

Evaluating Eliminative Reasoning in LLM-Based Differential Diagnosis

Stanford CS224N {Custom} Project

Tatiana Zhang
Department of Computer Science
Stanford University
tytzhang@stanford.edu

Seyun (Joy) Bang
Department of Computer Science
Stanford University
sjoybang@stanford.edu

Abstract

Large language models (LLMs) achieve strong performance on medical question-answering benchmarks, yet these benchmarks evaluate only final answer selection from a fixed candidate set, making it impossible to distinguish a model that reasons coherently from one that reaches the correct answer through clinically incoherent intermediate steps. We propose a process-oriented evaluation framework that measures eliminative reasoning in LLM-based differential diagnosis under sequential and counterfactual evidence presentation. We construct 60 staged clinical vignettes from MedQA-USMLE, each paired with a counterfactual variant introducing a discriminative finding that argues against the ground-truth diagnosis, and evaluate four LLMs — Claude Sonnet 4.5, Gemini 2.0 Flash, Gemini 2.5 Pro, and GPT-4o — using three newly defined metrics: Elimination Monotonicity, Evidence Sensitivity, and Cross-Step Consistency. Our results reveal two key findings: diagnostic accuracy and eliminative reasoning quality are substantially dissociated, and greater sensitivity to counterfactual evidence does not predict better diagnostic performance, suggesting that reactivity to perturbation and clinically coherent reasoning are not equivalent. These findings demonstrate that process-oriented metrics surface reasoning failures that accuracy-based evaluation cannot detect.

1 Key Information to include

- Mentor: Nevin George
- External Collaborators (if you have any): No
- Sharing project: No
- Late Days (if you are using any):

Team member	Late days remaining
<Tatiana Zhang>	<3>
<Joy Bang>	<3>
Total pooled	<6>
Team size	<2>
Late days applied for the report	<3>

2 Introduction

Large language models have demonstrated impressive performance on medical question-answering (QA) benchmarks such as MedQA, a dataset of multiple-choice clinical questions drawn from the United States Medical Licensing Examination (USMLE) — the standardized exam required for medical practice in the US. Recent models achieve near-expert accuracy on these examinations [1, 2]. However, these benchmarks evaluate only final-answer accuracy — whether the model selects the

correct diagnosis from a set of options. In clinical practice, arriving at the right answer is necessary but not sufficient. Physicians engage in a structured process of eliminative reasoning: that is, maintaining a differential diagnosis, revising it as new evidence emerges, and permanently excluding hypotheses when contradicted by discriminative findings. A model that reaches the correct final diagnosis through inconsistent intermediate reasoning — for example, eliminating a diagnosis at one stage only to reintroduce it later without justification — may be relying on surface-level pattern matching rather than coherent clinical logic.

This distinction matters because downstream applications of LLMs in medicine, such as clinical decision support and diagnostic assistance, require not just correct outputs but trustworthy reasoning processes. If a model’s reasoning cannot be relied upon to behave consistently under new or contradictory evidence, its utility in safety-critical settings is fundamentally limited. Yet existing evaluation frameworks offer no way to detect these failures: a model that reasons incoherently but arrives at the right answer receives the same score as one that reasons faithfully throughout.

We address this gap by proposing a process-oriented evaluation framework for eliminative reasoning in differential diagnosis. Our approach has three components. First, we construct staged clinical vignettes drawn from MedQA [3], decomposed into sequential evidence stages with minimally contrasting counterfactual variants designed to rule out specific diagnoses. Second, we define three evaluation metrics — elimination monotonicity, evidence sensitivity, and cross-step consistency — that measure whether models appropriately revise their beliefs as evidence accumulates, drawing on tools from information theory [3], information retrieval [4], and consistency evaluation in NLP [5, 6]. Third, we evaluate multiple LLMs including GPT-4o, Claude Sonnet 4.5, Gemini 2.0 Flash, and Gemini 2.5 Pro using zero-shot chain-of-thought prompting [7].

