BEYOND ANSWER CONSISTENCY: EVALUATING THE STABILITY OF EXPLICIT CLINICAL REASONING ACROSS REPEATED LLM GENERATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Final-answer consistency does not necessarily imply stability in the explicit reasoning produced by a large language model (LLM). This distinction is particularly important in clinical diagnosis, where repeated generations may reach the same conclusion while relying on substantially different evidence and interpretations. We study clinical reasoning stability by representing generated diagnostic reasoning as Clinical Reasoning Graphs (CRGs) and comparing their explicit support structures using a symmetric structural distance, CA-CRED. To distinguish prompt-associated changes from ordinary generation variability, we introduce a leave-one-case-out calibration procedure and a case-level calibrated Structural Reasoning Divergence score (cSiRD).

Using 120 clinical cases and 720 formal generations from Qwen3-8B, we find substantial structural variability even under repeated generation of the same case. Importantly, the apparent increase in structural dispersion under Structured Reflection largely disappears after masking diagnostic conclusion identity: the Baseline–Structured within-condition gap decreases from 0.0661 to approximately 0.0020. Likewise, the strong association between conclusion-aware cSiRD and answer changes (AUC $= 0.888$) is markedly reduced under support-only calibration (AUC $= 0.599$), indicating that final diagnosis identity accounts for a substantial portion of the original separation. Nevertheless, semantically consistent diagnostic answers can still coexist with highly divergent support structures. Finally, in an exploratory multi-level analysis, support-only similarity follows a consistent hierarchy across cases, with same-case similarity (0.7280) exceeding same-diagnosis similarity (0.7005), which in turn exceeds different-diagnosis similarity (0.6952). Together, these results suggest that answer consistency, conclusion variation, and support-structure stability capture distinct aspects of observable clinical reasoning behavior.

1 INTRODUCTION

Large language models (LLMs) have shown strong performance in clinical question answering and diagnostic reasoning (Singhal et al., 2023), yet they are still typically evaluated at the level of final answers. Final-answer consistency, however, does not imply consistency in the reasoning that is explicitly produced. Repeated generations for the same case may emphasize different evidence or interpretations while reaching the same diagnosis, and multiple reasoning trajectories may converge to the same answer (Wang et al., 2023). Moreover, generated explanations should not be treated as direct observations of a model's latent computation (Turpin et al., 2023). We therefore study the stability of *observable reasoning structure*, rather than claiming access to the model's internal reasoning process.

This problem is difficult because clinical reasoning is generated as free-form text: wording, ordering, and level of detail naturally vary across generations. Structural comparison is also confounded by the final diagnosis itself. A change in conclusion text can mechanically increase graph distance even if the supporting evidence changes only modestly, whereas answer-level comparison can miss large changes in support structure when the diagnosis remains unchanged. We address these issues by representing generated reasoning as deterministic Clinical Reasoning Graphs (CRGs), comparing

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.

them with a symmetric structural distance (CA-CRED), and calibrating case-level changes against repeated-generation variability using leave-one-case-out (LOSO) reference distributions and cSiRD. We further separate conclusion-aware from support-only divergence and examine structural similarity across the same case, cases sharing a diagnosis, and cases from different diagnosis groups.

On 120 clinical cases and 720 Qwen3-8B generations, we find substantial structural variation even under repeated generation of the same case. The apparent Structured–Baseline dispersion gap drops from 0.0661 to approximately 0.0020 after masking diagnostic conclusion identity, while answer-status discrimination falls from AUC 0.888 to 0.599. Nevertheless, support-only similarity retains a graded structure: $S_{\text{same-case}} = 0.7280$, $S_{\text{same-diagnosis}} = 0.7005$, and $S_{\text{different-diagnosis}} = 0.6952$. These results indicate that conclusion identity, answer consistency, and support structure capture distinct aspects of observable clinical reasoning behavior.

Contributions. We make three contributions. First, we introduce a calibrated framework that combines deterministic CRGs, CA-CRED, and LOSO calibration to evaluate reasoning structure relative to natural repeated-generation variability. Second, we distinguish conclusion-aware from support-only divergence, reducing the direct lexical contribution of diagnosis identity to characterize residual variation in the expressed supporting structure. Third, we evaluate reasoning consistency jointly at within-case and cross-case levels, linking calibrated repeated-generation divergence with multi-level structural similarity across clinical cases. Figure 1 summarizes the framework.

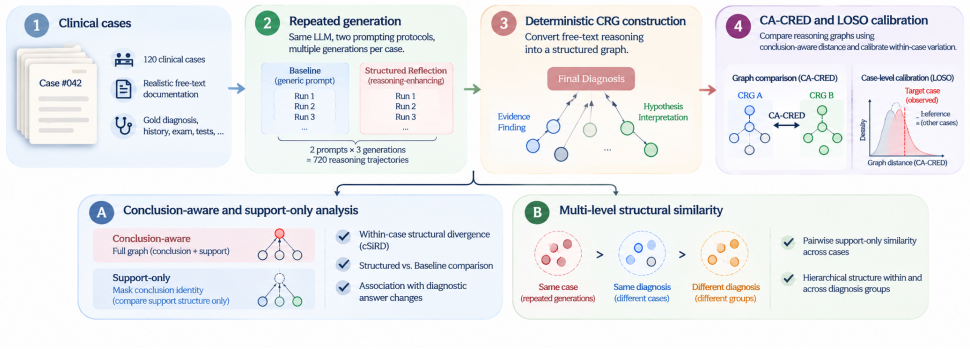

Figure 1: Overview of the framework. Repeated clinical reasoning is converted into deterministic CRGs, compared with CA-CRED, and analyzed through calibrated within-case divergence and cross-case structural similarity.

