Original Paper

Benchmark Integrity and Reasoning-Trace Errors in Medical Question Answering With Large Language Models: Mixed Methods Study With Sparse Autoencoders

Jialin Liu[1,2*], MD; Siru Liu[3,4*], PhD; Adam Wright[3,5], PhD

[1]Department of Medical Informatics, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[2]Department of Otolaryngology-Head and Neck Surgery, West China Hospital of Sichuan University, Chengdu, Sichuan, China
[3]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, TN, United States
[4]Department of Computer Science, Vanderbilt University, Nashville, TN, United States
[5]Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, United States
*these authors contributed equally

Corresponding Author:

Siru Liu, PhD
Department of Biomedical Informatics
Vanderbilt University Medical Center
2525 West End Ave
Nashville, TN 37203
United States
Phone: 1 615 936 6867
Email: siru.liu@vumc.org

Abstract

Background: Large language models (LLMs) show promise for enhancing diagnostic accuracy and clinical decision-making. However, prevailing evaluations rely on examination-based benchmarks such as MedQA. Furthermore, the internal mechanisms driving both correct and incorrect reasoning in LLMs remain poorly understood, limiting opportunities for targeted improvement.

Objective: This study aimed to investigate failure modes of reasoning-based LLMs in medicine by (1) auditing the integrity of the MedQA benchmark, (2) developing a clinically informed taxonomy of reasoning errors across multiple major LLMs, and (3) testing a mechanistic intervention using sparse autoencoders (SAEs) to modulate reasoning characteristics and improve accuracy in medical question answering benchmarks.

Methods: We evaluated OpenAI o1 on the MedQA and cross-referenced incorrect answers against original source platforms to identify benchmark flaws including missing figures and postrelease ambiguity corrections. For the 37 confirmed model failures remaining after exclusion of flawed items, we developed a reasoning error taxonomy through iterative inductive coding by 2 independent reviewers (JL and SL) and validated it on three major LLMs (ie, OpenAI GPT-4.5, OpenAI o3-mini, and DeepSeek-R1). We then trained an SAE on the DeepSeek-R1-Distill-Llama-8B model using MedQA-derived reasoning traces. Reasoning-specific features were identified using ReasonScore and subjected to activation steering at 2 strengths. Model accuracy, reasoning trace length, and hallucination metrics were measured across MedQA, MedMCQA, and PubMedQA. Hallucinations were evaluated using an LLM-as-a-judge (OpenAI GPT-5-mini) and validated on a stratified manual sample of 100 claims.

Results: Forty-one percent of initial OpenAI o1 errors reflected benchmark problems, including missing figures (22%) and ambiguities subsequently corrected on the source platforms (19%). Neither OpenAI o1 nor OpenAI o3-mini explicitly flagged these flawed items, while GPT-5.2 identified a small subset, suggesting that question-integrity recognition remains limited and model-dependent. Among the 37 confirmed errors, our taxonomy classified failures into four categories: Information Synthesis Errors, Therapeutic Decision Errors, Diagnostic Reasoning Errors, and Foundational Principle Errors. Activation steering of reasoning-specific SAE features improved accuracy on MedQA and PubMedQA, with a consistent positive trend on MedMCQA. The greatest gains were observed at steering strength 2 (MedQA: 0.568-0.597 and PubMedQA: 0.708-0.739). Steering also increased reasoning-trace length substantially, with no significant correlation between verbosity and accuracy. Five functional feature categories were identified, with alignments to the error taxonomy.

Conclusions: These findings reveal two distinct sources of unreliability in medical LLM evaluation: benchmark-level integrity gaps that misattribute model failure and recurrent model-level reasoning patterns potentially amenable to mechanistic correction. Notably, the benchmark issues identified here do not reflect static flaws in the original source platforms, which have since corrected many problematic items, but rather a failure to propagate those corrections to derived benchmarks. The alignment between SAE-identified feature categories and the error taxonomy further suggests that reasoning failures reflect structured internal processes that can be targeted at the feature level.

J Med Internet Res 2026;28:e90061; doi: 10.2196/90061

Keywords: large language model; medical question answering; sparse autoencoders; benchmark evaluation; reasoning errors; clinical decision support; mechanistic interpretability; activation steering; hallucination; artificial intelligence in medicine

Introduction

The integration of artificial intelligence (AI) into medicine presents a profound opportunity to enhance diagnostic accuracy and clinical decision-making. At the forefront of this transformation are large language models (LLMs), which, after training on vast corpora of medical literature and clinical guidelines, have demonstrated remarkable capabilities [1]. These models show promise in diverse applications, from critiquing clinical decision support systems to drafting patient communications and summarizing clinical notes [2-6].

