 10^{-4}$) and Spearman $\rho = 0.668$. The matched relationship between γ_{prior} and Δ_P is weak and not significant ($r = -0.143$, $p = 0.535$). The evidence coefficient is therefore the quantity for which the intervention supports the intended interpretation. We use the prior coefficient as a conditioning coordinate and do not make a symmetric validity claim for it.

Random assignment supports a causal statement about the effect of changing the prompt evidence on the model output within this experiment. It does not establish an internal causal representation or human-like causal reasoning.

E SEQUENTIAL DIAGNOSTIC REVISION

The factorial experiment compares completed prompts. We next measure whether a model revises a diagnosis when evidence arrives in two steps. The experiment contains 82 target and confounder pairs. Each pair is presented in two orders. In the confounder-first order, the first block supports the confounder and the second adds target-specific findings. In the target-first order, the same two blocks are reversed. The union of evidence at step two is identical across orders.

Let $Y_{m,j,t}^{C\to T}$ and $Y_{m,j,t}^{T\to C}$ be binary target-choice indicators. The recovery rate is

$$R_m = \frac{\sum_j \mathbf{1}[Y_{m,j,1}^{C\to T}=0]Y_{m,j,2}^{C\to T}}{\sum_j \mathbf{1}[Y_{m,j,1}^{C\to T}=0]}. \tag{16}$$

The final-step order effect is

$$O_m = \frac{1}{J}\sum_j \left(Y_{m,j,2}^{C\to T} - Y_{m,j,2}^{T\to C}\right). \tag{17}$$

Seventeen models have at least five cases in the denominator of R_m. Baseline γ_{evidence} predicts recovery with Pearson $r = 0.860$ ($p = 9.8 \times 10^{-6}$) and Spearman $\rho = 0.811$. All 17 models have a positive mean order effect. Under the sign convention above, this means target evidence has greater influence when it appears second. We therefore describe the effect as recency or path dependence rather than persistent anchoring.

F MODEL PANEL AND INCLUSION RULES

The released model panel contains 25 open-weight checkpoints. It deliberately includes both general-purpose models and models specialized for medicine, since a diagnostic benchmark should not assume that medical specialization is the only route to strong performance. Both groups receive the same diagnostic prompts, candidate vocabulary, scoring rules, and rank estimator. Twenty-one checkpoints provide enough valid ranked outputs to identify the two coefficients in the primary analysis. Later experiments use different response formats, so their model counts are reported separately rather than forcing all comparisons onto one reduced subset.

Table 1 lists the released panel. The table is intentionally descriptive: it records checkpoint identity, model family, and parameter count, while experiment-specific coverage is reported with the analysis that uses it. This keeps the model panel separate from the statistical inclusion rules and avoids attaching qualitative labels to incomplete outputs.

Table 1: Open-weight checkpoints in the released model panel. General-purpose and medically specialized checkpoints are evaluated with the same diagnostic protocol.

Checkpoint	Family	Params. (B)
General-purpose checkpoints		
Qwen2.5-0.5B	Qwen	0.5
Qwen2.5-1.5B	Qwen	1.5
Qwen2.5-3B	Qwen	3.0
Qwen2.5-7B	Qwen	7.0
Mistral-7B-Instruct	Mistral	7.3
Phi-3.5-mini	Phi	3.8
Yi-1.5-6B	Yi	6.0
Yi-1.5-9B	Yi	9.0
OLMo-2-7B	OLMo	7.0
DeepSeek-LLM-7B	DeepSeek	7.0
Falcon-7B	Falcon	7.0
StableLM-Zephyr-3B	StableLM	3.0
Granite-3.1-2B	Granite	2.0
Granite-3.1-8B	Granite	8.0
Zephyr-7B	Mistral	7.2
OpenChat-3.5	Mistral	7.0
Vicuna-7B	Llama2	7.0
TinyLlama-1.1B	Llama	1.1
GLM-4-9B	GLM	9.0
Nous-Hermes-2-7B	Mistral	7.0
Medically specialized checkpoints		
MedAlpaca-7B	Llama2	7.0
OpenBioLLM-8B	Llama3	8.0
Asclepius-7B	Mistral	7.0
Med42-8B	Llama3	8.0
BioMistral-7B	Mistral	7.0

A checkpoint enters the primary coefficient analysis when the two-parameter likelihood is identifiable and there are enough valid ranked responses to estimate the coefficient. We do not impute missing experiments. Every cross-model statistic therefore reports its own model count. This matters most for checkpoints whose free-text output does not consistently follow the requested ranking format.

