

Same Diagnosis, Different Evidence Maps: A “Rashomon Effect” in Clinical Evidence Appraisal

Anonymous Authors
Affiliation withheld for review

Abstract—Large language models (LLMs) are increasingly evaluated on their diagnostic accuracy, yet a correct diagnosis does not necessarily indicate a consistent interpretation of clinical evidence. This limitation matters because answer-level evaluations may overlook important differences in how models assess the relevance and diagnostic value of individual findings. To investigate this issue, we use Script Concordance Test–inspired ratings to construct finding-by-option evidence maps for eight LLMs on MedQA and MedMCQA, comparing model pairs that both answer the same question correctly. Despite agreeing on the diagnosis, models frequently differ in the relevance, strength, and diagnostic specificity they assign to clinical findings, particularly in how findings support competing diagnoses. Furthermore, removing findings rated as strongly supportive reduces diagnostic accuracy more than removing neutral findings, suggesting that these ratings capture meaningful diagnostic information. These results reveal a diagnostic “Rashomon effect” in which identical correct answers conceal different interpretations of clinical evidence. Our approach provides a complementary means of evaluating clinical reasoning beyond diagnostic accuracy.

Index Terms—large language models, clinical reasoning, evidence appraisal, Script Concordance Test

I. INTRODUCTION

Clinical multiple-choice question (MCQ) accuracy is widely used to assess medical capabilities of large language models (LLMs) [1], [10]. These benchmarks provide a reproducible measure of answer accuracy, but reveal little about how models interpret the clinical evidence leading to a diagnosis. Clinical LLM evaluation has therefore expanded to clinician-rated reasoning, simulated care workflows, and structured clinical cases [10], [12], [13]. Nevertheless, answer accuracy remains the predominant evaluation dimension [11].

Despite the widespread use of answer accuracy in clinical LLM evaluation, correctness alone does not reveal how a model appraises the clinical findings underlying its diagnosis. Different models can select the same diagnosis while disagreeing about the contribution of individual findings to that decision. A finding that one model views as supporting the selected diagnosis may be considered uninformative by another, or even as favoring a competing diagnosis, while the remaining evidence still leads each model to the same overall diagnosis. We refer to this as evidence-appraisal disagreement: agreement at the level of the final decision despite disagreement in how the underlying evidence is interpreted. This possibility reflects the Rashomon effect, a classic phenomenon in machine learning: models with similar predictive performance can assign different importance to the same inputs [14], [15]. Outside clinical diagnosis, models can agree on an

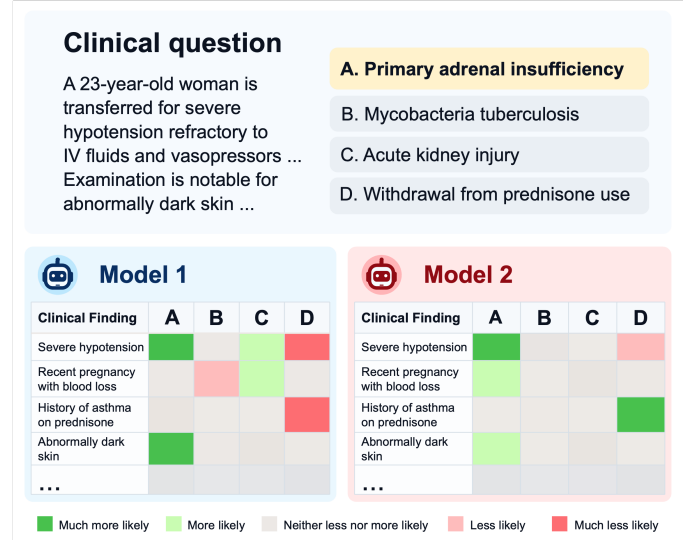


Fig. 1. Conceptual illustration of evidence-appraisal disagreement. Two models reach the same correct diagnosis (A) but assign different evidence-impact ratings to individual clinical findings across candidate diagnoses. Ratings are illustrative.

output while differing in the inputs they treat as important [16], [17]. Whether evidence-appraisal disagreement also occurs in clinical diagnosis remains unclear.

In differential diagnosis, evidence-appraisal disagreement arises when models that reach the same diagnosis differ in how they interpret individual findings in relation to competing diagnostic hypotheses. Because the final diagnosis reflects the evidence considered across the case, agreement on that diagnosis does not require agreement on the diagnostic contribution of every finding. The same finding may be interpreted as specifically supporting the correct diagnosis, supporting several plausible diagnoses, favoring a competing diagnosis, or undermining a competing hypothesis. Answer accuracy cannot distinguish among these patterns, as illustrated in Fig. 1.

We therefore represent these relations with structured evidence-impact ratings adapted from the Script Concordance Test (SCT) [18]. SCT asks how new clinical information changes the plausibility of a diagnostic hypothesis [18]–[20], and has recently been used to benchmark LLM clinical reasoning against expert responses [21]. We extend its directional response format to every eligible finding and every answer option. The resulting finding-by-option matrix forms a differential-evidence map. However, the numeric rating vector

alone does not directly identify the diagnostic function of a finding, such as whether it specifically supports the correct diagnosis, supports a competing diagnosis, or weakens one. To make these relationships interpretable, we introduce a deterministic six-role framework that assigns each finding a diagnostic role relative to the ground-truth diagnosis and competing answer options. Together, the differential-evidence map and six-role framework provide an analytic method for systematically comparing evidence appraisal across models.

We compare models only on questions that both answered correctly, with a further analysis of questions answered correctly by all eight models. This design isolates differences in evidence appraisal after controlling for agreement on the final diagnosis.

This study makes three main contributions. First, we introduce an analytic framework that combines finding-by-option differential-evidence maps with a six-role characterization of each finding’s diagnostic function. Second, we show that models reaching the same correct diagnosis can assign different relevance, strength, and diagnostic specificity to the same clinical finding. Third, we examine whether models’ self-assigned evidence-impact ratings are reflected in their diagnostic behavior through controlled evidence-removal experiments.

II. RELATED WORK

A. Clinical LLM Evaluation Beyond Accuracy

Clinical LLM evaluation has largely emphasized answer correctness. A systematic review found that 95.4% evaluated accuracy, whereas calibration, uncertainty, and deployment considerations were rarely assessed [11]. Several studies have broadened this scope. Singhal et al. supplemented medical question-answering benchmarks with clinician assessments of factuality, reasoning, potential harm, and bias [10]. Hager et al. evaluated models in simulated clinical workflows, identifying limitations in information gathering, laboratory interpretation, and guideline adherence that examination performance did not capture [12]. More recently, MedR-Bench assessed written reasoning for factual accuracy, completeness, and efficiency across examination recommendation, diagnosis, and treatment planning [13]. These approaches extend evaluation from selecting a correct answer to assessing how models interpret and use clinical information.