Recent LLM releases differ less in core architecture than in training incentives and inference-time policies that govern how much computation a model allocates before producing an answer [7]. Reasoning models are optimized (eg, via reinforcement learning) to perform better on multistep problems and often generate longer intermediate rationales, whereas general-purpose chat models may produce shorter justifications by default. Importantly, generated rationales are not guaranteed to be faithful readouts of causal reasoning [8].

However, the prevailing methods for evaluating these advanced models remain a critical weak point. Current assessments predominantly rely on examination-based benchmarks, with a mere 5% using real patient data [9,10]. MedQA, a widely used benchmark derived from the United States Medical Licensing Examination (USMLE), exemplifies this issue [11]. Developed in 2020 from questions scraped from public websites, its multiple-choice format often tests rote memorization over nuanced clinical judgment [12]. While some have questioned the validity of using examination questions to evaluate clinical AI [13], there has been little systematic investigation into the intrinsic quality and reliability of the benchmark questions themselves. Prior work has shown that LLM performance drops substantially when multiple-choice shortcuts are disrupted, further underscoring the need to examine benchmark quality beyond surface-level accuracy [14].

A crucial factor in the performance of reasoning LLMs is their explicit, stepwise thought process, often refined through reinforcement learning. This can be optimized using outcome-based rewards, which assess only the final answer's correctness [15], or process-based rewards, which evaluate the entire reasoning sequence [16]. In deterministic fields such as mathematics, outcome-based rewards suffice. In medicine, however, the clinical logic must be as sound as the conclusion, making the analysis of reasoning errors essential for developing safe and effective clinical LLMs. While reasoning errors in LLMs have been studied in general domains, clinically grounded taxonomies of reasoning failures across multiple frontier models remain limited. Recent work in mechanistic interpretability has demonstrated that sparse autoencoders (SAEs) can identify and steer interpretable features in LLMs [17,18], but this approach has not yet been applied to medical reasoning.

We address 3 linked gaps in evaluating and improving medical question-answering (QA) LLMs. First, we audit MedQA test-set integrity by reconciling incorrect model outputs with the original source platforms to identify missing modalities and postrelease corrections. Second, focusing on questions that remain well-specified, we develop an initial clinically informed taxonomy of observable reasoning-trace failures and examine how its distribution varies across several frontier models. Third, we test whether a mechanistic intervention, steering SAE features enriched around reasoning-trace tokens, can measurably shift accuracy and reasoning-trace properties across multiple medical QA datasets.

Methods

Reasoning Error Analysis and Taxonomy Development

We selected the MedQA test dataset (N=1273) of USMLE-style questions as our primary benchmark. We used OpenAI o1, a representative state-of-the-art reasoning model, to generate answers and a corresponding chain-of-thought reasoning process for each question. LLM prompts for answer and reasoning generation are listed in Multimedia Appendix 1. We identified incorrectly answered questions by comparing the model's output with the provided answer key.

For each incorrect answer, we located the original question in its source examination bank (Medbullets, AMBOSS, or Lecturio). This allowed us to identify discrepancies, such as missing figures or subsequent question updates made by the source platforms to correct ambiguities. During reconciliation with source platforms, we also tracked whether models explicitly signaled uncertainty or unanswerability, such as noting missing information or required figures, or identifying ambiguity in the question stem. To assess whether newer models demonstrate improved question-integrity recognition, we additionally tested GPT-5.2-pro-2025-12-11 on the identified flawed items post hoc. This source reconciliation

was feasible for MedQA because its questions could be traced to actively maintained platforms. Analogous auditing was not performed for MedMCQA or PubMedQA, as MedMCQA was compiled from various open websites and books without a single maintained source platform, and PubMedQA derives questions and answers from PubMed abstracts rather than from a curated examination bank.

After excluding questions with these external quality issues, we developed an error taxonomy using an inductive coding approach [19]. Two coders (SL and JL) independently analyzed the reasoning processes for a subset of 5 incorrect answers, using open coding to identify error themes. The coders then met to discuss their findings and create a consensus-based coding guideline. This iterative process was repeated with new batches of 5 questions until all reasoning traces for the incorrectly answered questions were coded.