Because low output coverage can increase variance, we repeat the principal association after imposing stricter coverage thresholds. Requiring at least 50% valid rankings leaves 17 models and increases the γ_{evidence}-accuracy correlation from $r = 0.843$ to $r = 0.974$. Thresholds at 25% and 10% maximum failure give the same qualitative conclusion. A descriptive model-level regression that includes γ_{evidence}, log parameter count, output-failure rate, and model-family fixed effects retains a significant coefficient on γ_{evidence} ($p = 0.0028$). These checks are not used to redefine the panel; they test whether the cross-model relationship is driven by a small set of poorly formatted outputs.

G Robustness of the Rank Model

G.1 Alternative choice likelihood

The main estimator uses Plackett-Luce because the model returns an ordered list. We also fit a Bradley-Terry pairwise likelihood to eight representative checkpoints. The absolute scale changes

Table 2: Representative estimates from the primary ranking protocol. Top-1 accuracy and the two fitted coefficients are shown for checkpoints spanning the observed range of evidence responsiveness. The complete table is released in machine-readable form.

Checkpoint	Type	Top-1 (%)	γ_{prior}	γ_{evidence}
Qwen2.5-0.5B	General	0.1	-0.107	-0.072
Qwen2.5-1.5B	General	6.2	0.007	0.076
Qwen2.5-3B	General	34.7	0.043	0.212
Qwen2.5-7B	General	41.1	0.059	0.230
OLMo-2-7B	General	53.0	0.053	0.280
Mistral-7B-Instruct	General	62.4	0.112	0.325
Phi-3.5-mini	General	69.8	0.150	0.362
Med42-8B	Medical	70.0	0.098	0.342
MedAlpaca-7B	Medical	2.4	0.381	0.018

under the different likelihood, but the ordering is similar. For Qwen2.5-1.5B, Qwen2.5-3B, Qwen2.5-7B, OLMo-2-7B, Mistral-7B-Instruct, and Phi-3.5-mini, the Bradley-Terry evidence coefficients are 0.086, 0.196, 0.238, 0.246, 0.280, and 0.294, respectively. The sequence agrees with the increase seen under Plackett-Luce.

G.2 Teacher-forced candidate scoring

Free-text parsing could confound the analysis if checkpoints with cleaner formatting are also more accurate. For three checkpoints with accessible logits, we score candidate diagnoses directly under teacher forcing and regress the resulting score differences on prior and evidence. The direct evidence coefficient is positive for Qwen2.5-1.5B (0.019, $p = 1.5 \times 10^{-8}$), Mistral-7B-Instruct (0.127, $p < 10^{-200}$), and Phi-3.5-mini (0.065, $p < 10^{-180}$). These coefficients are not numerically comparable to the ranking coefficients because the dependent variable is now a logit-derived score, but the result removes free-text parsing from the measurement.

G.3 Feature-shuffling test

We independently permute one feature within each case and refit the rank model. Shuffling the evidence score reduces γ_{evidence} from 0.230 to 0.003 for Qwen2.5-7B, from 0.362 to 0.008 for Phi-3.5-mini, from 0.325 to -0.012 for Mistral-7B-Instruct, and from 0.212 to 0.003 for Qwen2.5-3B. Shuffling the prior leaves γ_{evidence} nearly unchanged. The selective collapse is consistent with the two features contributing distinct information to the fitted ranking.

G.4 Prevalence representation

The continuous prevalence representation is not necessary for the evidence result. Across eight representative models, replacing continuous prevalence with within-set rank or three rarity bins changes γ_{evidence} only slightly. For Qwen2.5-7B the values are 0.230, 0.229, and 0.229; for Phi-3.5-mini they are 0.362, 0.361, and 0.361. The prior coefficient changes more because the feature scale has been changed directly, which is why we do not compare the magnitude of γ_{prior} across encodings.

G.5 Candidate-set dependence

Choice coefficients depend on the alternatives available to the model. We recompute the rank fit using the full 82-disease set, candidates from the same organ system, and the same-organ set augmented with globally common diseases. Absolute coefficients shrink under the restricted sets, but the ordering of representative models is broadly preserved. Qwen2.5-7B has evidence coefficients 0.230, 0.155, and 0.160 across the three candidate sets; Phi-3.5-mini has 0.362, 0.239, and 0.245. This dependence is expected and is the reason we interpret γ_{evidence} comparatively under a fixed evaluation protocol.