2 Related Work

Clinical LLM evaluation. Evaluation of medical and clinical large language models has primarily focused on answer-level performance, including medical question answering, diagnostic accuracy, and clinical decision making. MultiMedQA established a broad evaluation framework for medical question answering and demonstrated the capability of large language models to encode clinically relevant knowledge (Singhal et al., 2023). More recent benchmarks have moved toward realistic clinical tasks. ClinicBench evaluates clinical decision-making abilities across a diverse collection of scenarios (Liu et al., 2024), while MedCaseReasoning focuses specifically on diagnostic reasoning from published case reports and provides detailed reasoning statements derived from these reports alongside diagnostic labels (Wu et al., 2025). Savage et al. further studied diagnostic-reasoning prompting, showing that GPT-4 can produce rationales patterned after common clinical reasoning strategies without sacrificing diagnostic accuracy, while noting that such explanations may still reflect post-hoc rationalization (Savage et al., 2024). Complementing these efforts, MedR-Bench evaluates reasoning-enhanced LLMs across examination recommendation, diagnostic decision-making, and treatment planning, and explicitly assesses written reasoning in terms of efficiency, factuality, and completeness (Qiu et al., 2025). Moving beyond static case evaluation, AMIE evaluates diagnostic dialogue along clinically meaningful dimensions including history-taking, diagnostic accuracy, management, communication, and empathy, illustrating the shift toward more interactive and clinically realistic evaluation settings (Tu et al., 2025). A complementary study evaluates AMIE for

differential diagnosis on challenging real-world case reports and further examines its ability to assist clinicians in constructing differential-diagnosis lists (McDuff et al., 2025). Together, these studies broaden clinical LLM evaluation beyond conventional answer-only benchmarks by assessing richer diagnostic and reasoning behaviors. However, they primarily characterize the correctness or quality of individual outputs rather than the stability of expressed reasoning structure across repeated generations. Our focus is complementary: we evaluate how consistently the same clinical case gives rise to explicit reasoning structures across repeated generations and prompting conditions.

Reasoning consistency, repeatability, and faithfulness. Variation across repeated generations is a fundamental property of stochastic language models. Self-consistency exploits this variation by sampling multiple reasoning trajectories and aggregating their final answers, demonstrating that distinct reasoning traces can converge to the same solution (Wang et al., 2023). In clinical settings, repeated outputs have also been studied from the perspective of repeatability and reproducibility; Shyr et al. develop a statistical framework for characterizing variability in diagnostic LLM outputs across repeated runs (Shyr et al., 2025). These lines of work establish that repeated generation is informative, but final-answer agreement alone does not reveal whether the supporting reasoning remains stable. At the same time, generated chain-of-thought or explanatory text should not be interpreted as a transparent record of latent model computation. Prior work has shown that natural-language explanations can be unfaithful to the mechanisms that produced an answer (Turpin et al., 2023), and broader surveys similarly emphasize the distinction between generated explanations and faithful model explanations (Lyu et al., 2024). We therefore deliberately restrict our claims to *observable* or *expressed* reasoning structure rather than latent reasoning processes.

Structured evaluation of clinical reasoning. Free-form rationales are difficult to compare directly because lexical choice, ordering, and level of detail can vary even when the underlying clinical content is similar. Structured representations provide a way to move beyond surface-form comparison. Most directly related to our setting, Clinical Reasoning Graphs (CRGs) organize diagnostic reasoning into structured representations that enable comparison of clinical reasoning traces (Patel, 2026). Our work builds on this structured perspective but addresses a different evaluation question. We are not primarily concerned with whether a single reasoning trace is clinically competent; instead, we study whether the *expressed support structure* is stable across repeated generations of the same case and across prompting conditions.

Positioning of our work. The resulting evaluation problem lies at the intersection of repeated-generation analysis and structured reasoning comparison. Three distinctions are central. First, we compare repeated outputs for the *same case*, rather than relying only on cross-case similarity. Second, we calibrate prompt-associated structural changes against the model's own repeated-generation variability using a leave-one-case-out reference distribution, so that non-zero graph distance is not automatically treated as abnormal. Third, we explicitly distinguish conclusion-aware divergence from support-only divergence by neutralizing the direct lexical identity of the diagnostic conclusion. This separation allows us to examine whether apparent structural changes primarily reflect a changed diagnostic answer or variation in the remaining expressed evidence and interpretation structure. We additionally study the organization of support-only similarity across repeated generations of the same case, different cases sharing a diagnosis, and cases from different diagnosis groups.

3 METHOD

3.1 PROBLEM FORMULATION

Let x_i denote a clinical case. Under prompting condition p and generation repetition r, the LLM produces a free-form reasoning output

$$y_{i,p,r} = f(x_i, p; \xi_r), \tag{1}$$

where ξ_r captures decoding stochasticity. We consider Baseline and Structured Reflection prompting, with three independent generations per condition.