To validate the taxonomy, we assessed the performance of other advanced LLMs (OpenAI GPT-4.5, OpenAI o3-mini, and DeepSeek-R1) on the same set of questions. For models that do not natively output their reasoning, we used a step-by-step chain-of-thought prompt to elicit it. All models were accessed via their respective application programming interfaces (APIs). The specific API versions queried were OpenAI o1 (o1-preview, queried January 11, 2025), OpenAI o3-mini (o3-mini-2025-01-31, queried March 25, 2025), GPT-4.5 (gpt-4.5-preview-2025-02-27, queried March 25, 2025), and DeepSeek-R1 (queried March 25, 2025). Throughout the remainder of the paper, these models are referred to as o1, o3-mini, GPT-4.5, and DeepSeek-R1, respectively. Each question was evaluated in a single-shot setting with temperature set to 0. Other decoding parameters (top-p, seed, max tokens) were not explicitly set and followed the provider default. We then applied our taxonomy to classify the failure modes in each model's incorrect reasoning traces and manually compared reasoning traces from different models. Notably, DeepSeek-R1 was deliberately included as one of the validation models because the SAE analysis model (DeepSeek-R1-Distill-Llama-8B) was derived from it via knowledge distillation.

SAE Development and Feature Steering

To further interpret the reasoning process quantitatively, we used an SAE, a method to decompose an LLM's internal activations into a sparse set of interpretable features [17]. The core idea is that a neural network's internal representations are dense and difficult to interpret directly, as individual neurons often respond to multiple unrelated concepts (polysemanticity). An SAE addresses this by learning to reconstruct the model's activations through a higher-dimensional but sparse intermediate representation, where each dimension (or feature) ideally corresponds to a single interpretable concept. SAEs were chosen over other interpretability methods (eg, attention analysis and probing classifiers) because they enable both interpretation and intervention: once reasoning-relevant features are identified, their activations can be directly modified to test effects on model behavior.

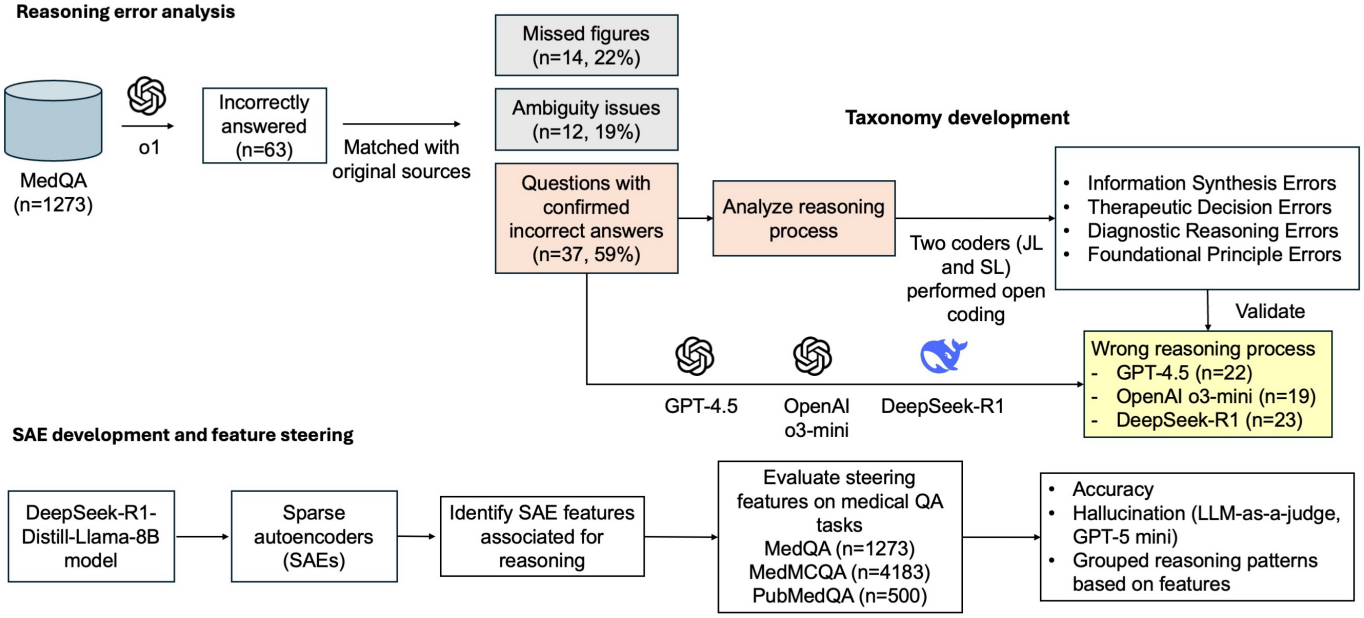
We trained an SAE using the DeepSeek-R1-Distill-Llama-8B model. This model was chosen for its open-source availability, strong reasoning capabilities, and moderate size, which balanced performance with experimental efficiency. The training data included general conversation logs (LMSys-Chat-1M) [20], general reasoning traces (OpenThoughts-114k) [21], and reasoning traces leading to correct answers in the training dataset of MedQA generated by DeepSeek-R1. Notably, the MedQA training and testing sets are different, and the SAE evaluation was conducted exclusively on the test set. Following established methods, we extracted activations from the 19th layer of the model to train the SAE [18].