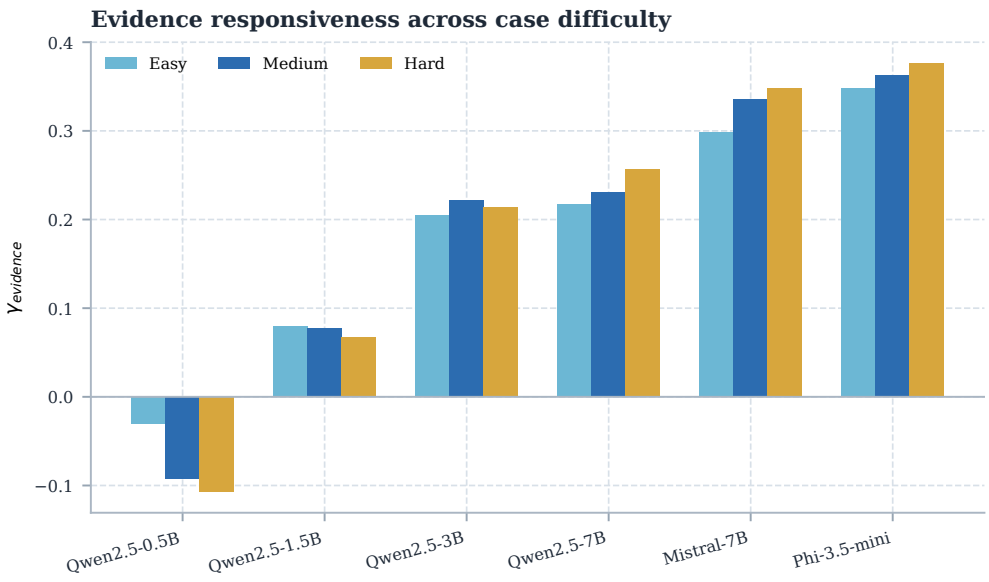

Figure 5: Evidence responsiveness stratified by case difficulty.

G.6 Irrelevant-text negative control

A checkpoint that changes its answer in response to any prompt edit could appear evidence responsive without using diagnostic content selectively. We append one clinically irrelevant sentence to every primary case and measure the top-1 flip rate. Across eight models, baseline γ_{evidence} is negatively correlated with irrelevant-text sensitivity ($r = -0.753$, $p = 0.031$). Models with larger evidence coefficients are therefore not simply more unstable under textual changes.

G.7 Matched rare and common targets

Rare cases could mechanically appear less evidence responsive if they contain fewer or weaker findings. We form 432 rare and common matched pairs controlling organ system, difficulty, number of findings, and target evidence. Mean finding counts are 9.46 for rare cases and 9.35 for common cases; mean target-evidence scores are 16.10 and 15.70. Six of eight tested models have lower γ_{evidence} on the rare subgroup, Phi-3.5-mini is approximately equal, and Qwen2.5-0.5B is unstable because its task performance is near floor. The result is therefore not explained simply by rare cases containing less information.

G.8 Sensitivity across evidence parameters and organ systems

We fit organ-specific coefficients and a hierarchical shrinkage model to check whether the global result is dominated by one clinical domain. The ordering remains visible across organ systems, although smaller strata have wider uncertainty. We also sweep the evidence-score parameters and recompute the model-level association over the full grid. The main relationship remains stable across this range.

H Additional Model-Level Analyses

H.1 Parameter count and medical specialization

Parameter count alone does not explain the main cross-model relationship. A regression using only log parameter count over 25 models has $R^2 = 0.109$ and $p = 0.107$. In the 21-model regression that contains γ_{evidence}, log parameter count, parsing failure rate, and model-family fixed effects,