Our goal is not to infer latent internal reasoning, but to quantify the stability of the *explicit reasoning structure* expressed in generated outputs. We ask three questions: (i) how much structure varies across repeated generations of the same case, (ii) whether prompt-associated changes exceed ordinary generation variability, and (iii) whether structural similarity exhibits an interpretable hierarchy across the same case, cases sharing a diagnosis, and cases from different diagnosis groups.

3.2 Clinical Reasoning Graph Construction

Each generated response is deterministically converted into a Clinical Reasoning Graph (CRG). The graph contains FACT nodes for explicitly stated clinical observations, INTERPRETATION nodes for explicit diagnostic or supportive statements, and one CONCLUSION node representing the final diagnosis. Retained FACT and INTERPRETATION nodes connect directly to the CONCLUSION through SUPPORTS edges.

The mapping is rule based and frozen before formal analysis. Candidate spans are normalized, deduplicated, and assigned roles using fixed lexical cues; at most 15 FACT and 8 INTERPRETATION nodes are retained. The conclusion is identified using a fixed ordered set of answer-anchor patterns, without post-hoc synonym expansion or semantic repair. The resulting graph is therefore a constrained star-shaped representation of explicitly expressed support organization rather than a reconstruction of latent multi-step causal reasoning. Full construction rules are reported in Appendix Table A1.

3.3 CA-CRED Structural Distance

Given two CRGs, G_a and G_b, CA-CRED compares the explicit support paths terminating at their conclusion nodes. Node matching cost combines lexical token-set distance with role consistency. For nodes u and v,

$$c(u, v) = 0.7\, d_{\text{sem}}(u, v) + 0.3\, d_{\text{role}}(u, v), \tag{2}$$

where d_{sem} is Jaccard distance over lowercased alphanumeric token sets, and d_{role} is 0 for matching roles and 0.5 otherwise.

Support paths are matched by minimum-cost correspondence, first at the path level and then across graphs. Let $\Delta(G_a \to G_b)$ denote the average minimum path-matching cost from G_a to G_b. The final symmetric distance is

$$D_{\text{CA-CRED}(G_a, G_b)} = \frac{\Delta(G_a \to G_b) + \Delta(G_b \to G_a)}{2}. \tag{3}$$

Lower values indicate greater similarity in the expressed reasoning structure. The complete node-, path-, and graph-level matching equations are provided in the Appendix.

3.4 LOSO Calibration and cSiRD

Raw CA-CRED values are difficult to interpret because repeated generations of the same case already exhibit non-zero structural variation. For each case, three Baseline and three Structured Reflection generations produce six within-condition distances, which characterize ordinary repeated-generation variability. The nine Baseline–Structured pairwise distances characterize the prompt-associated shift, summarized by their median C_i.

To avoid self-calibration, the reference distribution $\mathcal{R}_{-i}$ is formed from the six within-condition distances of all other cases. With 120 cases, each reference contains $119 \times 6 = 714$ distances. We define

$$\text{cSiRD}_i = \frac{1}{|\mathcal{R}_{-i}|} \sum_{d \in \mathcal{R}_{-i}} \mathbf{1}[d \le C_i]. \tag{4}$$

Thus, cSiRD is the empirical percentile of the observed cross-condition shift relative to natural repeated-generation variability. Under the frozen protocol, $cSiRD > 0.95$ defines a high-divergence case. cSiRD measures calibrated structural divergence, not clinical correctness or reasoning quality.

3.5 Conclusion-Aware and Support-Only Analysis

Because every support path terminates at the diagnostic conclusion, the original CA-CRED distance reflects variation in both the supporting structure and final diagnosis identity. We therefore analyze two complementary views.

The conclusion-aware analysis compares the original graphs directly. For the support-only analysis, a masking transformation $M(\cdot)$ replaces every CONCLUSION-node text with the same constant

token while preserving graph topology, node roles, and all FACT and INTERPRETATION content:

$$D_{\text{support}}(G_a, G_b) = D_{\text{CA-CRED}}\big(M(G_a), M(G_b)\big). \tag{5}$$

This reduces the direct lexical contribution of diagnosis identity without removing the conclusion node itself. Conclusion-aware distance therefore captures joint diagnostic-output divergence, whereas support-only distance emphasizes variation in the expressed evidence and interpretation structure.

3.6 Multi-Level Structural Similarity

To examine cross-case organization without directly rewarding shared diagnosis text, we use support-only similarity

$$S(G_a, G_b) = 1 - D_{\text{support}}(G_a, G_b). \tag{6}$$

For each target case, we estimate three levels: repeated generations of the same case, different cases within the same diagnosis group, and cases from different diagnosis groups. Comparisons use only generations from the same prompting condition. For same-diagnosis comparisons, comparator cases receive equal weight; for different-diagnosis comparisons, similarities are first averaged within each comparator diagnosis group and then averaged equally across groups. Baseline and Structured Reflection estimates are finally averaged at the case level. We test the hierarchy using two contrasts,

$$D_1 = S_{\text{same-case}} - S_{\text{same-diagnosis}}, \qquad D_2 = S_{\text{same-diagnosis}} - S_{\text{different-diagnosis}}. \tag{7}$$

Positive D_1 and D_2 correspond to the ordering $S_{\text{same-case}} > S_{\text{same-diagnosis}} > S_{\text{different-diagnosis}}$. Full aggregation equations and statistical details are deferred to the Appendix.