3 Related Work

LLMs in Medical Question Answering. Recent work has demonstrated strong LLM performance on medical QA benchmarks. Nori et al. [1] showed GPT-4 achieving passing scores on USMLE examinations, while Singhal et al. [2] introduced Med-PaLM, demonstrating that LLMs can encode substantial clinical knowledge. These evaluations focus on final-answer accuracy, leaving open the question of whether models reason coherently along the way. Our work complements these benchmarks by evaluating the reasoning process rather than just the outcome.

Chain-of-Thought and Faithful Reasoning. Wei et al. [7] showed that chain-of-thought prompting improves performance on complex reasoning tasks by eliciting step-by-step explanations. However, subsequent work has questioned whether these explanations faithfully reflect the model’s internal reasoning process. Turpin et al. [8] demonstrated that chain-of-thought explanations can be systematically unfaithful, with models producing plausible-sounding justifications that do not correspond to their actual decision process. Our framework provides a concrete way to test faithfulness in the clinical domain: if a model claims to eliminate a diagnosis but later reintroduces it, the reasoning is demonstrably inconsistent.

Counterfactual Evaluation. Kaushik et al. [9] introduced counterfactually augmented data to test model robustness, creating minimally edited examples that change the ground-truth label. Ribeiro et al. [10] proposed CheckList, a behavioral testing framework using perturbations to evaluate NLP models beyond aggregate accuracy. Our counterfactual variants follow a similar philosophy but are designed specifically for clinical reasoning: we alter a single discriminative finding to test whether the model revises its differential appropriately.

Consistency and Process-Oriented Evaluation. Our evaluation metrics draw on several lines of prior work. Elimination Monotonicity builds on consistency evaluation in pretrained language models — Elazar et al. [5] measured whether models give consistent answers to semantically equivalent queries, finding substantial inconsistency even in large models. We adapt this idea from paraphrase consistency to sequential consistency: rather than asking whether a model answers the same question the same way twice, we ask whether a model that eliminates a diagnosis at one stage keeps it eliminated at the next. Cross-Step Consistency extends this further, drawing on Wang et al. [6], who introduced self-consistency as a decoding strategy by sampling multiple reasoning paths. Where they measure agreement across parallel reasoning paths, we measure agreement across sequential stages of a single case. Evidence Sensitivity uses KL divergence [3] — a standard information-theoretic measure of distributional difference — applied to rank-derived probability

distributions using reciprocal rank weighting, a technique from information retrieval [4]. The specific combination of these tools into a unified framework for evaluating eliminative reasoning in staged clinical diagnosis, and the counterfactual perturbation methodology used to stress-test belief revision, are our contributions.

4 Approach

4.1 Problem Formulation

Each clinical case is revealed in 4 stages. After each stage t , the model produces a ranked differential $D_t = (d_1, \dots, d_5)$ and a natural-language justification, along with an explicit list of eliminated diagnoses. We evaluate whether models appropriately revise beliefs as new evidence is introduced, particularly when evidence contradicts a previously plausible diagnosis.

4.2 Dataset Construction

We construct our dataset from MedQA-USMLE [11], a publicly available collection of over 10,000 USMLE-style multiple-choice questions. We load the dataset from HuggingFace (GBaker/MedQA-USMLE-4-options) and filter for clinical vignette questions using heuristic rules — specifically, we retain questions whose stems contain age-gender patterns (e.g., “45-year-old woman”) or phrases like “presents with” or “comes to,” filtering out pure pharmacology or anatomy questions that lack clinical narratives. This yields a pool of candidate cases from which we sample.

Our dataset construction pipeline, implemented entirely by us, uses Gemini 2.0 Flash via the Vertex AI API as an LLM assistant for two generation steps:

Stage Decomposition. For each selected vignette, we prompt Gemini to restructure the question into 4 sequential evidence stages that simulate how a clinician would receive information over time. Stage 1 presents demographics and chief complaint only. Stage 2 adds vital signs and physical examination findings. Stage 3 introduces laboratory results and imaging findings. Stage 4 provides confirmatory information such as culture results, biopsy findings, or specialist evaluations. If the original question lacks stage-4 information, the model synthesizes a clinically plausible confirmatory finding consistent with the ground truth diagnosis. We instruct the model to never reveal the diagnosis in any stage. only clinical findings are included.