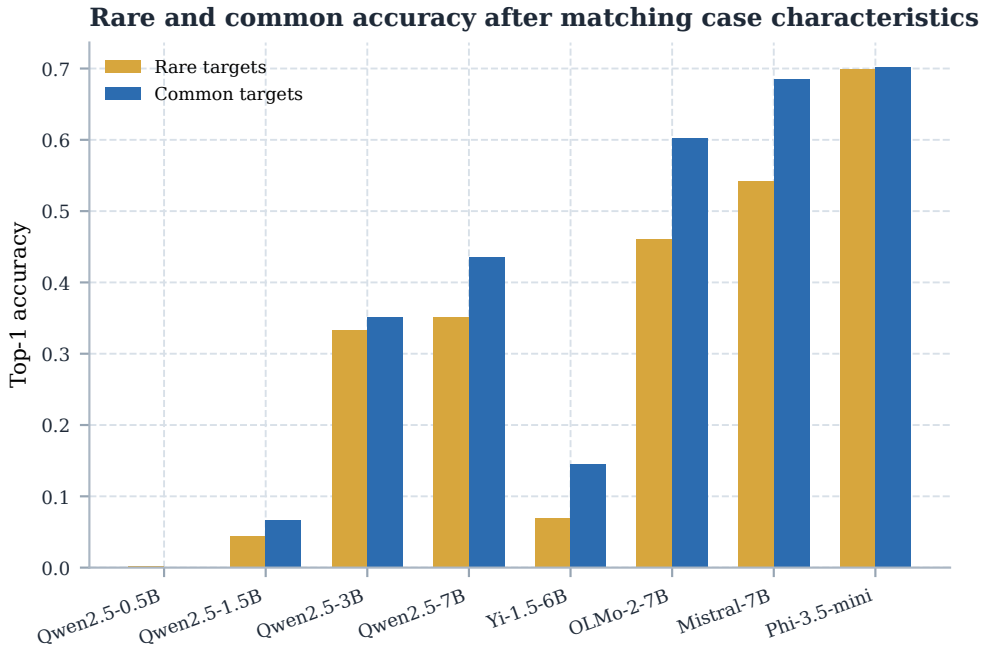

Figure 6: Rare and common accuracy after matching organ system, difficulty, finding count, and target evidence.

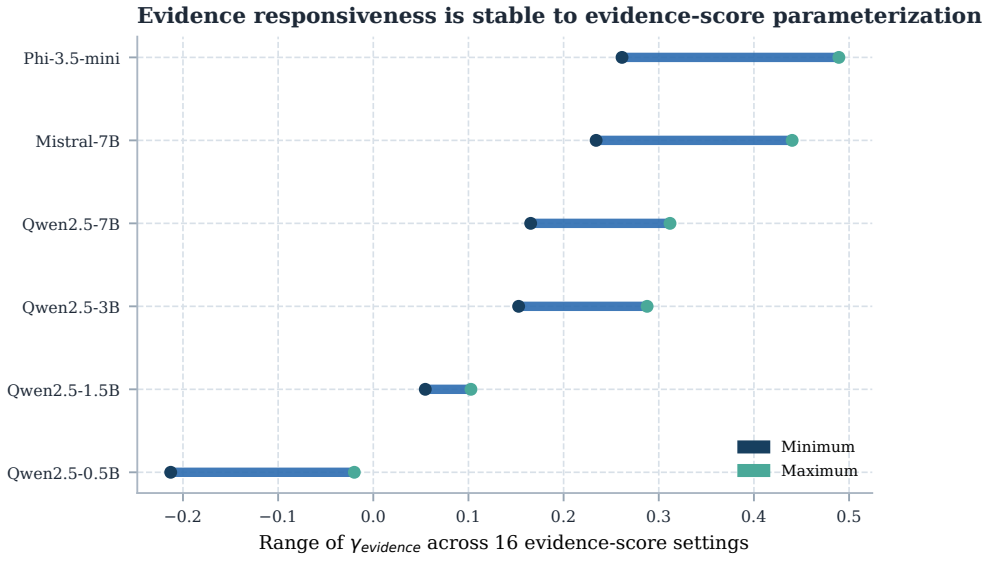

Figure 7: Sensitivity of the cross-model result to alternative evidence-score parameterizations.

γ_{evidence} remains significant ($p = 0.0028$), while log parameter count does not ($p = 0.886$). These comparisons are observational and are not interpreted as causal effects of scale.

We also compare medically specialized and general checkpoints in a restricted panel. After accounting for γ_{evidence} and parameter count, the specialization indicator is not significant ($p = 0.242$). The available medically specialized checkpoints differ in base architecture, data, and training procedure, so this comparison is confounded and underpowered. We report it only as a check that the main result is not produced by a simple specialized-versus-general partition.