4 Experiments

4.1 Experimental Setup

Dataset and generation protocol. We evaluate on 120 cases from MedCaseReasoning (Wu et al., 2025), organized into 12 frozen diagnosis groups with 10 cases per group. Cases used in earlier pilot experiments are excluded from the formal evaluation set. For each case, Qwen3-8B (Yang et al., 2025) is evaluated under two prompting conditions: a Baseline prompt and a Structured Reflection prompt. Each case–condition pair is generated independently three times, yielding $120 \times 2 \times 3 = 720$ formal generations in total. All formal analyses reported in this study are based on this fixed Qwen3-8B evaluation set. We use Qwen3-8B with model-native thinking disabled. Generation uses stochastic sampling with temperature 1.0, top-$p = 0.95$, top-$k = 20$, and a maximum of 8192 newly generated tokens. All cases and repetitions use the same frozen generation configuration. Each generation job is assigned a deterministic seed derived from the model identifier, model revision, case identifier, prompting condition, and repetition index. The corresponding seed is applied consistently to the Python, NumPy, and model-generation random number generators, providing three reproducible, seed-controlled stochastic draws for each case–condition pair.

The Baseline prompt requests a ranked differential diagnosis, a leading diagnosis, and step-by-step supporting reasoning. Structured Reflection adds a three-phase deliberation procedure consisting of an initial problem representation, a structured second look that stress-tests the leading diagnosis and considers alternatives, and a revised final assessment. The frozen prompting protocol is described in Appendix A.1.

The repeated-generation design provides multiple structural comparisons within each case. Three generations per condition yield three pairwise distances within Baseline and three within Structured Reflection, giving six within-condition distances per case. The two sets of three generations further produce nine Baseline–Structured pairwise distances. These repeated comparisons form the basis of the within-case variability estimate and the LOSO-calibrated cross-condition analysis described in Section 3.4.

Graph construction and evaluation. All generated outputs are converted into deterministic V5.4.1 Clinical Reasoning Graphs using the frozen procedure in Section 3.2. We evaluate both conclusion-aware and support-only CA-CRED distances, LOSO-calibrated cSiRD, and multi-level support-only

similarity. Final answers are separately assigned to SAME, DIFFERENT, or UNCERTAIN categories using blinded model-assisted semantic adjudication performed without access to structural-distance or cSiRD results. These answer-status labels are used only for secondary analyses of the relationship between diagnostic consistency and structural divergence.

Statistical analysis. The case is the primary unit of analysis. Diagnosis groups are treated as clusters when testing group-level consistency of the observed effects. For the multi-level similarity contrasts, we use exact cluster sign-flip tests over the 12 diagnosis groups together with hierarchical bootstrap confidence intervals. The balanced 10-case-per-group design also allows same-diagnosis and different-diagnosis comparisons to be aggregated without allowing larger diagnosis groups to dominate the analysis. The Appendix provides the frozen prompting protocol, full graph-construction rules, aggregation equations, and statistical details.

4.2 Structural Stability under Repeated Generation

Conclusion-aware CA-CRED revealed substantial repeated-generation variability. Mean within-condition distance was 0.4374 for Baseline and 0.5035 for Structured Reflection, while mean cross-condition distance was 0.4995. Structured Reflection therefore showed a mean within-condition increase of 0.0661; this difference was positive in 86/120 cases and 11/12 diagnosis groups (cluster sign-flip $p = 0.00146$; 95% bootstrap CI $[0.0325, 0.0983]$). Because this metric includes the final diagnosis, the difference should be interpreted as dispersion in the complete expressed output structure rather than as evidence of poorer clinical reasoning.

LOSO calibration further showed that prompt-associated shifts were usually moderate relative to natural generation variability. Mean and median cSiRD were 0.596 and 0.581, respectively, and only 12/120 cases (10.0%) exceeded the frozen high-divergence threshold of 0.95. We next separate conclusion identity from the remaining support structure.

4.3 Conclusion-Aware and Support-Only Analysis

Masking diagnostic conclusion identity substantially changed the observed pattern. Support-only within-condition distances were 0.2710 for Baseline and 0.2730 for Structured Reflection, reducing the Structured–Baseline gap from 0.0661 to approximately 0.0020. Likewise, separation between SAME and DIFFERENT answers decreased sharply after masking: AUC fell from 0.8882 under conclusion-aware calibration to 0.5985 under support-only calibration (Table 1).

Table 1: Quantitative comparison of conclusion-aware and support-only structural analyses on the Qwen3-8B evaluation set.