We used ReasonScore, a quantitative metric, to identify reasoning-specific features from a trained SAE [18]. The metric works by identifying features that activate most strongly around predefined reasoning words within a fixed-width context window and using an entropy penalty to penalize features that activate only on limited reasoning words. After identifying the top 100 features with the highest ReasonScore, we manually reviewed their activations and logits using an SAE dashboard [22]. For the top 15 features, 2 reviewers (SL and JL) independently examined the top-activating contexts, activation patterns, and logit effects for each feature, then met to discuss and reconcile their functional labels through consensus. This process yielded 5 functional categories, which were then compared with the independently derived error taxonomy.

To test the impact of these features, we used activation steering [23,24]. This method involves intentionally modifying the model's reasoning by adding a positive bias to the activations of top-performing features (using strengths of 2 and 4). Boosting a feature's activation encourages the model to more strongly incorporate that feature's specific concept into its subsequent processing. We measured the accuracy of the steered model on three medical benchmarks: MedQA, which served as the primary benchmark across all study components; and 2 other widely used benchmarks, MedMCQA (multiple-choice questions from Indian medical entrance examinations) and PubMedQA (yes, no, or maybe questions derived from PubMed abstracts). MedMCQA and PubMedQA were included to test whether steering effects generalize beyond the dataset used for taxonomy development and SAE training. The 26 MedQA items identified as having benchmark integrity issues in the audit phase were retained in the SAE evaluation, as these items affect all steering conditions equally and represent approximately 2% of the test set. In addition, we used an LLM-as-a-judge (OpenAI GPT-5-mini) to assess hallucinations in the reasoning process, categorizing them as factual hallucinations (contradicting known facts) or input hallucinations (contradicting the question prompt) and assigning a severity score from 1 to 3. LLM-generated labels were validated manually on a stratified sample of 100 claims across datasets and severity levels. Expert validation was performed by JL, a physician with an MD degree and clinical experience as an attending in otolaryngology–head and neck surgery, as well as research experience in medical informatics. All

disagreement cases were independently reviewed by SL, an investigator with a PhD in biomedical informatics and expertise in clinical AI evaluation. The severity correlation was computed at the claim level between the LLM judge's severity rating and the human annotator's independently assigned rating using the same rubric. To assess feature generalizability, we ranked features by their poststeering accuracy (exact match) within each dataset and steering strength, selected the top K features (for K=5, 10, 15, and 20), and computed the set intersection count between each dataset pair. Then, we selected the top 15 features, manually compared reasoning patterns before and after steering, and grouped them based on their potential function in reasoning. We used the *SAELens* package to train the SAE [25], and the *SAEDashboard* package to visualize identified features and their impact in Python 3 [22]. The overall project pipeline is shown in Figure 1.

Figure 1. Overview of the project pipeline for reasoning-error taxonomy development. LLM: large language model; QA: question answering.



Statistical Analysis

The same questions were evaluated under all steering conditions, yielding paired observations. For accuracy (binary outcome), McNemar tests were used for pairwise comparisons between steering strengths. For reasoning token counts, Wilcoxon signed-rank tests were used. For hallucination counts, where the per-question hallucination set differs across conditions, Mann-Whitney *U* tests were used. All pairwise tests were corrected for multiple comparisons using the Holm method. Chi-square tests were used to compare the distribution of hallucination types across conditions. Pearson correlation coefficients were calculated between reasoning token length and performance metrics (accuracy, hallucination frequency, and severity), with statistical significance assessed using *t* tests on the correlation coefficients. All statistical analyses were performed using Python 3 (Python Software Foundation).