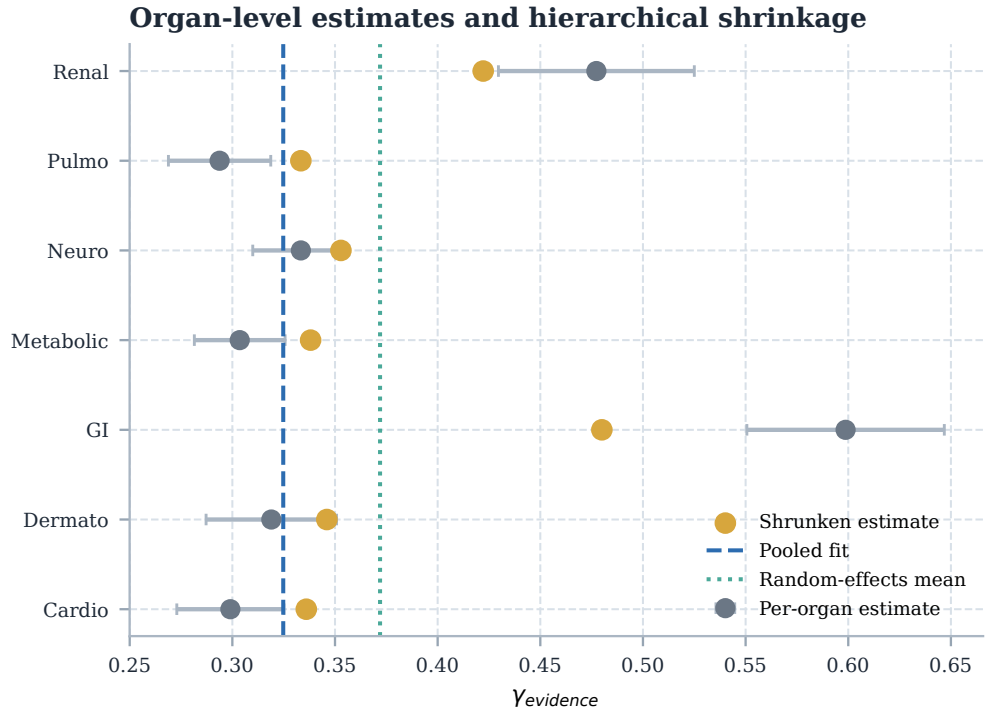

Figure 8: Organ-level estimates and hierarchical shrinkage.

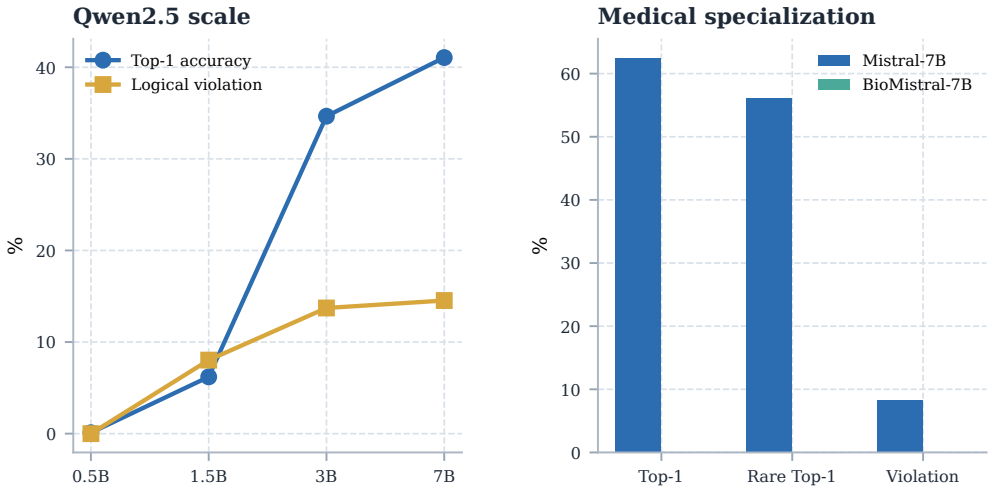

Figure 9: Descriptive analyses of parameter count and medical specialization.