Metric	**Conclusion-aware**	**Support-only**
Structural distances		
Baseline within-condition distance	0.4374	0.2710
Structured within-condition distance	0.5035	0.2730
Cross-condition distance	0.4995	0.2888
Structured–Baseline gap	0.0661	0.0020
Semantic answer-status separation		
SAME mean cSiRD	0.3638	0.8557
DIFFERENT mean cSiRD	0.7308	0.8913
DIFFERENT–SAME difference	0.3669	0.0355
95% bootstrap CI	[0.2875, 0.4434]	$[-0.0119, 0.0851]$
Permutation p	$< 10^{-5}$	0.1367
AUC	0.8882	0.5985
Cliff's δ	0.7763	0.1971
High-divergence cases among SAME answers		
Cases with cSiRD > 0.95	0/23 (0.0%)	5/23 (21.7%)

Note. Conclusion-aware and support-only cSiRD values use separately calibrated leave-one-case-out reference distributions; their absolute values and high-divergence rates therefore should not be interpreted as directly comparable scales.

These results indicate that final diagnosis identity accounts for a substantial part of the conclusion-aware divergence. At the same time, support-only variation does not disappear: 5/23 semantically consistent cases exceeded $cSiRD > 0.95$ after separate support-only recalibration, showing that semantically consistent diagnostic answers can coexist with markedly different expressed support structures. Because the two cSiRD variants use different reference distributions, their absolute values and threshold exceedance rates are not directly comparable.

Because this decomposition was introduced during the Qwen3-8B analysis, we treat this support-only decomposition as exploratory rather than confirmatory.

4.4 Multi-Level Structural Similarity

Support-only similarity exhibited a graded hierarchy across the 120 evaluation cases:

$$S_{\text{same-case}} = 0.7280 > S_{\text{same-diagnosis}} = 0.7005 > S_{\text{different-diagnosis}} = 0.6952.$$

The same ordering held separately for Baseline ($0.7290 > 0.7012 > 0.6958$) and Structured Reflection ($0.7270 > 0.6998 > 0.6947$), as shown in Figure 2.

Both hierarchy contrasts were positive at the diagnosis-group level: $D_1 = 0.0275$ for same-case versus same-diagnosis similarity and $D_2 = 0.00528$ for same-diagnosis versus different-diagnosis similarity (both one-sided cluster sign-flip $p = 2.44 \times 10^{-4}$).

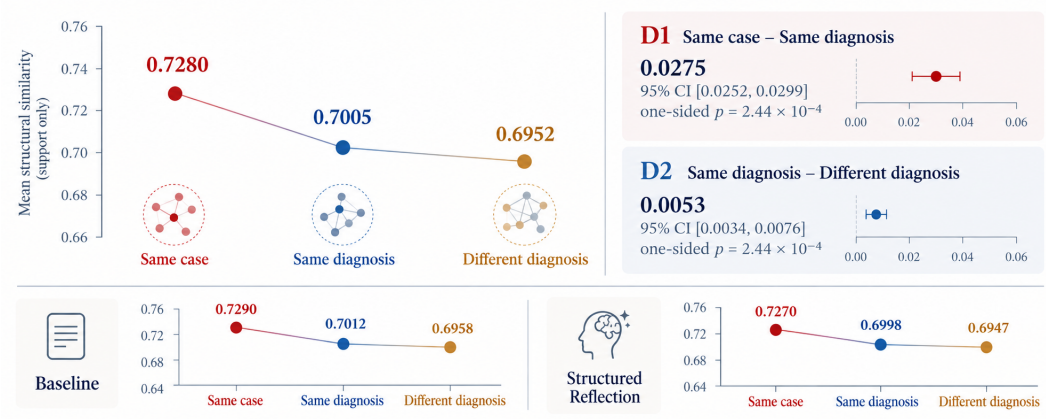

Figure 2: Multi-level support-only structural similarity on the Qwen3-8B evaluation set. Similarity decreases from repeated generations of the same case, to different cases within the same diagnosis group, and finally to cases from different diagnosis groups. Both hierarchy contrasts are positive across all 12 diagnosis groups.

The substantially larger D_1 indicates strong case-specific structural preservation, whereas the smaller but consistent D_2 suggests a modest diagnosis-level regularity after direct conclusion identity is removed. Detailed case counts and confidence intervals are reported in the Appendix.

Because the multi-level protocol was specified after earlier Qwen3-8B analyses, we interpret the observed hierarchy as exploratory evidence rather than an independent confirmatory result.

4.5 Sensitivity to the cSiRD Threshold

We next examine whether the identified high-divergence subset depends strongly on the frozen cSiRD threshold. Under conclusion-aware calibration, the number of flagged cases decreased monotonically as the threshold became more stringent: 24/120 cases (20.0%) exceeded cSiRD > 0.90, 12/120 (10.0%) exceeded > 0.95, 8/120 (6.7%) exceeded > 0.975, and 4/120 (3.3%) exceeded > 0.99. Thus, the main observation that extreme prompt-associated structural divergence is concentrated in a relatively small subset of cases is not specific to a single threshold choice.

The primary analysis retains the frozen decision rule of cSiRD > 0.95; the alternative thresholds are used only as a sensitivity analysis and do not alter the primary definition of a high-divergence

case. Category-specific counts for SAME, DIFFERENT, and UNCERTAIN answer-status groups are reported in Appendix Table A2.