B. Cross-Model Disagreement in Evidence Attribution

Agreement on a prediction does not necessarily imply agreement on its explanation. Research on the Rashomon effect examines how similarly performing models can yield different feature attributions. Laberge et al. addressed this problem by identifying attribution relationships that remain consistent across models in a Rashomon set [15]. In language modeling, Bogaert et al. compared token-level attributions across fine-tuned CamemBERT instances in news-versus-opinion classification, showing that agreement on classification outputs does not ensure agreement on explanatory tokens [16]. They explicitly connected this variability to the Rashomon effect. Kulmizev et al. examined label and explanation variation

within and across LLMs on natural language inference, finding model-specific judgment patterns even when model ensembles approximated human label distributions [17]. These studies distinguish agreement on outcomes from agreement on their justification. In clinical diagnosis, this distinction raises the question of whether models reaching the same correct answer assign similar significance to individual findings, particularly their ability to distinguish competing diagnoses.

C. Script Concordance and Diagnostic Evidence Appraisal

SCT assesses clinical reasoning under uncertainty by examining how additional information changes the plausibility of a diagnostic hypothesis [18]. Each item links a clinical scenario, a proposed hypothesis, and a new finding; respondents indicate the direction and strength of the finding’s effect on an ordered scale [19]. Conventional scoring assigns credit according to an expert panel’s response distribution, accommodating variation in expert judgments rather than requiring unanimous agreement [20]. SCT thus targets a specific component of clinical reasoning: interpreting evidence in relation to a hypothesis. McCoy et al. applied SCT to evaluate how LLMs update diagnostic and therapeutic judgments in response to new clinical information, using expert-panel responses as the scoring reference [21]. This literature provides a clinical foundation for assessing evidence–hypothesis relationships separately from final diagnostic accuracy.

III. METHODS

A. Study Design

We investigate three research questions:

RQ1: Evidence-map concordance. When LLMs converge on the same correct diagnosis, to what extent do they construct concordant differential-evidence maps across candidate diagnostic hypotheses?

RQ2: Evidence-disagreement patterns. When those maps disagree, how do their implications differ for the correct diagnosis and competing diagnostic hypotheses?

RQ3: Evidence-removal sensitivity. Are models’ self-assigned evidence-impact ratings associated with how their diagnostic decisions change when the corresponding evidence is removed?

We evaluated each diagnostic MCQ using eight models: four closed-source models, three general-purpose open-weight models, and one medically specialized open-weight model. Each model first selected one option using a standard MCQ prompt, then provided a differential-evidence map. The dataset-labeled correct answer is termed the gold option, and the remaining answer choices are termed competing options.

Each question stem is divided into an initial clinical scenario, intermediate sentences, and a final question instruction. Each intermediate sentence is treated as one clinical finding for analysis and is termed an eligible finding. Only eligible findings receive evidence-impact ratings; the initial scenario and final question instruction remain unrated.

To construct these maps, we adapted SCT [18], [20]. We treated each answer option as a candidate diagnosis and each

eligible finding as new clinical information. For every finding-option pair, the model rated the change in diagnostic likelihood on a five-point scale from -2 to $+2$. These ratings formed a numeric finding-by-option matrix, termed a differential-evidence map.

For **evidence-map concordance**, we compare maps only on questions that both models in a pair answered correctly, with an additional analysis restricted to questions answered correctly by all eight models. Finding-level concordance is measured at three comparison levels: gold-option rating agreement, six-role agreement, and full-vector agreement. **Evidence-disagreement patterns** are classified separately at each comparison level. For **evidence-removal sensitivity**, eligible findings are removed according to each model’s own ratings, and changes in diagnostic choice are measured. We additionally illustrate changes in normalized answer-option probabilities using an open-weight model case. All analyses were conducted separately on MedQA [1] and MedMCQA [2].

B. Datasets and Models

We derived candidate cases from the MedQA [1] and MedMCQA [2] subsets of MINT [22]. We first retained four-option questions that met our definition of diagnostic MCQs and contained at least three eligible findings in the stem. From the resulting eligible pools, we then randomly sampled 200 MedQA and 122 MedMCQA questions for analysis. MedQA comprises United States Medical Licensing Examination (USMLE)-style clinical vignettes. MedMCQA comprises multi-subject questions from Indian medical entrance examinations. The number of eligible findings per question ranged from 3 to 18 (median 7) in MedQA and from 3 to 8 (median 4) in MedMCQA.

The four closed models were GPT-4.1 [3], GPT-5.5 [4], Claude Sonnet 5 [5], and Claude Opus 4.8 [6]. The four open-weight models were Qwen3-8B, Qwen3-32B [7], Gemma-3-27B-IT [8], and the medically specialized MedGemma-27B-text-IT [9].

C. Constructing Differential-Evidence Maps

We construct each map from SCT-style ratings of how one new finding changes the likelihood of a diagnostic hypothesis [18], [20]. In a conventional SCT item, a respondent first considers a diagnostic hypothesis and then observes one additional finding. The respondent rates the direction and magnitude of the finding’s effect on the plausibility of the diagnostic hypothesis. Each answer option serves as a diagnostic hypothesis, and each eligible finding serves as the additional clinical information.

For each eligible finding i and candidate option j , the model receives the initial clinical scenario, finding i , and option j framed as the diagnostic hypothesis under consideration. It is asked: *Given the initial clinical scenario, how does this finding affect the plausibility of the diagnostic hypothesis?* The model assigns

$$s_{ij} \in \{-2, -1, 0, +1, +2\},$$

with the following anchors:

- -2 : much less likely; strong evidence against the hypothesis;
- -1 : less likely; modest evidence against the hypothesis;
- 0 : neither less nor more likely; no meaningful impact;
- $+1$: more likely; moderate support for the hypothesis;
- $+2$: much more likely; strong support for the hypothesis.

For a question with m eligible findings and k candidate options, the differential-evidence map is

$$S \in \{-2, -1, 0, +1, +2\}^{m \times k}.$$

Each eligible finding receives one rating for every candidate option.

D. Six-Role Framework for Finding-Level Evidence Appraisal

The differential-evidence map preserves the option-level effects of each finding, but the numeric vector itself does not directly identify the finding’s diagnostic function. We therefore map each vector to one of six mutually exclusive roles.

Three roles capture **positive support**: correct-option-specific (COS), when a finding preferentially supports the correct diagnosis; distractor-specific (DS), when it preferentially supports a competing diagnosis; and shared supportive (Shared), when it provides equally strong positive support for the correct diagnosis and the highest-rated competing diagnosis. Two roles capture **negative evidence**: correct-undermining (CU), when a finding weakens the correct diagnosis, and distractor-undermining (DU), when it weakens a competing diagnosis. Findings that do not exhibit any of these preferential relationships are classified as non-discriminative (ND).

These roles distinguish different mechanisms through which a finding can favor one diagnostic hypothesis over another. For example, COS and DU can both favor the correct diagnosis, but COS does so by directly supporting it, whereas DU does so by weakening a competing diagnosis. Similarly, DS and CU can both favor a competing diagnosis, either by supporting the competing diagnosis or by weakening the correct diagnosis.