Counterfactual Generation. For each base case, we generate one counterfactual variant by prompting Gemini to introduce a single discriminative finding at Stage 2 or Stage 3 that argues strongly against the ground truth diagnosis. The altered finding must be clinically plausible and meaningfully discriminative, sufficient that a competent clinician would move the ground truth diagnosis down or off their differential entirely. All stages not being altered remain identical to the base case. Each counterfactual is annotated with the altered stage number, a description of the alteration, and the diagnosis it rules out.

Outputs are validated for structural completeness (valid JSON format with all four stages present). Cases that fail structural validation are discarded and replaced. We do not perform automated clinical validation of counterfactual plausibility; we rely on the LLM’s instruction-following to produce clinically meaningful alterations. This is a limitation we discuss in Section 7.

Our final dataset consists of $N_{base} = 60$ base cases and $N_{cf} = 60$ counterfactual variants (one per base case). All dataset construction code is implemented by us; only the underlying LLM API is external.

4.3 Evaluation Metrics

We define three process-oriented evaluation metrics adapted from standard information-theoretic and consistency measures, designed specifically for evaluating eliminative reasoning in sequential differential diagnosis. Our metrics adapt standard tools to a novel setting. Evidence Sensitivity applies KL divergence [3] to rank-derived distributions using reciprocal rank weighting from information retrieval [4]. Elimination Monotonicity relates to consistency evaluation in LLMs [5], and Cross-Step Consistency to self-consistency in chain-of-thought reasoning [6], though both are reformulated here for sequential eliminative reasoning over staged clinical evidence. The specific metric definitions and

their application to differential diagnosis under counterfactual perturbation are our contributions. All metric computation code is implemented by us.

Elimination Monotonicity. Let E_t denote diagnoses eliminated at stage t . The elimination monotonicity rate is:

$$RE = \frac{|\{d \in E_t \mid d \in D_{t+1}\}|}{|E_t|} \quad (1)$$

We compute this for stages $t = 1, 2, 3$ and average over all stages with at least one elimination. Lower values indicate more consistent eliminative reasoning. This metric is computed on base cases only.

Evidence Sensitivity. We convert ranked differentials into probability distributions via rank-based weighting:

$$P(d_i) = \frac{1/\text{Rank}(d_i)}{\sum_j 1/\text{Rank}(d_j)} \quad (2)$$

and measure belief revision between original and counterfactual cases using KL divergence:

$$D_{KL}(P_{base} \| P_{cf}) = \sum_d P_{base}(d) \log \frac{P_{base}(d)}{P_{cf}(d)} \quad (3)$$

We compute this at each stage from the altered stage onward (since earlier stages are identical) and average. Larger divergence indicates stronger belief revision under discriminative evidence. We apply epsilon smoothing ($\varepsilon = 10^{-8}$) to handle diagnoses that appear in one distribution but not the other.

Cross-Step Consistency. For each diagnosis eliminated at stage t , we check whether it remains absent from the differential at all subsequent stages ($t+1$ through 4). The metric is the proportion of all such elimination events where the diagnosis stays excluded. Higher values indicate more consistent reasoning. This metric is computed on base cases only.

4.4 Models

We evaluate four models spanning frontier commercial LLMs: GPT-4o (via OpenAI API), Claude Sonnet 4.5 (via Anthropic API), Gemini 2.0 Flash and Gemini 2.5 Pro (via Vertex AI). As our goal is to evaluate and compare existing models rather than propose a new model, we do not define a trained baseline; the models themselves serve as points of comparison against each other. A simpler alternative analysis would be to rank models by final-answer accuracy alone (i.e. the standard MedQA protocol), but as our results demonstrate, accuracy differences carry no information about reasoning process quality and cannot detect the failure modes our metrics are designed to surface.