5 DISCUSSION

Our results highlight a distinction between final-answer consistency and the stability of explicitly generated clinical reasoning. Repeated generations of the same case exhibit substantial non-zero structural variation, and prompting can further alter the observed reasoning structure. This variation should not, however, be interpreted directly as instability of a model's latent reasoning process. Our framework evaluates the structure expressed in generated reasoning traces—which clinical facts and interpretations are explicitly represented and how they are organized relative to the final diagnosis. Two generations can therefore reach the same diagnostic conclusion while exposing different support structures, or reach different conclusions while sharing substantial portions of their expressed evidence organization.

The conclusion-aware and support-only analyses show that diagnosis identity and support-structure variation should not be conflated. Under the conclusion-aware metric, Structured Reflection exhibited greater within-condition dispersion than Baseline, with a mean gap of 0.0661. After neutralizing the lexical identity of the diagnostic conclusion, this gap decreased to approximately 0.0020, a reduction of about 97%. Similarly, discrimination between semantically consistent and answer-changing cases decreased from AUC $= 0.888$ under conclusion-aware calibration to AUC $= 0.599$ under support-only calibration. These findings suggest that final diagnosis identity accounts for a substantial portion of the separation captured by the original metric. Conclusion-aware CA-CRED should therefore be interpreted as measuring variation in the joint diagnostic output structure, whereas the support-only analysis provides a complementary view that emphasizes variation remaining after the direct lexical identity of the diagnostic conclusion is neutralized.

Importantly, neutralizing conclusion identity does not eliminate all structural organization. Support-only similarity followed the graded ordering $S_{\text{same-case}} > S_{\text{same-diagnosis}} > S_{\text{different-diagnosis}}$. The same-case contrast was substantially larger, indicating strong preservation of case-specific support structure across repeated generations. The same-diagnosis advantage over different-diagnosis cases was considerably smaller ($D_2 = 0.00528$), but remained positive in $116/120$ cases and all 12 diagnosis groups. We therefore interpret this result as evidence for a modest but consistent diagnosis-level regularity in the observable support representation, rather than as evidence for sharply separated diagnostic reasoning schemas.

LOSO calibration provides a complementary case-level interpretation by placing an observed cross-condition shift relative to the model's own repeated-generation variability on other cases. This avoids treating every non-zero graph distance as anomalous. At the frozen threshold of cSiRD > 0.95, only $12/120$ evaluation cases were classified as high-divergence cases. As shown in Section 4.5, this conclusion is qualitatively stable across a range of more permissive and more stringent thresholds. More generally, cSiRD should be interpreted as a calibrated measure of structural divergence in generated reasoning, not as a measure of diagnostic correctness, clinical quality, or latent reasoning fidelity.

Implications for clinical LLM evaluation. Taken together, our results suggest that answer-level consistency alone provides an incomplete view of reasoning stability. At least three sources of variation should be distinguished: whether the final diagnostic answer changes, whether the lexical identity of the conclusion changes the measured structural distance, and whether the remaining expressed support structure itself varies. These dimensions can behave differently. In our analysis, much of the apparent prompt-associated increase in conclusion-aware divergence was attenuated after neutralizing conclusion identity, yet support-only variability did not vanish, and semantically consistent answers could still exhibit substantial structural differences. This suggests that evaluations of clinical prompting strategies should avoid treating a stable final answer as sufficient evidence of stable reasoning behavior. A more informative assessment should jointly consider answer consistency, calibrated structural divergence, and the organization of the explicitly expressed support structure.

Several limitations define the scope of these findings. First, the deterministic V5.4.1 CRG construction intentionally favors reproducibility over expressive graph complexity. Its star-shaped support

topology does not reconstruct latent multi-step causal reasoning. Second, lexical and rule-based extraction can miss clinically equivalent expressions that fall outside the frozen patterns. Moreover, because the masked conclusion node remains in every support path and diagnosis-related language may still occur in FACT or INTERPRETATION nodes, the support-only metric reduces rather than eliminates diagnosis-related influence. Third, semantic answer status is determined using blinded model-assisted adjudication rather than independent expert clinical annotation and should therefore be treated as supporting analysis rather than a clinical gold standard. Finally, all formal experiments in this study are conducted with Qwen3-8B. The extent to which the observed stability patterns generalize across model families therefore remains untested. Replication across independently selected LLMs is an important direction for future work.

6 Conclusion

We introduced a framework for evaluating consistency in explicitly generated clinical reasoning beyond final-answer agreement. By combining deterministic Clinical Reasoning Graphs, CA-CRED structural comparison, and leave-one-case-out calibration, the framework separates ordinary repeated-generation variability from larger case-level structural shifts. Our results show that conclusion identity substantially influences conclusion-aware divergence, while support-only analysis reveals residual variation in evidence and interpretation structure even when diagnostic answers remain semantically unchanged. An exploratory multi-level analysis further reveals a graded hierarchy in support-only similarity, from repeated generations of the same case, to cases sharing a diagnosis, to cases from different diagnosis groups. Together, these results motivate evaluating clinical LLMs not only by whether they reach the same answer, but also by how consistently they express the support structure leading to that answer.

Ethics Statement

This study uses the open-access MedCaseReasoning benchmark, which is derived from published clinical case reports, and involves no new patient recruitment or data collection. The work evaluates model-output stability and is not intended as clinical decision support or medical advice.