Because a single rating vector may satisfy more than one of these relations, we apply the role-assignment rules in a fixed order to obtain a unique label. Positive-support roles are evaluated first, followed by undermining roles, with ND assigned otherwise. Thus, a finding that supports the correct diagnosis more strongly than any competing diagnosis is labeled COS even if it also weakens a competing diagnosis.

Formally, each rule is defined using three ratings from the finding’s option-level vector: s_* for the gold option, $s_{\max o}$ for the highest-rated competing option, and $s_{\min o}$ for the lowest-rated competing option. Table I lists the assignment rules in priority order; the first satisfied rule determines the role. Each finding therefore receives exactly one role.

E. Cross-Model Evidence-Map Concordance (RQ1)

Among the eight models, we considered all 28 possible model pairs. For each pair, we compared their differential-evidence maps only on questions that both models answered correctly. We grouped the pairs as closed-closed, open-open,

TABLE I
SIX-ROLE PARTITION OF MODEL-IMPLIED FINDING-LEVEL EVIDENCE. HERE, s_* IS THE GOLD-OPTION RATING, AND $s_{\max o}$ AND $s_{\min o}$ ARE THE HIGHEST AND LOWEST COMPETING-OPTION RATINGS.

Role	Clinical interpretation	Assignment rule
Correct-option-specific (COS)	Positive evidence supports the correct diagnosis more strongly than any competing diagnosis	$s_* \geq 1 \wedge s_* - s_{\max o} \geq 1$
Distractor-specific (DS)	Positive support preferentially identifies a competing diagnosis	$s_{\max o} \geq 1 \wedge s_{\max o} - s_* \geq 1$
Shared supportive (Shared)	Positive evidence supports the correct diagnosis and the highest-rated competing diagnosis equally	$s_* = s_{\max o} \geq 1$
Correct-undermining (CU)	Negative evidence weakens the correct diagnosis relative to the highest-rated competing diagnosis	$s_* \leq -1 \wedge s_* < s_{\max o}$
Distractor-undermining (DU)	Negative evidence weakens at least one competing diagnosis more than the correct diagnosis	$s_{\min o} \leq -1 \wedge s_* > s_{\min o}$
Non-discriminative (ND)	No preferential relation is captured by the preceding roles	none of the above

or closed-open according to whether each model was closed or open-weight.

The unit of comparison is one eligible finding. Each model assigns a k -dimensional rating vector to that finding, with one rating for each candidate option. We report exact-agreement rates at three comparison levels. Each rate is the proportion of eligible findings, pooled across questions answered correctly by both models, for which the two models agree at that comparison level.

Gold-option rating agreement requires the two models to assign the same rating s_* to the gold option. This comparison assesses whether they agree on how the finding changes the plausibility of the correct diagnosis, without considering ratings for competing options.

Six-role agreement requires the two models to assign the same role from Table I. The role depends on s_* and on its contrast with the highest- and lowest-rated competing options. Two models can therefore assign the same gold-option rating to a finding but assign it different roles, for example COS versus Shared. The same role does not require the same gold-option rating: ratings of +2 and +1 can both yield COS.

Full-vector agreement requires the two models to assign identical ratings to every option for a given finding. This is the strictest agreement criterion. Two models can share a gold-option rating and even a role while differing in a rating for a competing option.

On questions answered correctly by all eight models, we also report unanimous agreement at each comparison level. A finding counts as unanimous only when all eight models match at that comparison level.

The same three agreement measures were also calculated at the question level. A question counted as a match only when every eligible finding matched at the corresponding comparison level.

F. Taxonomy of Evidence Disagreement (RQ2)

When two models disagreed on an eligible finding, we classified the disagreement for the gold-option rating and the full vector, and then for the finding’s role. We reused the eligible findings from questions that both models answered correctly. For a given finding, models may disagree at more than one comparison level, or agree at one level and disagree at

another. Table II defines the classes. Fixed role-disagreement labels use “gold” for the correct diagnosis and “distractor” or “competitor” for competing diagnoses. In the labels ending in “versus none,” “none” denotes ND, not necessarily zero ratings or an absence of evidence.

When the gold-option ratings differed, the disagreement was classified as magnitude-only change, neutral versus nonzero, or positive versus negative. These three classes cover every case in which the gold-option ratings differ.

Full-vector mismatches were divided into two classes: distractor-column only, where the gold-option ratings agree but at least one competing-option rating differs; and gold column differs, where the gold-option ratings differ, regardless of whether competing-option ratings also differ. Here, “distractor columns” denote the competing-option columns, and “gold column” denotes the gold-option column.

The role is the diagnostic direction read from the gold-option rating together with the highest and lowest competing-option ratings. It does not keep every other cell in the vector. The six roles in Table I yield 15 unordered transitions, and each transition was assigned to exactly one named class. Two models can therefore share a gold-option rating and still disagree in role. Sign conflicts in gold-option ratings can occur across different role-disagreement classes.

G. Self-Assigned Evidence-Impact Ratings and Ablation Effects (RQ3)

To examine whether models’ self-assigned evidence-impact ratings were associated with diagnostic changes following evidence removal, we performed self-ablation on questions each model initially answered correctly. Using each model’s gold-option ratings, we defined four deletion conditions: removing eligible findings rated 0, rated +1, rated +2, or rated +1 and +2 together. The initial clinical scenario and final question instruction were retained in all conditions. Accuracy was calculated only for questions with at least one eligible finding removed. We compared post-ablation accuracy across conditions to determine whether removing evidence rated as supportive disrupted diagnostic decisions more than removing evidence rated as neutral. For each model, dataset, and deletion condition, we calculated 95% percentile intervals from 2,000 question-level bootstrap resamples.

TABLE II

DISAGREEMENT TAXONOMY FOR THE GOLD-OPTION RATING, THE FULL VECTOR, AND THE FINDING’S ROLE, SCORED SEPARATELY. THE ROLE IS READ FROM THE GOLD-OPTION RATING TOGETHER WITH THE HIGHEST AND LOWEST COMPETING-OPTION RATINGS. GOLD, VECTOR, AND ROLE DENOTE THE GOLD-OPTION RATING, FULL VECTOR, AND SIX-ROLE ASSIGNMENT, RESPECTIVELY. ROLE TRANSITIONS ARE UNORDERED AND APPEAR IN ONLY ONE ROLE CLASS.