5 Experiments

5.1 Data

We evaluate on our constructed dataset of 60 staged clinical vignettes from MedQA-USMLE (Section 3.2). The final dataset comprises 60 base cases and 60 counterfactual variants, for 120 total vignettes.

5.2 Evaluation Method

We evaluate models along two axes: diagnostic accuracy and process quality; full metric definitions and equations are provided in Section 3.3. Diagnostic accuracy is measured by Top-1 and Top-5 accuracy at the final stage (Stage 4), indicating whether the ground-truth diagnosis appears as the top-ranked or top-five-ranked item in the final differential, respectively. Process-oriented metrics measure the quality of eliminative reasoning across stages: Elimination Monotonicity ($\downarrow$) captures the proportion of eliminated diagnoses that reappear in a subsequent differential; Evidence Sensitivity ($\uparrow$) measures the average KL divergence between base and counterfactual rank-derived distributions at and after the alteration stage; and Cross-Step Consistency ($\uparrow$) measures the proportion of eliminated diagnoses that remain absent from all subsequent differentials.

5.3 Experimental Details

Each model receives an identical zero-shot prompt instructing it to produce (1) a ranked differential of exactly five diagnoses with one-sentence justifications per item, and (2) an elimination list of diagnoses ruled out by the evidence so far. All models are evaluated in zero-shot chain-of-thought mode with temperature set to 0 and greedy decoding. The prompt is applied stage-by-stage with cumulative context: at stage t , the model receives all evidence from stages 1 through t . Outputs are parsed automatically via regex to extract ranked differentials and elimination lists for downstream metric computation.

We note that KL divergence values differ substantially from the preliminary figures reported at the milestone stage (range 0.19–0.43). This is attributable to two factors: expansion from an initial pilot of 20 cases to the full 60-case dataset, and a correction to the KL computation that applies epsilon smoothing ($\varepsilon = 10^{-8}$) over the union of diagnoses across base and counterfactual differentials, resolving numerical instability in the earlier unsmoothed implementation.

5.4 Results

Table 1 reports performance across four frontier LLMs on our staged differential diagnosis dataset. Two primary quantitative patterns emerge.

Table 1: Results on the staged differential diagnosis dataset ($N = 60$ base cases per model). $\uparrow$ = higher is better, $\downarrow$ = lower is better. **Bold** = best per column.

Model	Top-1 $\uparrow$	Top-5 $\uparrow$	Elim. Mono. $\downarrow$	KL Div. $\uparrow$	Consistency $\uparrow$
Claude Sonnet 4.5	0.200	0.200	0.015	9.526	0.985
Gemini 2.0 Flash	0.400	0.433	0.050	7.908	0.950
Gemini 2.5 Pro	0.217	0.217	0.037	10.098	0.956
GPT-4o	0.333	0.383	0.026	7.026	0.971

Accuracy and process quality are dissociated. Gemini 2.0 Flash achieves the highest Top-1 accuracy (0.400) and Top-5 accuracy (0.433), yet its elimination monotonicity (0.050) and cross-step consistency (0.950) are the weakest among all evaluated models. Conversely, Claude Sonnet 4.5 achieves the lowest elimination monotonicity (0.015) and the highest cross-step consistency (0.985) while producing the lowest Top-1 accuracy (0.200). This dissociation confirms that diagnostic accuracy is an insufficient proxy for clinically coherent reasoning: under the standard MedQA evaluation protocol, which serves as our implicit baseline, these process-level failures would be entirely invisible, as all models would be ranked solely by final-answer accuracy.