Ethical Considerations

This study used publicly available, deidentified data and did not require ethics approval.

Results

Benchmark Quality Analysis

The OpenAI o1 model answered 63 of 1273 questions incorrectly, for an accuracy of 95%. Upon cross-referencing these 63 questions with their original examination banks, we found that 14 (22%) questions were missing figures that were essential for arriving at the correct answer (Textbox 1). For instance, one question about a patient's pressure-volume loop explicitly mentioned a figure (Figure 2), while another describing a patient with retrosternal burning implicitly relied on a figure showing diffuse lung fibrosis to make the correct diagnosis. A complete list of questions that should include figures is provided in Multimedia Appendix 2.

An additional 12 (19%) questions contained ambiguities that have since been corrected in the source question banks (Textbox 1). For example, a question about a 3-month-old infant with a holosystolic murmur was updated to include "tetany is noted when taking the blood pressure," a hallmark sign that clarifies the diagnosis as 22q11 deletion syndrome (DiGeorge syndrome) rather than fetal alcohol syndrome. In another instance, a question involving a 78-year-old woman was revised. The version in MedQA implied that acute cognitive changes suggested a symptomatic urinary tract infection (warranting treatment), whereas the updated version, by noting that the patient continued to exhibit symptoms

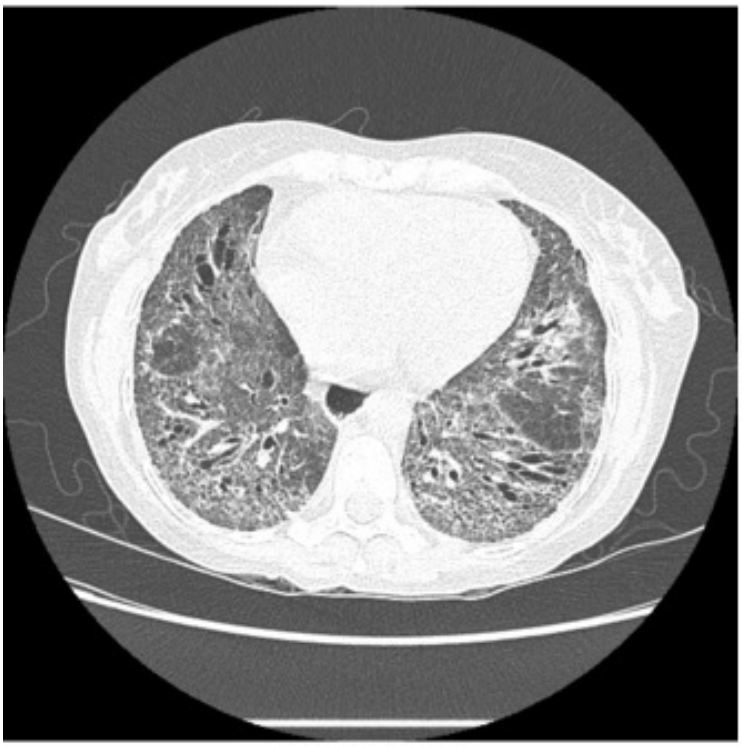
consistent with Alzheimer dementia, supported a decision for no treatment. Twelve questions (19%) that had serious ambiguity issues, resulting in answers that differ from those provided as correct, are listed in Multimedia Appendix 3.

Textbox 1. Examples of MedQA questions missing original figures and examples of questions with ambiguity issues. Bold text indicates updated information in the current online questions compared with the original questions in MedQA.