Level	Class	Interpretation	Rule or transitions
Gold	Magnitude-only change	Both models assign ratings with the same nonzero sign to the gold option but disagree on magnitude.	Same sign, different nonzero gold-option rating
Gold	Neutral versus nonzero	One model assigns a nonzero gold-option rating; the other assigns zero, indicating no meaningful impact on the correct diagnosis.	One gold-option rating is 0 and the other is nonzero
Gold	Positive versus negative	The models assign opposite implications to the same finding for the correct diagnosis.	Opposite signs in the gold-option ratings
Vector	Distractor-column only	The models agree on the gold-option rating but differ on at least one competing option.	Same gold-option rating; at least one other option rating differs
Vector	Gold column differs	The gold-option ratings differ. Remaining option ratings may or may not also differ.	The gold-option ratings differ
Role	Specificity of gold support	Both models assign positive support to the correct diagnosis, but disagree on whether it exceeds or equals the strongest competing support.	COS↔Shared
Role	Preferred diagnosis	Both models treat the finding as positive evidence, but they disagree on which diagnosis, if any, it preferentially identifies.	COS↔DS, Shared↔DS
Role	Positive evidence versus none	One model assigns a supportive role; the other treats the finding as non-discriminative.	COS↔ND, Shared↔ND, DS↔ND
Role	Affirmation versus elimination	One model interprets the finding as supporting the correct diagnosis; the other interprets it as weakening a competing diagnosis. Both readings can favor the correct diagnosis.	COS↔DU, Shared↔DU
Role	Gold support versus undermining	One model reads the finding as supporting the correct diagnosis; the other reads it as weakening that diagnosis.	COS↔CU, Shared↔CU
Role	Undermining target	Both models treat the finding as negative evidence, but they disagree on whether it primarily weakens the correct diagnosis or a competing diagnosis.	CU↔DU
Role	Undermining versus none	One model treats the finding as undermining a diagnosis; the other treats it as non-discriminative.	CU↔ND, DU↔ND
Role	Competitor support versus gold undermining	One model interprets the finding as supporting a competing diagnosis; the other interprets it as weakening the correct diagnosis. Both readings can favor a competing diagnosis.	DS↔CU
Role	Distractor support versus undermining	One model interprets the finding as supporting a competing diagnosis; the other interprets it as weakening a competing diagnosis.	DS↔DU

We illustrate this effect with an open-weight model case in which the selected answer remained unchanged. For this example, we normalized the log-likelihoods of the four option letters at the answer position using a softmax transformation, applying the same scoring procedure before and after ablation:

$$p_k(x) = \frac{\exp(\ell_k(x))}{\sum_{j \in \{A, B, C, D\}} \exp(\ell_j(x))}, \quad k \in \{A, B, C, D\}. \quad (1)$$

Here, x denotes the original or ablated input, and $\ell_k(x)$ is the log-likelihood of option letter k at the answer position. These probabilities quantify relative preference among the four options; reported percentages equal $100p_k(x)$.

IV. RESULTS AND DISCUSSION

The section begins with diagnostic accuracy to establish the jointly correct cases used for comparison. It then examines evidence-map concordance and evidence-disagreement patterns, followed by an illustrative case study. The final analysis evaluates evidence-removal sensitivity to assess whether models’ self-assigned evidence-impact ratings are reflected in their diagnostic behavior.

A. Diagnostic Accuracy

All eight models achieved high diagnostic accuracy on both MedQA and MedMCQA. Table III(a) reports per-model

accuracy on the MedQA and MedMCQA questions, ranging from 76.0% to 98.0% on MedQA and from 86.1% to 96.7% on MedMCQA. Table III(b) lists the both-correct sample for each of the 28 pairs in Fig. 2. Those jointly correct sets contained 137–191 MedQA questions and 99–117 MedMCQA questions. All eight models were jointly correct on 130 MedQA questions and 95 MedMCQA questions. Later sections compare evidence maps only on these jointly correct questions.

B. Concordance of Differential-Evidence Maps

We next compared differential-evidence maps on the jointly correct questions. Finding-level concordance was measured at three comparison levels: gold-option rating agreement, six-role agreement, and full-vector agreement. Fig. 2 reports these rates for the 28 pairs, grouped as closed-closed, open-open, or closed-open.

Gold-option rating agreement was the highest of the three, ranging from 0.53 to 0.82. Six-role agreement was lower, indicating that a shared gold-option rating did not necessarily imply the same diagnostic function. Full-vector agreement was the lowest of the three agreement measures in every pair class and did not exceed 0.44, as competing-option ratings often differed even when the gold-option rating and role matched.

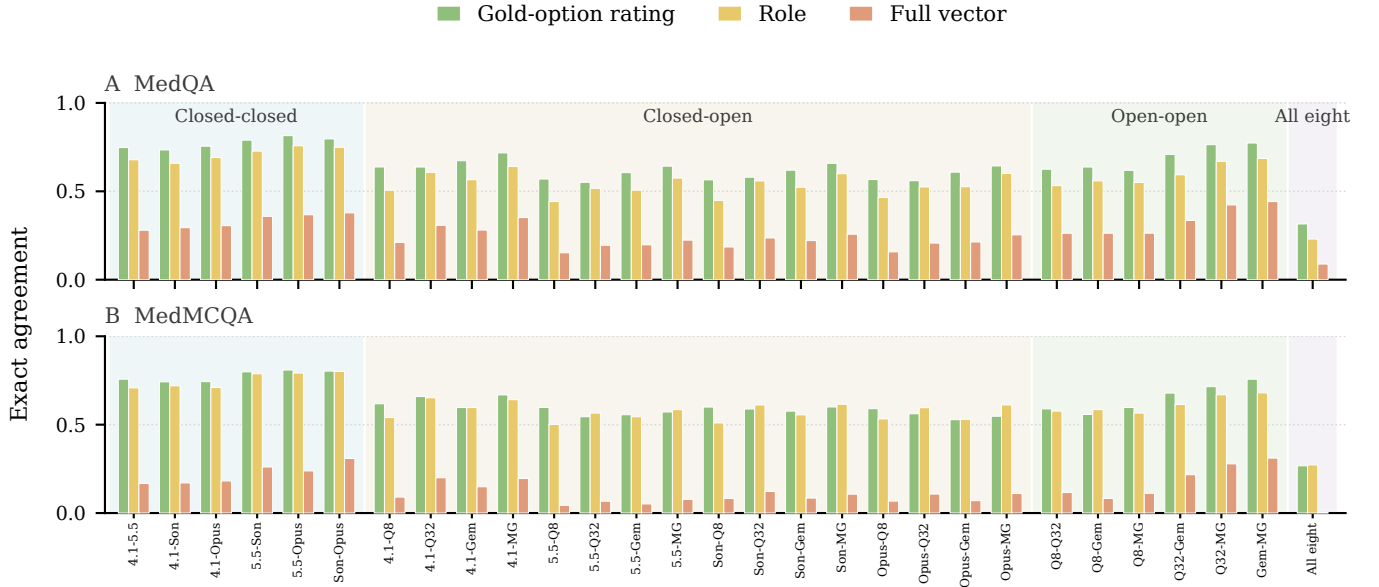


Fig. 2. Finding-level exact agreement at three comparison levels. Gold-option rating agreement: the two models assigned the same -2 to $+2$ rating to the gold option. Six-role agreement (labeled Role in the figure): they assigned the same role from Table I to the finding. Full-vector agreement: they assigned identical ratings to every answer option. The 28 pairwise clusters use questions both models answered correctly; the final cluster uses the questions answered correctly by all eight models (MedQA, $n = 130$; MedMCQA, $n = 95$). Shading distinguishes closed-closed, closed-open, open-open, and all-eight groups. Tick labels: 4.1, GPT-4.1; 5.5, GPT-5.5; Son, Claude Sonnet 5; Opus, Claude Opus 4.8; Q8, Qwen3-8B; Q32, Qwen3-32B; Gem, Gemma-3-27B; MG, MedGemma-27B. (A) MedQA. (B) MedMCQA.