Greater evidence sensitivity does not predict better diagnostic or eliminative performance. Gemini 2.5 Pro produces the largest distributional shift in response to counterfactual perturbations (KL: 10.098), yet achieves the second-lowest Top-1 accuracy (0.217) and lower cross-step consistency than both GPT-4o and Claude Sonnet 4.5. Conversely, GPT-4o produces the smallest KL divergence (7.026) while achieving the second-highest accuracy (0.333) and strong consistency (0.971). This inverse relationship between evidence sensitivity and downstream performance suggests that large belief revisions under counterfactual pressure may reflect instability in the model’s ranking rather than genuine discriminative reasoning.

The relatively low absolute accuracy values (Top-1 range: 0.200–0.400) were expected given the open-ended generation setting: unlike standard MedQA evaluation where models select from four provided answer choices, our framework requires free-text diagnosis generation with no candidate list, substantially increasing task difficulty. The magnitude of the accuracy–consistency dissociation, however, was more pronounced than anticipated, with the highest-accuracy model simultaneously producing the weakest elimination behavior.

6 Analysis

6.1 Dissociation Between Accuracy and Eliminative Reasoning

The most striking qualitative pattern is the systematic dissociation between final-answer accuracy and the quality of intermediate reasoning. Claude Sonnet 4.5 exemplifies this pattern: its differentials are highly stable across stages, and diagnoses once eliminated are rarely reintroduced. But the model is slower to commit the correct diagnosis to rank 1, distributing probability mass across a broader set of plausible alternatives throughout later stages. Gemini 2.0 Flash exhibits the inverse pattern: it aggressively and correctly promotes the ground-truth diagnosis early, but does so inconsistently with respect to its own elimination decisions, producing the highest elimination monotonicity score. These patterns suggest that elimination stability and top-rank precision reflect distinct reasoning behaviors that do not co-vary across models. A concrete example of this behavior appears in `case_001`, discussed in Section 6.2 and Section 6.3.

6.2 Counterfactual Sensitivity: Rank Shifts vs. Genuine Belief Revision

A consistent pattern across models is that counterfactual alterations produce reliable surface-level rank reorderings without consistently triggering explicit elimination updates. In `case_001` (ground truth: Cystitis), the counterfactual introduces marked CVA tenderness at Stage 2, a finding that argues strongly against uncomplicated cystitis and toward pyelonephritis. All four models respond with strong rank revision: Pyelonephritis moves to rank 1 in the counterfactual condition across all models, producing non-zero KL divergence. However, only Claude Sonnet 4.5 explicitly eliminates “Uncomplicated cystitis” from the differential at Stage 3, with the justification that “presence of costovertebral angle tenderness indicates upper urinary tract involvement.” Gemini 2.0 Flash, GPT-4o, and Gemini 2.5 Pro all continue to list Cystitis in their counterfactual differentials through Stage 4 without any elimination entry, despite the discriminative evidence introduced at Stage 2 directly arguing against it. This dissociation between rank revision and elimination behavior represents a failure mode that accuracy-only evaluation would not detect: all four models produce high KL divergence on this case, yet three fail to update their elimination lists accordingly.

6.3 Elimination Monotonicity: Failure Mode Analysis

Elimination monotonicity failures occur most frequently at the Stage 2 to Stage 3 transition, coinciding with the introduction of laboratory and imaging findings that partially overlap semantically with previously excluded diagnoses. This pattern is illustrated concretely by `case_001` (ground truth: Cystitis), in which all four models correctly eliminate Pyelonephritis at Stage 2 given absent fever and CVA tenderness. At Stage 3, however, three models reintroduce it without referencing their prior elimination: Gemini 2.0 Flash (rank 3: “can sometimes present without CVA tenderness”), GPT-4o (rank 3: “could still suggest early or atypical pyelonephritis”), and Gemini 2.5 Pro (rank 2, persisting through Stage 4). Claude Sonnet 4.5 is the only model that maintains the elimination throughout.