Example 1. Figure explicitly mentioned and needed to answer correctly, but not present in the benchmark.
Question: A 72-year-old woman is admitted to the intensive care unit for shortness of breath and palpitations. A cardiac catheterization is performed, and measurements of the left ventricular volume and pressure at different points in the cardiac cycle are obtained. The patient's pressure-volume loop (gray) is shown with a normal pressure-volume loop (black) for comparison. Which of the following is the most likely underlying cause of this patient's symptoms?
Options:
A: Mitral valve regurgitation
B: Increased systemic vascular resistance
C: Increased ventricular wall stiffness
D: Impaired left ventricular contractility
Example 2. Figure not explicitly mentioned and needed to answer correctly, but not present in the benchmark.
Question: A 43-year-old woman presents with complaints of retrosternal burning associated with eating. It has persisted for the past several years but has been getting worse. Her past medical history is unknown and this is her first time seeing a doctor. She states she is otherwise healthy and review of systems is notable for episodic hand pain that is worse in the winter as well as a chronic and severe cough with dyspnea, which she attributes to her smoking. Her temperature is 97.7 °F (36.5 °C), blood pressure is 174/104 mmHg, pulse is 80/min, respirations are 22/min, and oxygen saturation is 92% on room air. Physical exam is notable for a young appearing woman with coarse breath sounds. Laboratory studies and urinalysis are ordered and currently pending. Which of the following is the pathophysiology of this patient's chief complaint?
Options:
A: Decreased lower esophageal tone
B: Esophageal fibrosis
C: Increased lower esophageal tone
D: Spastic cricopharyngeal muscle
Note: The original question included a chest computed tomography image (Figure 2), which is necessary to answer this question.
Example 3. A question with ambiguity issues.
MedQA question: A 3-month-old infant is brought to her pediatrician because she coughs and seems to have difficulty breathing while feeding. In addition, she seems to have less energy compared to other babies and appears listless throughout the day. She was born by cesarean section to a G1P1 woman with no prior medical history and had a normal APGAR score at birth. Her parents say that she has never been observed to turn blue. Physical exam reveals a high-pitched holosystolic murmur that is best heard at the lower left sternal border. The most likely cause of this patient's symptoms is associated with which of the following abnormalities?
Options:
A: 22q11 deletion
B: Deletion of genes on chromosome 7
C: Lithium exposure in utero
D: Maternal alcohol consumption
Updated question: A 3-month-old infant is brought to her pediatrician because she coughs and seems to have difficulty breathing while feeding. In addition, she seems to have less energy compared to other babies and appears listless throughout the day. She was born by cesarean section to a G1P1 woman with no prior medical history and had a normal APGAR score at birth. Her parents say that she has never been observed to turn blue. Physical exam reveals a high-pitched holosystolic murmur that is best heard at the lower left sternal border. Tetany is noted when taking the blood pressure. The most likely cause of this patient's symptoms is associated with which of the following abnormalities?
MedQA question: A 78-year-old woman presents to the office for an annual health check-up with her family physician accompanied by her daughter. She has no complaints during this visit but her daughter states that she is having difficulty locating objects such as the television remote, car keys, and her purse. Her medical history is significant for Alzheimer's dementia, coronary artery disease, diabetes mellitus, hypothyroidism, congestive heart failure, osteoarthritis, and centrilobular emphysema. The patient takes memantine, atorvastatin, metformin, levothyroxine, lisinopril, aspirin, albuterol, and ipratropium. The patient's vitals are within normal limits today. Physical exam reveals an elderly female in no acute distress, oriented to person, place, and year, but not to month or day of the week. She has a 3/6 holosystolic murmur at the left sternal border along with an S3 gallop. There are mild crackles at the lung bases. The remainder of the exam is normal. A previous urine culture reports growth of > 100,000 CFU of Enterobacter. Urinalysis findings are offered below:
Leukocyte esterase positive

WBCs 50-100 cell/HPF
Nitrites positive
RBCs 2 cell/HPF
Epithelial cells 2 cell/HPF
Urine pH 5.7
Which of the following is the most appropriate next step?
A: TMP-SMX
B: Nitrofurantoin
C: Levofloxacin
D: No treatment is necessary
Updated question: A 78-year-old woman presents with her daughter for an annual health check-up with her family physician. The patient has no complaints during this visit but her daughter states that the patient continues to manifest symptoms consistent with her known diagnosis of Alzheimer dementia, including having difficulty locating objects such as her television remote, car keys, and purse. Her other medical history is significant for coronary artery disease, diabetes mellitus, hypothyroidism, congestive heart failure, osteoarthritis, and centrilobular emphysema. The patient takes memantine, atorvastatin, metformin, levothyroxine, lisinopril, aspirin, albuterol, and ipratropium. The patient's vitals are within normal limits. Physical exam reveals no acute distress. She is oriented to person, place, and year, but not to month or day of the week. She has a 3/6 holosystolic murmur at the left sternal border along with an S3 gallop. There are mild crackles at the lung bases. The remainder of the exam is normal. Two urine cultures performed in one week each reported growth of > 100,000 CFU of Enterobacter. Urinalysis findings are presented below:
Leukocyte esterase positive
WBCs 50-100 cell/HPF
Nitrites positive
RBCs 2 cell/HPF
Epithelial cells 2 cell/HPF
Urine pH 5.7
Which of the following is the most appropriate next step?
A: TMP-SMX
B: Nitrofurantoin
C: Levofloxacin
D: No treatment is necessary