Closed-closed pairs were the most concordant on gold-option ratings and roles. Closed-open pairs were the least concordant. Open-open pairs were intermediate in gold-option rating agreement and six-role agreement. Full-vector agreement remained low in every pair class.

On the 130 MedQA and 95 MedMCQA questions answered correctly by all eight models, a unanimous full vector occurred for only 0.088 and 0.003 of eligible findings. Unanimous gold-option ratings (0.315 and 0.267) and unanimous roles (0.230 and 0.272) were higher, but remained far from complete.

Whole-question agreement was much lower at all three comparison levels. Whole-question gold-option rating agreement ranged from 0.05 to 0.25 on MedQA and from 0.11 to 0.48 on MedMCQA. Whole-question six-role agreement was lower. A fully identical evidence map occurred for at most 0.02 of questions. A shared correct diagnosis therefore did not imply a shared differential-evidence map.

C. Sources of Evidence-Map Disagreement

Agreement at one comparison level did not imply agreement at the others. Two models could assign the same gold-option rating to a finding but different ratings to a competing option. We therefore typed mismatches in the gold-option rating and the full vector before turning to the role (Table II). Fig. 3 shows how those gold-option rating and full-vector mismatches break down for every pair. The role is the diagnostic direction read from the gold-option rating together with the highest and lowest competing-option ratings.

Most gold-option rating mismatches reflected differences in relevance or magnitude rather than direction. Across both datasets and all model pairs in Fig. 3, Neutral versus nonzero and Magnitude-only change accounted for the majority of mismatches, whereas Positive versus negative was consistently the least common category. In other words, models more often disagreed on whether a finding was informative, or on how strongly it supported the correct diagnosis, than on whether it supported or opposed that diagnosis. Although rare, Positive versus negative mismatches represent a qualitatively different failure mode because the two models assign opposite directional implications to the same finding.

Most full-vector mismatches preserved agreement on the gold-option rating and arose only from differences in the competing-option columns. Distractor-column-only mismatches were the majority for every closed-closed pair. Among closed-open pairs, however, differences in gold-option ratings became more prominent and, in some cases, exceeded distractor-column-only mismatches.

Role disagreements were concentrated within the positive-evidence branch of the taxonomy (Table II), with pairwise rates reported in Table IV. The dominant class, Specificity of gold support, reflects disagreement over whether positive support is stronger for the correct diagnosis than for every competing diagnosis or equal to the strongest competing support (COS $\leftrightarrow$ Shared). Other common disagreements similarly concerned which diagnosis positive evidence favored or whether the finding expressed a preferential diagnostic

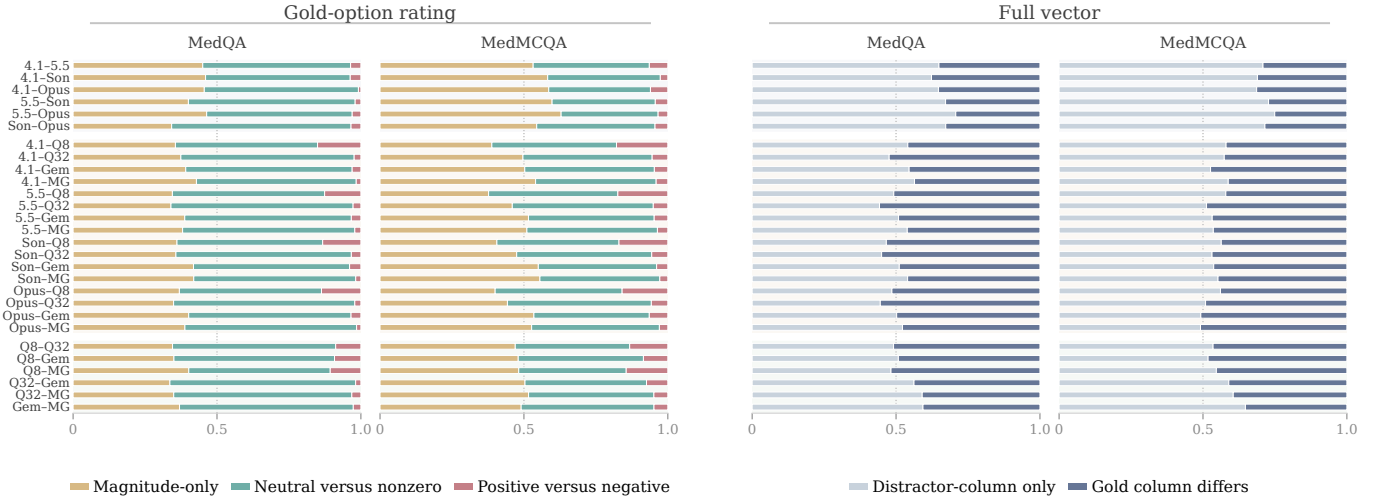


Fig. 3. Composition of gold-option rating and full-vector mismatches for the 28 pairs in Fig. 2 order. Gold-option rating mismatches are divided into Magnitude-only (magnitude-only change), Neutral versus nonzero, and Positive versus negative; full-vector mismatches are divided into Distractor-column only and Gold column differs. Each cell is a 100% stack over eligible findings on which the models disagreed at that comparison level. The left and right panels show MedQA and MedMCQA, respectively, within each comparison level.

TABLE III

DIAGNOSTIC ACCURACY AND BOTH-CORRECT SAMPLE SIZES.

(A) PER-MODEL ACCURACY, WITH THE NUMBER OF CORRECT QUESTIONS ON MEDQA ($n = 200$) AND MEDMCQA ($n = 122$). (B) QUESTIONS BOTH MODELS ANSWERED CORRECTLY, IN THE SAME ORDER AS FIG. 2. PAIR LABELS USE 4.1 FOR GPT-4.1, 5.5 FOR GPT-5.5, SON FOR CLAUDE SONNET 5, OPUS FOR CLAUDE OPUS 4.8, Q8 FOR QWEN3-8B, Q32 FOR QWEN3-32B, GEM FOR GEMMA-3-27B, AND MG FOR MEDGEMMA-27B.