Qualitative inspection suggests two primary failure modes. First, models occasionally eliminate a diagnosis prematurely and then correctly restore it when confirmatory findings arrive, a case where reintroduction may be clinically appropriate, pointing to a limitation of our metric in distinguishing justified from erroneous reintroductions. Second, and more commonly, models reintroduce a previously excluded diagnosis without grounding the reversal in any new evidence, suggesting that elimination behavior is not maintained as a persistent belief state but recomputed locally at each prompt call. Gemini 2.5 Pro additionally reintroduces Interstitial Cystitis at Stage 2 despite eliminating it at Stage 1 as “inconsistent with the patient’s acute one-day history,” a rationale unchanged by any intervening evidence. GPT-4o exhibits a related degradation in the counterfactual condition, eliminating “Fever” and “Fever-related infections” — symptoms rather than diagnoses — suggesting the elimination mechanism can collapse into surface-level keyword matching under altered evidence conditions.

6.4 Model Comparison Summary

Taken together, the four evaluated models occupy distinct positions in the accuracy–consistency tradeoff space. Gemini 2.0 Flash is the strongest accuracy-oriented reasoner but the least consistent eliminator. Claude Sonnet 4.5 is the most consistent eliminator but the least accurate. GPT-4o presents

the most balanced profile across all five metrics. Gemini 2.5 Pro produces the largest distributional response to counterfactual evidence but does not translate this into accuracy gains, suggesting that large belief revisions under counterfactual perturbation and improved diagnostic precision are not equivalent.

7 Conclusion

We introduced a process-oriented evaluation framework for measuring eliminative reasoning in LLM-based differential diagnosis. By constructing 60 staged clinical vignettes with counterfactual variants from MedQA-USMLE, we exposed systematic failures in belief revision invisible to accuracy-based benchmarks. Diagnostic accuracy and process quality are substantially dissociated across models: the strongest accuracy-oriented systems exhibit the weakest elimination consistency, while the most stable eliminators rank the correct diagnosis less aggressively. Qualitative analysis further reveals that rank changes under counterfactual evidence frequently occur without explicit elimination updates, indicating that surface-level reranking and genuine belief revision are distinct phenomena. Together, these findings demonstrate that process-oriented metrics surface clinically meaningful reasoning failures that accuracy alone cannot detect.

Our work has three main limitations. First, open-ended diagnosis generation substantially increases task difficulty and likely depresses absolute accuracy; future work should incorporate embedding-based soft matching to reduce noise from lexically distinct but semantically equivalent diagnosis names. Second, our pipeline does not track the implied ground truth of counterfactual variants, conflating failure to revise beliefs with correct revision toward an unrecognized new diagnosis; future work should elicit counterfactual ground truths at construction time. Third, dataset construction relies entirely on Gemini 2.0 Flash without human clinical validation, meaning incoherent counterfactuals may produce artificially low KL divergence scores; expert review of a representative subset is needed to bound this noise. Finally, both domain-adapted models we attempted to evaluate (BioGPT and Meditron) failed to produce parseable outputs under zero-shot chat prompting, suggesting that structured evaluation frameworks of this kind require instruction-following capability not yet reliably present in smaller domain-adapted models.

Team contributions (Required for multi-person team)

Tatiana and Joy contributed equally to this project. Tatiana led the design and implementation of the evaluation framework. Joy led the design and implementation of the evaluation metrics. Both team members contributed to the data construction pipeline, overall project design, experimental methodology, and final report editing.

References

- [1] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023.
- [2] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- [3] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [4] Nick Craswell. Mean reciprocal rank. In *Encyclopedia of Database Systems*. Springer, 2009.
- [5] Yanai Elazar, Nora Kassner, Shauli Shaber, Yoav Goldberg, and Daniel Roth. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.
- [6] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

- [7] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837, 2022.
- [8] Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*, 2023.
- [9] Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*, 2020.
- [10] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.
- [11] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421, 2021.