Figure 2. Chest computed tomography image originally accompanying example 2 in MedQA. This figure was present in the source question bank but absent from the MedQA benchmark dataset.



Across the 26 flawed MedQA items, neither o1 nor o3-mini explicitly identified missing figures or ambiguity in any case, suggesting that earlier-generation reasoning models lack reliable question-integrity recognition. In contrast, GPT-5.2-pro-2025-12-11 flagged 5 questions as missing required figures and identified 1 question with unresolved ambiguity, indicating an emerging but still limited ability to detect underspecified or unanswerable items. Overall, explicit detection of unanswerable or underspecified questions was infrequent, indicating that question-integrity recognition remains limited and model-dependent for medical questions.

Taxonomy of Reasoning Errors

After excluding questions with missing figures or ambiguity, 37 questions (14 from USMLE step 1 examinations and 23 from USMLE step 2 and step 3 examinations) remained for our error analysis. From these, we identified four major error categories: Information Synthesis Errors, Therapeutic Decision Errors, Diagnostic Reasoning Errors, and Foundational Principle Errors. The definitions and examples for each category are detailed in Textbox 2.

Textbox 2. Taxonomy of reasoning errors with definitions and examples. COPD: chronic obstructive pulmonary disease.

Information Synthesis Errors:

- Misjudgment of Clinical Feature Importance:
  - Definition: overemphasizes irrelevant details while ignoring critical ones.
  - Example 1: in a lung cancer case involving a 76-year-old man with COPD and asbestos exposure, the reasoning process overemphasized the patient's asbestos exposure and pleural plaques, while neglecting symptoms such as weight loss and anemia, and his 60-pack-year smoking history that was more indicative of malignancy.
  - Example 2: in a case of chronic abdominal pain with multiorgan involvement, the focus was incorrectly placed on a recent impetigo episode and the possibility of poststreptococcal glomerulonephritis, thereby overlooking systemic evidence pointing to a chronic condition such as secondary amyloidosis.

Therapeutic Decision Errors:

- Misapplication of Evidence-Based Guidelines:
  - Definition: fails to select or correctly apply clinical guidelines.
  - Example 1: the reasoning process overemphasized the use of an adjunctive drug for tumor lysis syndrome while neglecting intravenous hydration, the primary strategy for managing this condition.
  - Example 2: the reasoning process incorrectly selected griseofulvin despite guidelines recommending itraconazole as the first-line therapy for patients with the fungal infection *tinea corporis*.
- Inadequate Dynamic Risk-Benefit Assessment:
  - Definition: fails to weigh evolving clinical risks and benefits.

- Example 1: in a case of circulatory electrolyte imbalance in hepatic encephalopathy, the reasoning process prioritized correcting hypoglycemia while neglecting hypokalemia.
- Example 2: in a case of hemoptysis with thrombolytic therapy, the reasoning process did not adequately account for the bleeding risks associated with thrombolytic therapy, underestimating the potential for life-threatening hemorrhage.

- Misinterpretation of Pharmacologic Mechanisms:
- Definition: misunderstands a medication's mechanism of action.
- Example 1: in a patient with seasonal allergies, the reasoning process incorrectly identified the drug's mechanism as competitive blockade of muscarinic receptors (used in asthma management) instead of recognizing that the appropriate decongestant works as an α-adrenergic agonist.
- Example 2: in a gout prophylaxis scenario, pancytopenia was erroneously attributed to colchicine toxicity rather than correctly identifying the mechanism of a xanthine oxidase inhibitor and its potential interaction with immunosuppressive agents.

- Premature Cognitive Closure:
- Definition: a reasoning error in which the model quickly settles on a diagnosis or management plan without sufficiently exploring alternative explanations or contributing factors.
- Example: in a case of erectile dysfunction in a patient on selective serotonin reuptake inhibitors with significant vascular risk factors, the reasoning process prematurely closed off further investigation (eg, nocturnal penile tumescence testing) that could better clarify the underlying cause.