I Model and Inference Interventions

I.1 QLoRA adaptation

We adapt Qwen2.5-7B with QLoRA on 72 examples spanning six diseases. The training examples use the same ranked-diagnosis format as the primary evaluation but do not include the randomized or sequential prompts. On the original 696-case split, top-1 accuracy increases from 40.9% to 63.1%, while γ_{evidence} increases from 0.215 to 0.237. We also evaluate 216 cases from organ systems absent from the adaptation set. Accuracy increases from 43.1% to 63.0%, and γ_{evidence} increases from

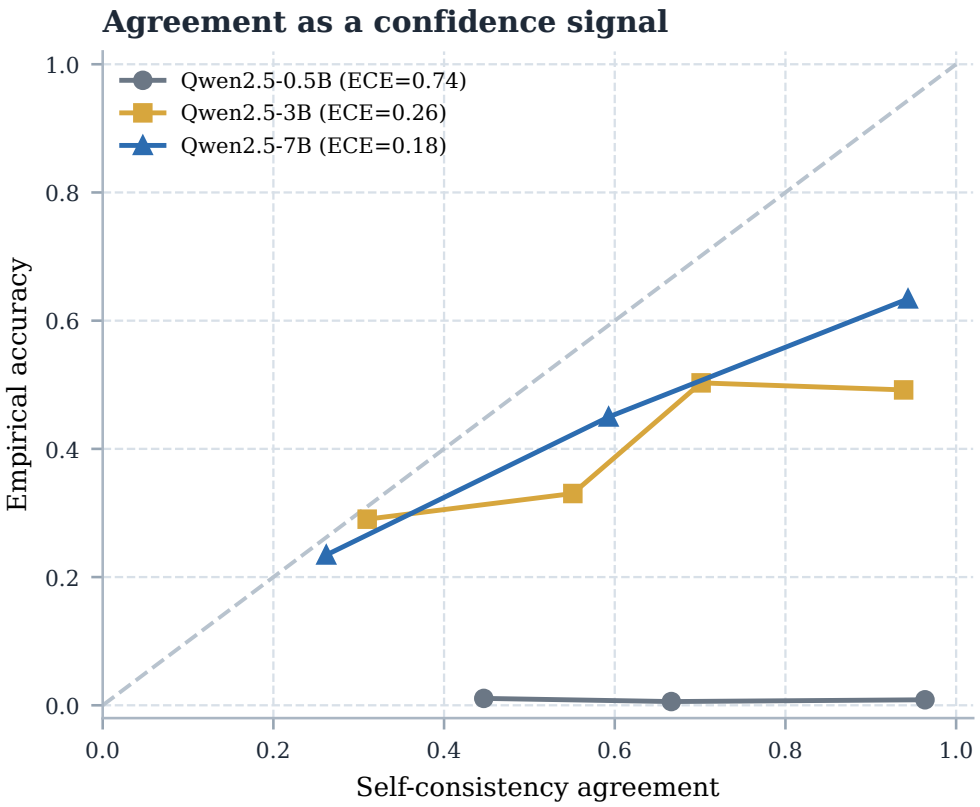

Figure 10: Self-consistency agreement as a confidence signal for the Qwen2.5 checkpoints.

0.306 to 0.386. The effect therefore transfers beyond the organ systems represented in the adaptation examples.

I.2 Self-consistency

For three Qwen2.5 checkpoints, we sample five independent ranked outputs and aggregate them. The 0.5B model remains near floor. For Qwen2.5-3B, top-1 accuracy increases from 34.7% to 41.3%, while γ_{evidence} changes from 0.212 to 0.222. For Qwen2.5-7B, accuracy increases from 41.1% to 47.5%, while γ_{evidence} changes from 0.230 to 0.244. Thus, aggregation improves final correctness more than it changes the estimated dependence of the ranking on patient evidence.

I.3 Prompt and chat-template sensitivity

We evaluate alternative instruction wording and chat-template application for representative checkpoints. A poorly matched template can reduce absolute accuracy, but the evidence-response pattern is stable under the intended chat template for each checkpoint. These runs are diagnostics of protocol sensitivity; prompts are not tuned separately to maximize each model's score.

I.4 Qwen3 reasoning budget

Qwen3-8B supports thinking and no-thinking modes within the same checkpoint, which permits a within-model comparison. With thinking disabled, the model completes nearly all primary ranking prompts and obtains 74.4% top-1 accuracy, $\gamma_{\text{evidence}} = 0.334$, $\Delta_E = 0.309$, and sequential recovery of 0.975. Under a 512-token thinking budget, all 984 primary ranking prompts exhaust the budget and only four produce usable rankings. Under a 2,048-token budget, 89.3% of 300 evaluated primary prompts exhaust the budget and 47 yield usable rankings. The resulting baseline coefficients are