A Appendix (optional)

A Dataset Construction Prompts

A.1 Stage Decomposition System Prompt

The following system prompt was used to instruct Gemini 2.0 Flash to decompose each MedQA-USMLE vignette into four sequential evidence stages:

```

You are a medical education expert. Your job is to take a
USMLE-style clinical vignette and restructure it into 4 sequential
evidence stages that simulate how a clinician would receive
information over time in a real encounter.

Rules:
  • Stage 1: Chief complaint, basic demographics only. No exam, no
    labs.
  • Stage 2: Vital signs and physical examination findings.
  • Stage 3: Laboratory results and/or imaging findings.
  • Stage 4: Any final/confirmatory information (e.g. culture
    result, biopsy, specialist finding). If the original question
    has no stage-4 info, synthesize a clinically plausible one
    consistent with the ground truth diagnosis.
  • Each stage must be a self-contained clinical narrative paragraph
    (not a list).
  • Do NOT reveal the diagnosis in any stage. Only provide
    findings.
  • The ground_truth is the correct answer diagnosis -- extract
    it cleanly (e.g. "Community-acquired pneumonia", not "A.
    Community-acquired pneumonia").

Respond ONLY with valid JSON. No markdown fences. No preamble.
Schema:
{
  "ground_truth": "<diagnosis name>",
  "stage_1": "<chief complaint + demographics>",
  "stage_2": "<vitals + physical exam>",
  "stage_3": "<labs + imaging>",
  "stage_4": "<confirmatory finding>"
}

```

A.2 Counterfactual Generation System Prompt

The following system prompt was used to instruct Gemini 2.0 Flash to generate one counterfactual variant per base case:

You are a medical education expert designing counterfactual clinical cases for evaluating AI diagnostic reasoning.

Given a staged clinical vignette and its ground truth diagnosis, you will generate ONE counterfactual variant. A counterfactual introduces a single discriminative finding at Stage 2 or Stage 3 that strongly argues AGAINST the ground truth diagnosis, while keeping all other stages as close to the original as possible.

Requirements:

- Change only ONE stage (either stage_2 or stage_3).
- The changed finding must be clinically plausible and meaningfully discriminative.
- The change should cause a competent clinician to move the ground truth diagnosis DOWN in their differential -- ideally off the list entirely.
- Stages not being changed should remain identical to the original.
- altered_stage: the stage number being changed (2 or 3).
- alteration_description: a one-sentence summary of what changed and why it argues against the diagnosis.
- rules_out: the ground truth diagnosis that this finding argues against.

Respond ONLY with valid JSON. No markdown fences. No preamble.

Schema:

```
{
  "cf_id_suffix": "a",
  "altered_stage": 2,
  "alteration_description": "<one sentence: what changed and why
it argues against ground truth>",
  "rules_out": "<ground truth diagnosis>",
  "stage_1": "<identical to base or adjusted>",
  "stage_2": "<original or modified>",
  "stage_3": "<original or modified>",
  "stage_4": "<original or adjusted>"
}
```

B Evaluation Prompt

The following prompt was issued to all four models at each stage. {stage_num}, {total_stages}, and {case_text} are filled in programmatically at inference time, where case_text contains all evidence from stages 1 through t concatenated cumulatively.

You are a medical reasoning assistant. This is a research evaluation; do NOT give medical advice.

TASK:

Given the clinical case information revealed so far, output:

- (1) A ranked differential diagnosis list of EXACTLY 5 items.
- (2) A short 1-sentence justification for each item.
- (3) An ELIMINATED list of diagnoses ruled out by the evidence so far (0 or more).

OUTPUT FORMAT (follow exactly -- no extra text before or after):

DIFFERENTIAL:

1. <diagnosis>: <1 sentence justification>

2. <diagnosis>: <1 sentence justification>
3. <diagnosis>: <1 sentence justification>
4. <diagnosis>: <1 sentence justification>
5. <diagnosis>: <1 sentence justification>

ELIMINATED:

- <diagnosis>: <short reason>

CASE (Stage {stage_num} of {total_stages}):
{case_text}

CONSTRAINTS:

- Use concise standard medical diagnosis names.
- Do not include treatment or management.
- If uncertain, still list the 5 most plausible diagnoses.