(a)					
	MedQA		MedMCQA		
GPT-4.1	96.5%	(193)	95.1%	(116)	
GPT-5.5	98.0%	(196)	95.9%	(117)	
Claude Sonnet 5	96.5%	(193)	94.3%	(115)	
Claude Opus 4.8	97.5%	(195)	96.7%	(118)	
Qwen3-8B	76.0%	(152)	86.1%	(105)	
Qwen3-32B	89.0%	(178)	90.2%	(110)	
Gemma-3-27B	81.5%	(163)	89.3%	(109)	
MedGemma-27B	89.5%	(179)	93.4%	(114)	

(b)					
Pair	MedQA	MedMCQA	Pair	MedQA	MedMCQA
Closed-closed					
4.1-5.5	191	114	4.1-Son	188	113
4.1-Opus	190	115	5.5-Son	191	115
5.5-Opus	191	117	Son-Opus	191	115
Closed-open					
4.1-Q8	150	104	4.1-Q32	175	109
4.1-Gem	161	106	4.1-MG	175	113
5.5-Q8	151	104	5.5-Q32	178	109
5.5-Gem	163	106	5.5-MG	177	114
Son-Q8	149	103	Son-Q32	175	109
Son-Gem	161	105	Son-MG	175	113
Opus-Q8	150	104	Opus-Q32	175	110
Opus-Gem	160	106	Opus-MG	177	114
Open-open					
Q8-Q32	147	101	Q8-Gem	137	99
Q8-MG	148	103	Q32-Gem	154	104
Q32-MG	166	109	Gem-MG	157	106

All eight jointly correct: MedQA 130, MedMCQA 95.

relationship captured by the role framework, whereas disagreements involving undermining roles were consistently rare. Thus, models often agreed on the broad evidential orientation of a finding but still differed on its diagnostic specificity and target. This distinction is meaningful for differential diagnosis: evidence that generally supports several plausible diagnoses plays a different role from evidence that selectively favors the correct diagnosis over competing diagnoses.

D. Case Study

Fig. 4 uses a MedQA case to illustrate how models can reach the same correct answer while constructing different evidence maps. Although all eight models selected the gold option, they disagree on the highlighted findings in the specificity of support, the ratings assigned to competing options, and the sign of the gold-option rating.

For the finding in S3, all models assign a positive rating to the gold option but assign different roles. Some models assign stronger support to primary adrenal insufficiency than to any competing diagnosis, yielding COS. Others assign equally strong positive ratings to the gold option and a competing option, yielding Shared.

For the finding in S6, all models assign a positive gold-option rating, yet their full rating vectors differ. Even among models assigning the same gold-option rating of +2, ratings for tuberculosis and steroid withdrawal differ. Qwen3-8B assigns Shared, whereas the other models assign COS; full-vector disagreement also remains among models sharing the COS label.

The disagreement becomes more fundamental for the finding in S8, where models differ in the sign of the gold-option rating. Successful lactation is judged to support primary

TABLE IV

SIX-ROLE DISAGREEMENT FOR THE 28 PAIRS. ROWS ARE CLASSES; COLUMNS ARE PAIRS IN FIG. 2 ORDER. (A) MEDQA. (B) MEDMCQA. EACH CELL IS THE NUMBER OF FINDINGS IN THAT ROLE-DISAGREEMENT CLASS DIVIDED BY ALL ELIGIBLE FINDINGS IN THE PAIR'S BOTH-CORRECT QUESTIONS. CLASSES ARE EXCLUSIVE AND THEREFORE SUM TO THE OVERALL ROLE-DISAGREEMENT RATE, NOT TO 1. BOLD MARKS THE HIGHEST CLASS FOR THAT PAIR; AN UNDERLINE MARKS THE SECOND-HIGHEST.

(a) MedQA																													
Class	Closed-closed						Closed-open												Open-open										
	4.1-5.5	4.1-Son	4.1-Opus	5.5-Son	5.5-Opus	Son-Opus	4.1-Q8	4.1-Q32	4.1-Gem	4.1-MG	5.5-Q8	5.5-Q32	5.5-Gem	5.5-MG	Son-Q8	Son-Q32	Son-Gem	Son-MG	Opus-Q8	Opus-Q32	Opus-Gem	Opus-MG	Q8-Q32	Q8-Gem	Q8-MG	Q32-Gem	Q32-MG	Gem-MG	
Specificity of gold support	.113	.129	.114	.106	.103	.081	.140	.107	.144	.112	.158	.114	.159	.128	.153	.108	.156	.118	.149	.116	.156	.114	.114	.127	.133	.116	.069	.067	.062
Preferred diagnosis	.062	.063	.059	.039	.038	.037	.141	.079	.102	.061	.145	.101	.110	.062	.149	.094	.100	.060	.152	.099	.122	.062	.118	.127	.117	.069	.067	.062	
Positive evidence versus none	.076	.070	.071	.058	.048	.053	.111	.117	.109	.111	.133	.145	.122	.139	.126	.132	.111	.118	.129	.145	.116	.144	.130	.108	.122	.122	.096	.108	
Affirmation versus elimination	.022	.028	.031	.032	.018	.038	.028	.042	.017	.019	.028	.049	.023	.025	.032	.037	.024	.018	.033	.062	.024	.028	.022	.021	.018	.019	.007	.008	
Gold support versus undermining	.003	.004	.001	.001	.002	.002	.013	.005	.003	.004	.013	.006	.007	.006	.014	.008	.006	.006	.014	.006	.006	.005	.006	.009	.008	.004	.005	.004	
Undermining target	.001	.001	.002	.001	.003	.001	.004	.002	.004	.002	.004	.003	.002	.001	.003	.002	.003	.001	.002	.002	.002	.000	.008	.002	.003	.006	.005	.002	
Undermining versus none	.037	.037	.024	.031	.023	.033	.052	.038	.043	.046	.067	.058	.057	.057	.062	.051	.054	.070	.047	.042	.036	.040	.061	.028	.042	.044	.062	.021	
Competitor support versus gold undermining	.002	.001	.001	.001	.001	.001	.002	.002	.000	.000	.002	.000	.002	.002	.001	.000	.002	.002	.000	.000	.002	.001	.003	.004	.000	.001	.001	.000	
Distractor support versus undermining	.006	.008	.007	.004	.006	.004	.004	.002	.013	.005	.007	.008	.011	.006	.011	.009	.021	.008	.008	.004	.009	.005	.006	.016	.006	.026	.004	.013	
(b) MedMCQA																													
Class	Closed-closed						Closed-open												Open-open										
	4.1-5.5	4.1-Son	4.1-Opus	5.5-Son	5.5-Opus	Son-Opus	4.1-Q8	4.1-Q32	4.1-Gem	4.1-MG	5.5-Q8	5.5-Q32	5.5-Gem	5.5-MG	Son-Q8	Son-Q32	Son-Gem	Son-MG	Opus-Q8	Opus-Q32	Opus-Gem	Opus-MG	Q8-Q32	Q8-Gem	Q8-MG	Q32-Gem	Q32-MG	Gem-MG	
Specificity of gold support	.131	.128	.131	.107	.105	.081	.168	.128	.146	.126	.166	.131	.178	.143	.170	.117	.172	.133	.161	.121	.180	.135	.136	.128	.146	.140	.122	.126	
Preferred diagnosis	.056	.072	.051	.028	.032	.037	.122	.081	.111	.081	.149	.122	.131	.095	.150	.108	.136	.096	.140	.094	.150	.084	.131	.143	.133	.086	.076	.082	
Positive evidence versus none	.049	.027	.040	.031	.030	.028	.056	.049	.064	.056	.070	.074	.070	.093	.049	.060	.052	.071	.060	.071	.065	.086	.042	.060	.058	.074	.055	.058	
Affirmation versus elimination	.027	.029	.040	.022	.028	.031	.058	.049	.040	.040	.060	.071	.035	.042	.053	.057	.038	.031	.063	.080	.040	.051	.035	.033	.024	.021	.021	.007	
Gold support versus undermining	.007	.002	.011	.004	.000	.007	.017	.012	.007	.009	.012	.014	.009	.009	.015	.014	.005	.009	.014	.018	.009	.009	.025	.005	.015	.014	.007	.005	
Undermining target	.002	.004	.000	.002	.000	.002	.005	.005	.000	.004	.002	.005	.000	.002	.000	.005	.000	.007	.000	.000	.000	.004	.005	.003	.000	.000	.011	.005	
Undermining versus none	.011	.009	.011	.007	.009	.013	.022	.023	.014	.036	.024	.011	.021	.024	.027	.016	.014	.024	.024	.018	.019	.018	.035	.020	.044	.026	.032	.019	
Competitor support versus gold undermining	.000	.000	.000	.000	.002	.000	.000	.000	.002	.004	.000	.002	.000	.000	.002	.000	.002	.002	.002	.000	.002	.000	.005	.003	.000	.005	.005	.002	
Distractor support versus undermining	.009	.009	.007	.011	.002	.000	.012	.002	.019	.002	.014	.005	.012	.007	.024	.011	.026	.011	.002	.002	.005	.002	.010	.022	.015	.019	.002	.016	