Reproducibility Statement

We emphasize reproducibility throughout the evaluation pipeline. The clinical case selection and repeated-generation protocol are fixed before formal analysis, and all reported formal results use the same 120-case evaluation set. Section 3 specifies the deterministic CRG representation, CA-CRED, LOSO calibration, support-only masking, and multi-level similarity definitions. Section 4 describes the generation and statistical-analysis protocol, while Appendix A provides the complete CA-CRED equations, frozen CRG construction rules, threshold-sensitivity results, and multi-level aggregation and inference procedures.

AI Use Statement

Generative AI tools were used in an auxiliary role for methodological and experimental feedback, literature retrieval and organization, implementation and code assistance, interpretation support, and manuscript preparation. All AI-assisted outputs, including methodological suggestions, code, analyses, citations, and manuscript text, were reviewed and verified by the authors. The authors take full responsibility for the content of this work.

References

Fenglin Liu, Zheng Li, Hongjian Zhou, Qingyu Yin, Jingfeng Yang, Xianfeng Tang, Chen Luo, Ming Zeng, Haoming Jiang, Yifan Gao, Priyanka Nigam, Sreyashi Nag, Bing Yin, Yining Hua, Xuan Zhou, Omid Rohanian, Anshul Thakur, Lei Clifton, and David A. Clifton. Large language models are poor clinical decision-makers: A comprehensive benchmark. In *Proceedings of the*

2024 Conference on Empirical Methods in Natural Language Processing, pp. 13696–13710. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.759.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, 50(2):657–723, 2024. doi: 10.1162/coli_a_00511.

Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, Le Hou, Yong Cheng, Yun Liu, S. Sara Mahdavi, Sushant Prakash, Anupam Pathak, Christopher Semturs, Shwetak Patel, Dale R. Webster, Ewa Dominowska, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Jake Sunshine, Alan Karthikesalingam, and Vivek Natarajan. Towards accurate differential diagnosis with large language models. *Nature*, 642:451–457, 2025. doi: 10.1038/s41586-025-08869-4.

Nisarg A. Patel. Clinical reasoning graphs: Structured evaluation of llm diagnostic reasoning reveals competence without consistency. *arXiv preprint arXiv:2606.29876*, 2026.

Pengcheng Qiu, Chaoyi Wu, Shuyu Liu, Yanjie Fan, Weike Zhao, Zhuoxia Chen, Hongfei Gu, Chuanjin Peng, Ya Zhang, Yanfeng Wang, and Weidi Xie. Quantifying the reasoning abilities of LLMs on clinical cases. *Nature Communications*, 16:9799, 2025. doi: 10.1038/s41467-025-64769-1.

Thomas Savage, Ashwin Nayak, Robert Gallo, Ekanath Rangan, and Jonathan H. Chen. Diagnostic reasoning prompts reveal the potential for large language model interpretability in medicine. *npj Digital Medicine*, 7:20, 2024. doi: 10.1038/s41746-024-01010-1.

Cathy Shyr, Boyu Ren, Chih-Yuan Hsu, Chao Yan, Rory J. Tinker, Thomas A. Cassini, Rizwan Hamid, Adam Wright, Lisa Bastarache, Josh F. Peterson, Bradley A. Malin, and Hua Xu. A statistical framework for evaluating the repeatability and reproducibility of large language models. *medRxiv*, 2025. doi: 10.1101/2025.08.06.25333170. Preprint.

Karan Singhal, Shekoofeh Azizi, Tao Tu, et al. Large language models encode clinical knowledge. *Nature*, 620:172–180, 2023. doi: 10.1038/s41586-023-06291-2.

Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, Alan Karthikesalingam, and Vivek Natarajan. Towards conversational diagnostic artificial intelligence. *Nature*, 642:442–450, 2025. doi: 10.1038/s41586-025-08866-7.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023.

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.

An Yang, Anfeng Li, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

A Additional Implementation Details

A.1 Frozen Prompting Protocol

The two prompting conditions were frozen before formal generation and were used without condition-specific rewriting or post-hoc modification. The Baseline prompt requested a ranked top-five differential diagnosis, a leading diagnosis with estimated probability, and step-by-step reasoning describing findings supporting or arguing against each candidate.

Structured Reflection used a three-phase protocol. Phase 1 required an initial problem representation and ranked differential diagnosis. Phase 2 required a structured second look that re-examined the problem representation, stress-tested the leading diagnosis, considered the strongest alternative and a "can't-miss" diagnosis, and then defended or updated the initial ranking. Phase 3 required a revised differential and final leading diagnosis, together with an explicit statement of whether reassessment changed the diagnosis.

For both conditions, the clinical case text was supplied as the user message, while the corresponding frozen condition prompt was supplied as the system message. Prompt text and message construction were held fixed across all cases and repetitions.

All formal generations used Qwen3-8B at the fixed model revision `b968826d9c46dd6066d109eabc6255188de91218`.