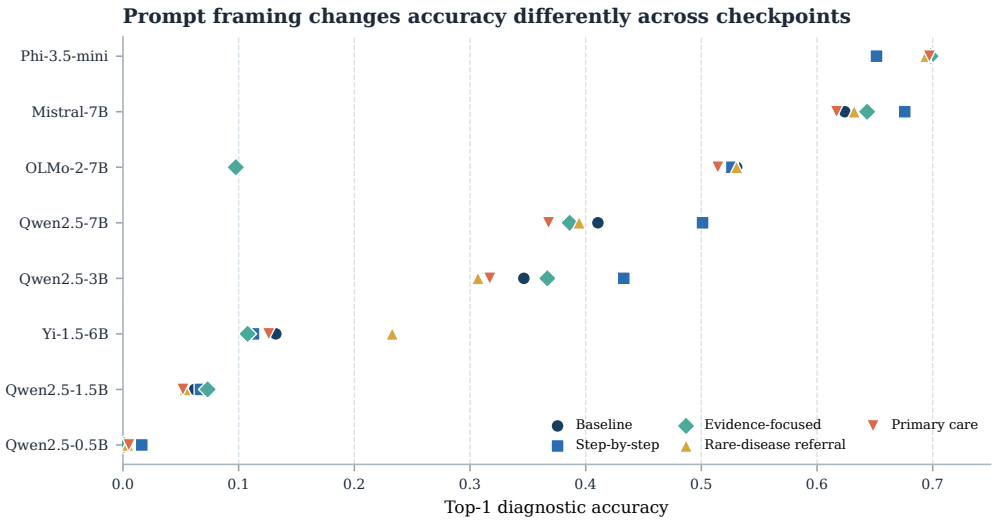

Figure 11: Prompt and decoding sensitivity across the evaluated checkpoints.

therefore not directly comparable to the no-thinking estimate because the usable case sets differ substantially.

The binary intervention tasks require less output. Their evidence effects are 0.309, 0.000, and 0.136 for no-thinking, 512-token, and 2,048-token conditions; recovery rates are 0.975, 0.983, and 0.962. Within this protocol, the tested reasoning budgets do not increase evidence responsiveness and introduce a large completion cost on the 82-way ranking task. The result is limited to this checkpoint, prompt format, and budget range.

J Prior and Evidence Coefficient Landscape

Figure 13 plots the two fitted coefficients for the open-weight panel. The plot is descriptive and is included to show how the two fitted coordinates vary across checkpoint families. The evidence coefficient has direct interventional validation; the prior coefficient is retained as the second coordinate of the rank model and does not receive the same interpretation.

K Prompt Protocols

K.1 Primary ranking prompt

The primary prompt presents positive findings, explicitly absent findings, and the fixed disease vocabulary. It asks for the three most likely diagnoses in ranked order and requests disease names only. Model-specific chat templates are used when provided by the checkpoint. No chain-of-thought explanation is requested or scored.

A representative prompt is:

> Given the clinical findings below, return the three most likely diagnoses from the supplied disease list. Use both the positive findings and the explicitly absent findings. Return a numbered list of three disease names, ordered from most to least likely.

The exact case findings and candidate list are inserted by the released code.

K.2 Randomized intervention prompt

The factorial prompt presents two diagnoses, an assigned statement about their prevalence, and findings corresponding to the assigned evidence level. The model must choose one diagnosis.