adrenal insufficiency, oppose it, or have no meaningful impact on its plausibility. The resulting vectors and roles also diverge, with one model assigning DS, indicating stronger support for a competing diagnosis. This pattern illustrates the less common positive-versus-negative conflict in Fig. 3: the same finding can be interpreted as supporting or weakening the correct diagnosis.

This example does not adjudicate the medically correct evidence map. It demonstrates that a shared correct answer can conceal disagreement ranging from the specificity of otherwise consistent support to a direct reversal in evidential direction.

E. Diagnostic Changes Following Self-Ablation

Self-ablation revealed greater diagnostic sensitivity to evidence rated as strongly supportive than to evidence rated as neutral or moderately supportive (Fig. 5). Among questions each model initially answered correctly, accuracy remained at 94%–100% after removing findings rated 0 and 87%–100% after removing those rated +1, but declined to 63%–88% after removing findings rated +2 and 59%–78% after removing findings rated +1 or +2. This pattern suggests that models' support ratings captured differences in the importance of evidence to their diagnostic decisions.

An unchanged answer does not necessarily indicate that a model is unaffected by evidence removal. In the MedMCQA example shown in Fig. 6, Qwen3-32B retained the gold option after the two findings with gold-option ratings of +1 and +2 were removed, but the normalized probability assigned to

that option fell from approximately 100% to 79.3%, while that of the competing option with the highest normalized probability rose to 20.7%. Evidence removal thus weakened the model's preference for the gold option without changing its final selection, an effect not captured by answer-level accuracy.

V. LIMITATIONS

Several limitations should be considered when interpreting these findings. First, each question was answered once and each evidence map was elicited from a single model run. Because both answer selection and evidence ratings may vary across repeated runs, the observed disagreements reflect one realization of model behavior and do not establish whether these differences are stable across repeated runs.

Second, the evidence ratings and derived roles were not independently validated by clinicians. The benchmark answer keys identify the correct option but do not establish how individual findings should affect each candidate diagnosis. Our analysis therefore characterizes where models differ in their appraisal of evidence, but cannot determine which appraisal is clinically preferable.

Finally, both MedQA and MedMCQA are public benchmarks on which the evaluated models achieved high accuracy. Some questions may have appeared in model pretraining or related training data. Such exposure could make agreement on the correct answer less informative about how models would appraise previously unseen clinical cases and may

Gold answer: A • All eight models selected A

A 23-year-old woman from Texas is transferred to the intensive care unit after delivering a child at 40 weeks gestation. The pregnancy was not complicated, and there was some blood loss during the delivery. [S3] The patient was transferred for severe hypotension refractory to IV fluids and vasopressors. She is currently on norepinephrine and vasopressin with a mean arterial pressure of 67 mmHg. Her past medical history is notable only for a recent bout of asthma treated with albuterol and a prednisone taper over 5 days for contact dermatitis. [S6] Physical exam is notable for abnormally dark skin for a Caucasian woman. The patient states she feels extremely weak. [S8] However, she did experience breastmilk letdown and was able to breastfeed her infant. Laboratory values are ordered as seen below.

Serum: Na+ 127 mEq/L; Cl- 92 mEq/L; K+ 6.1 mEq/L; HCO3- 22 mEq/L; BUN 20 mg/dL; Creatinine 1.1 mg/dL; Ca2+ 10.2 mg/dL

Which of the following is the most likely diagnosis?

- A. Primary adrenal insufficiency
- B. Mycobacteria tuberculosis
- C. Acute kidney injury
- D. Withdrawal from prednisone use

ID	GPT 4.1	GPT 5.5	Sonnet 5	Opus 4.8	Qwen3 8B	Qwen3 32B	Gemma 3	MedGemma
S3	[2,1,1,1] COS	[2,0,1,1] COS	[2,0,1,1] COS	[2,1,1,2] Shared	[1,-1,1,-1] Shared	[2,0,1,0] COS	[1,0,0,0] COS	[1,0,0,0] COS
S6	[2,0,0,0] COS	[2,1,0,-2] COS	[2,1,0,-1] COS	[2,1,0,-1] COS	[2,2,0,0] Shared	[2,0,0,0] COS	[2,0,0,0] COS	[1,0,0,0] COS
S8	[-1,0,0,-1] CU	[0,0,0,0] ND	[2,0,0,0] COS	[1,0,0,0] COS	[-1,-1,0,2] DS	[-1,0,0,0] CU	[0,0,0,0] ND	[1,0,0,0] COS

Fig. 4. Full stem and eight-model maps for three eligible findings from a jointly correct MedQA question. Each cell gives the -2 to +2 rating vector across answer options A-D and the implied role. Highlighted spans in the stem match the matrix rows. The figure’s “Gold answer” denotes the gold option.

also affect the disagreement patterns observed among jointly correct responses.