A.2 Full CA-CRED Definition

For a CRG G, let the set of explicit support paths be

$$\mathcal{P}(G) = \{(v, c) \mid (v, c, \text{SUPPORTS}) \in E\}. \tag{8}$$

For nodes u and v, the lexical token-set distance is

$$d_{\text{sem}}(u, v) = 1 - \frac{|T(u) \cap T(v)|}{|T(u) \cup T(v)|}, \tag{9}$$

where $T(\cdot)$ denotes the lowercased alphanumeric token set. Role distance is

$$d_{\text{role}}(u, v) = \begin{cases} 0, & \text{role}(u) = \text{role}(v), \\ 0.5, & \text{otherwise.} \end{cases} \tag{10}$$

The node matching cost is

$$c(u, v) = 0.7 d_{\text{sem}}(u, v) + 0.3 d_{\text{role}}(u, v). \tag{11}$$

For two support paths P_a and P_b, the directed distance is

$$\delta(P_a \to P_b) = \frac{1}{|P_a|} \sum_{u \in P_a} \min_{v \in P_b} c(u, v), \tag{12}$$

and the symmetric path distance is

$$d_{\text{path}}(P_a, P_b) = \frac{\delta(P_a \to P_b) + \delta(P_b \to P_a)}{2}. \tag{13}$$

The directed graph distance is

$$\Delta(G_a \to G_b) = \frac{1}{|\mathcal{P}(G_a)|} \sum_{P_a \in \mathcal{P}(G_a)} \min_{P_b \in \mathcal{P}(G_b)} d_{\text{path}}(P_a, P_b), \tag{14}$$

and CA-CRED is

$$D_{\text{CA-CRED}(G_a, G_b)} = \frac{\Delta(G_a \to G_b) + \Delta(G_b \to G_a)}{2}. \tag{15}$$

Table A1: Frozen deterministic CRG construction rules used throughout the V5.4.1 analyses.

Component	Frozen rule
Graph topology	One unique CONCLUSION node; retained FACT and INTERPRETATION nodes connect directly to the conclusion through SUPPORTS edges.
Node roles	FACT, INTERPRETATION, and CONCLUSION.
Maximum FACT nodes	15 per reasoning trace.
Maximum INTERPRETATION nodes	8 per reasoning trace.
FACT cues	age, year-old, history, symptom, finding, imaging, CT, MRI, level, elevation, marker, mass, lesion, absence, and presence.
INTERPRETATION cues	suggest, consistent, supports, indicate, compatible, and associated.
Dropped phrases	supporting findings; ranking reasoning; estimated probability; high/moderate/low probability; leading diagnosis; top 5.
Minimum retained line length	Lines shorter than 15 characters are discarded.
Duplicate handling	Duplicate lines are removed before graph construction.
Conclusion extraction	A fixed sequence of predefined answer-anchor patterns is applied; empty cleaned matches are skipped without introducing new patterns or synonyms.
Implicit edge creation	Disabled. No unsupported relation is inferred.
Graph construction	Deterministic; no additional LLM-based extractor is used.

Table A2: Sensitivity of high-divergence case counts to the cSiRD threshold in the conclusion-aware Qwen3-8B analysis.

cSiRD threshold	SAME	DIFFERENT	UNCERTAIN	All cases
> 0.90	0/23 (0.0%)	21/62 (33.9%)	3/35 (8.6%)	24/120 (20.0%)
> 0.95	0/23 (0.0%)	11/62 (17.7%)	1/35 (2.9%)	12/120 (10.0%)
> 0.975	0/23 (0.0%)	7/62 (11.3%)	1/35 (2.9%)	8/120 (6.7%)
> 0.99	0/23 (0.0%)	3/62 (4.8%)	1/35 (2.9%)	4/120 (3.3%)

Note. The primary analysis uses the frozen threshold of cSiRD > 0.95. The additional thresholds are reported only as a sensitivity analysis and do not alter the primary decision rule.

A.3 Multi-Level Aggregation and Statistical Inference

For prompting condition p, same-case similarity for case i is

$$S^{(p)}_{\text{same-case}}(i) = \frac{1}{3} \sum_{r<r'} S(G_{i,p,r}, G_{i,p,r'}). \tag{16}$$

Same-diagnosis similarity averages all 3×3 repetition combinations for each comparator case and then gives each comparator case equal weight:

$$S^{(p)}_{\text{same-diagnosis}}(i) = \frac{1}{|C_{g(i)} \setminus \{i\}|} \sum_{j \in C_{g(i)} \setminus \{i\}} \frac{1}{9} \sum_{r=1}^{3} \sum_{r'=1}^{3} S(G_{i,p,r}, G_{j,p,r'}). \tag{17}$$

For different-diagnosis similarity, comparator cases are first averaged within each diagnosis group and the resulting diagnosis-group means are then weighted equally:

$$S^{(p)}_{\text{different-diagnosis}}(i) = \frac{1}{|\mathcal{G} \setminus \{g(i)\}|} \sum_{g \neq g(i)} \frac{1}{|C_g|} \sum_{j \in C_g} \frac{1}{9} \sum_{r=1}^{3} \sum_{r'=1}^{3} S(G_{i,p,r}, G_{j,p,r'}). \tag{18}$$

Baseline and Structured Reflection values are averaged at the case level. Inference for D_1 and D_2 uses an exact one-sided sign-flip test over the 12 diagnosis-group mean effects ($2^{12} = 4096$ sign configurations). Confidence intervals use a 100,000-repetition hierarchical bootstrap that resamples diagnosis groups and then cases within groups.