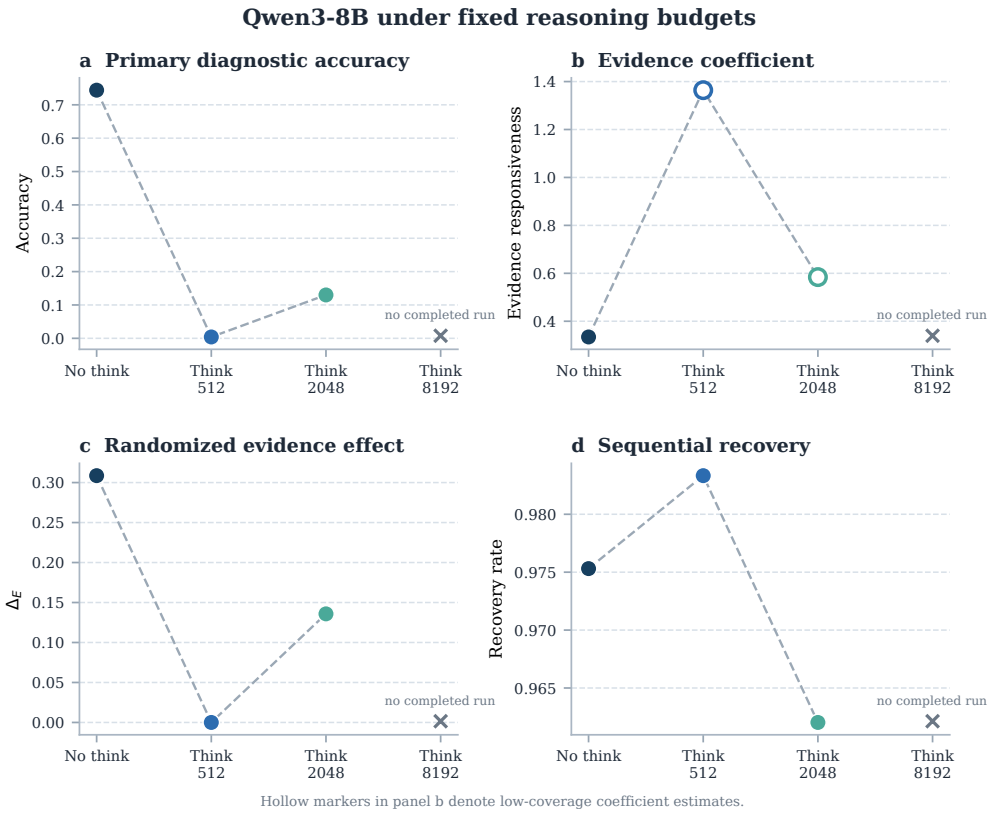

Figure 12: Qwen3-8B under fixed reasoning budgets. Baseline ranking coefficients become difficult to compare when the reasoning trace consumes the available budget before the requested ranked answer is produced.

Evidence and stated prevalence are assigned independently, so the prompt can pair strong target evidence with an unfavorable target prior.

K.3 Sequential prompt

The sequential task preserves conversational state. At step one, the model receives the first evidence block and returns a diagnosis. At step two, it receives the second block and is asked to update the diagnosis using all information seen so far. The reverse-order condition swaps the two blocks but gives the same union of evidence at the second step.

L Reproducibility

The benchmark data and analysis code are available in the supplementary appendix and will be released publicly upon acceptance.

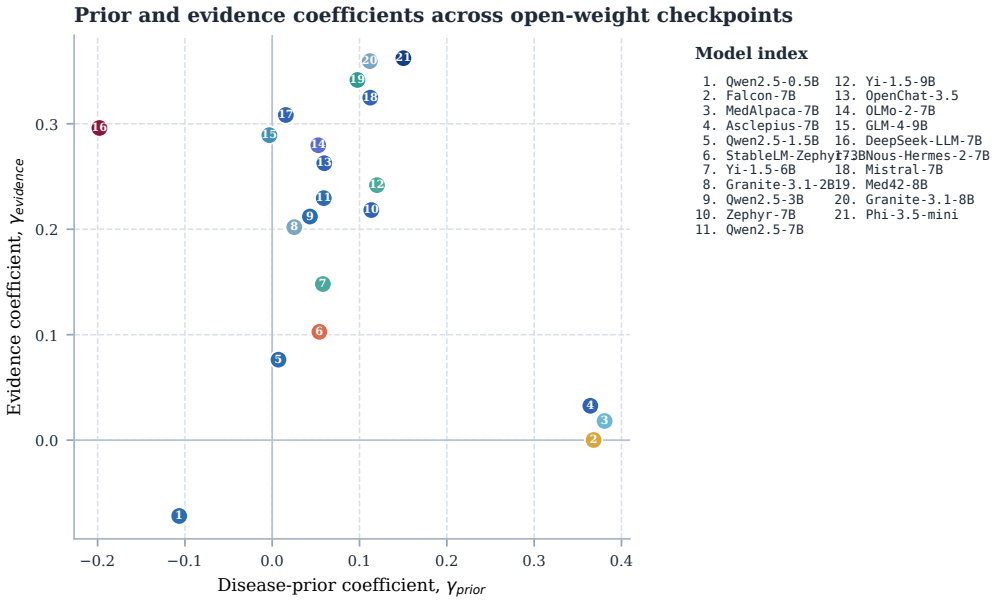

Figure 13: Prior and evidence coefficients across the open-weight model panel.