VI. CONCLUSION

This study shows that agreement on the correct diagnosis does not necessarily imply agreement on how the underlying clinical evidence is interpreted. Across eight LLMs, differential-evidence maps revealed substantial variation in whether findings were considered informative, how strongly they supported the correct diagnosis, and whether support was stronger for that diagnosis or equally strong for a competing diagnosis. Removing findings assigned stronger supportive ratings was associated with larger reductions in diagnostic accuracy than removing findings rated as neutral. Clinically, these results suggest that models reaching the same correct diagnosis may still rely on meaningfully different patterns of evidence appraisal. Evaluating clinical LLMs may therefore require assessing not only diagnostic accuracy, but also

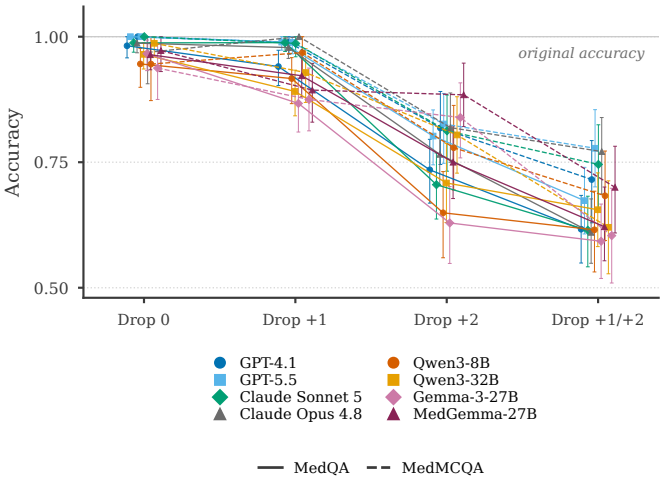


Fig. 5. Accuracy after removing findings the same model rated 0, +1, +2, or both +1 and +2 for the gold option. The cohort is questions the model originally answered correctly, so original accuracy is 1.0. Each condition includes only questions for which at least one eligible finding was removed. Colors identify models. Solid lines are MedQA; dashed lines are MedMCQA. Vertical segments are 95% percentile intervals from 2,000 question-level bootstrap resamples.

MedMCQA

A male patient 59 year old complains of multiple swellings in the neck, fever, and weight loss. He gives a history of hypertension treatment.

Removed +1
General examination reveals painless lymph nodal enlargement in the neck and on blood investigation, anaemia is revealed.

Removed +2
Lymph node biopsy shows typical owl eye-shaped appearance cells.

What can be the most probable diagnosis for this patient?

Options	Before (%)	After (%) Drop +1/+2
A. Multiple myeloma	0.0	0.0
B. Hodgkin's lymphoma (gold)	100.0	79.3
C. Non-Hodgkin's lymphoma	0.0	20.7
D. Burkitt's lymphoma	0.0	0.0

Answer unchanged (B) | B -20.7 pp; C +20.7 pp

Qwen3-32B

Fig. 6. One MedMCQA question in which Qwen3-32B kept its original answer after removal of findings with gold-option ratings of +1 or +2. The highlighted text corresponds to the removed findings. The table gives probabilities normalized across the four answer-option letters before and after deletion. The label “gold” identifies the gold option.

whether the evidence underlying their decisions is interpreted in a clinically coherent and reliable manner.

REFERENCES

[1] D. Jin, E. Pan, N. Oufattole, W.-H. Weng, H. Fang, and P. Szolovits, “What disease does this patient have? A large-scale open domain question answering dataset from medical exams,” *Applied Sciences*, vol. 11, no. 14, Art. no. 6421, 2021, doi: 10.3390/app11146421.

[2] A. Pal, L. K. Umapathi, and M. Sankarasubbu, “MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering,” in *Proc. Conf. Health, Inference, Learn.*, 2022, pp. 248–260.

- [3] OpenAI, "Introducing GPT-4.1 in the API," 2025. [Online]. Available: <https://openai.com/index/gpt-4-1/>
- [4] OpenAI, "Introducing GPT-5.5," 2026. [Online]. Available: <https://openai.com/index/introducing-gpt-5-5/>
- [5] Anthropic, "Introducing Claude Sonnet 5," 2026. [Online]. Available: <https://www.anthropic.com/news/claude-sonnet-5>
- [6] Anthropic, "Introducing Claude Opus 4.8," 2026. [Online]. Available: <https://www.anthropic.com/news/claude-opus-4-8>
- [7] Qwen Team, "Qwen3 technical report," arXiv:2505.09388, 2025.
- [8] Gemma Team, "Gemma 3 technical report," arXiv:2503.19786, 2025.
- [9] Google Research and Google DeepMind, "MedGemma technical report," arXiv:2507.05201, 2025.
- [10] K. Singhal *et al.*, "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.
- [11] S. Bedi *et al.*, "Testing and evaluation of health care applications of large language models: A systematic review," *JAMA*, vol. 333, no. 4, pp. 319–328, 2025, doi: 10.1001/jama.2024.21700.
- [12] P. Hager *et al.*, "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nature Medicine*, vol. 30, pp. 2613–2622, 2024, doi: 10.1038/s41591-024-03097-1.
- [13] P. Qiu *et al.*, "Quantifying the reasoning abilities of LLMs on clinical cases," *Nature Communications*, vol. 16, Art. no. 9799, 2025, doi: 10.1038/s41467-025-64769-1.
- [14] L. Breiman, "Statistical modeling: The two cultures," *Statistical Science*, vol. 16, no. 3, pp. 199–231, 2001, doi: 10.1214/ss/1009213726.
- [15] G. Laberge, Y. Pequignot, A. Mathieu, F. Khomh, and M. Marchand, "Partial order in chaos: Consensus on feature attributions in the Rashomon set," *Journal of Machine Learning Research*, vol. 24, no. 364, pp. 1–50, 2023.
- [16] J. Bogaert, L. Escoufflaire, A. Descampe, C. Fairon, M.-C. de Marneffe, and F.-X. Standaert, "Explanation variability in text classification: Humans vs. LLMs," *Computational Linguistics*, pp. 1–31, 2026, doi: 10.1162/coli.a.624.
- [17] A. Kulmizev, E. Lombart, P. Watrin, and M.-C. de Marneffe, "Label and explanation variation in LLM-based annotation: A case study in natural language inference," in *Proc. 64th Annu. Meeting Assoc. Comput. Linguistics*, 2026, pp. 16526–16543, doi: 10.18653/v1/2026.acl-long.752.
- [18] B. Charlin, L. Roy, C. Brailovsky, F. Goulet, and C. P. M. van der Vleuten, "The Script Concordance test: A tool to assess the reflective clinician," *Teaching and Learning in Medicine*, vol. 12, no. 4, pp. 189–195, 2000, doi: 10.1207/S15328015TLM1204_5.
- [19] J.-P. Fournier, A. Demeester, and B. Charlin, "Script Concordance Tests: Guidelines for construction," *BMC Medical Informatics and Decision Making*, vol. 8, Art. no. 18, 2008, doi: 10.1186/1472-6947-8-18.
- [20] S. Lubarsky, V. Dory, P. Duggan, R. Gagnon, and B. Charlin, "Script concordance testing: From theory to practice: AMEE Guide No. 75," *Medical Teacher*, vol. 35, no. 3, pp. 184–193, 2013.
- [21] L. G. McCoy *et al.*, "Assessment of large language models in clinical reasoning: A novel benchmarking study," *NEJM AI*, vol. 2, no. 10, 2025, doi: 10.1056/AIDBP2500120.
- [22] J. Fang, R. Chen, X. Yang, J. Yu, J. Xu, A. Vinod, W. Shi, T. Chen, H. Ji, C. Zhai, et al., "Benchmarking multi-turn medical diagnosis: Hold, lure, and self-correction," *arXiv preprint arXiv:2604.04325*, 2026.