Diagnostic Reasoning Errors:

- Failure to Integrate Pathophysiological Mechanisms:
- Definition: fails to synthesize and apply key pathophysiological principles, resulting in misattribution or misunderstanding of clinical findings.
- Example 1: when analyzing the relevant option, the model failed to link the characteristic facial anomalies (eg, low-set ears and retrognathia) with the Potter sequence.
- Example 2: the model overlooked that achlorhydria in a vasoactive intestinal peptide-secreting tumor (VIPoma) leads to impaired iron absorption, failing to integrate this pathophysiological mechanism into its diagnostic reasoning.

- Deviation from Prioritized Diagnostic Protocols:
- Definition: initiates treatment before completing necessary diagnostic workups.
- Example 1: when evaluating a 2-month-old infant with signs of head trauma and suspicious injury patterns, the model prioritized nonmedical interventions (eg, involving social services) over obtaining an urgent head computed tomography (CT) scan, delaying critical diagnostic evaluation.
- Example 2: in the initial management of a transient ischemic attack, if the CT scan is normal, a CT angiogram is indicated to further characterize the cerebral vessels. The reasoning process, however, wrongly opted for heparin therapy.

Foundational Principle Errors:

- Misapplication of Ethical Principles:
- Definition: fails to follow proper ethical protocols.
- Example 1: in the case of a surgical complication, the reasoning process failed to follow proper communication protocols by not discussing the complication with the attending physician before reporting it.
- Example 2: in the situation where a daughter refuses consent due to an abuse history, the reasoning process neglected to involve the appropriate legal guardian, instead opting to seek immediate court intervention without first contacting the next of kin.

- Misinterpretation of Statistical Concepts:
- Definition: misinterprets statistical principles.
- Example 1: the reasoning process failed to recognize that the *P* value is computed under the assumption that the null hypothesis is true.
- Example 2: the reasoning process failed to differentiate between lead-time bias and measurement bias.

Comparison of Errors Across Different LLMs

We tested other leading models on the 37 challenging questions. Accuracy was 49% for o3-mini, 41% for GPT-4.5, and 38% for DeepSeek-R1. Reasoning process lengths varied significantly, from an average of 1319 characters for o3-mini to 11,055 characters for DeepSeek-R1.

An analysis of the reasoning traces revealed that different models used highly distinct problem-solving strategies, especially on questions requiring multistep synthesis (Figures 3 and 4). For example, one question required understanding that immunoglobulin A (IgA) deficiency is associated with celiac disease and that these patients are typically deficient in fat-soluble vitamins.

As shown in Figure 3, the o1, GPT-4.5, and o3-mini models attempted a linear process but failed for different

reasons. They first identified the key patient details, then developed an initial differential diagnosis, and finally assessed the available answer options. The o1 model recognized the possibility of IgA deficiency and evaluated each option, linking option A (Hemolytic anemia and ataxia) with vitamin E deficiency. However, it did not connect the vitamin E deficiency back to IgA deficiency. In the GPT-4.5 process, the analysis of option A led directly to its association with certain hereditary disorders, such as Friedreich ataxia and other conditions involving neurological and hematological symptoms, resulting in the elimination of what was actually the correct choice. In contrast, the o3-mini model exhibited a shortcut: it noted, "Although the answer choices do not directly mention IgA deficiency or anti-IgA antibodies, the only option referring to antibodies (which in this context are responsible for transfusion reactions) is option D: anti-A, B, or O antibodies in the serum."

Meanwhile, DeepSeek-R1's process was fundamentally different (Figure 4). It used an iterative, 3-round analysis, repeatedly reassessing the problem, highlighting a completely distinct cognitive architecture from the other models. The detailed reasoning processes for each model are listed in Multimedia Appendix 4.

When tested on questions that o1 failed, the models displayed varied error patterns (Table 1). Information Synthesis Errors were most frequent in OpenAI o1 (11 errors), while Therapeutic Decision Errors were common across all models. These results highlight model-specific differences in reasoning failure modes.

Figure 3. Examples of identified errors in the reasoning traces of OpenAI o3-mini, OpenAI o1, and GPT-4.5. BP: blood pressure; IgA: immunoglobulin A; IV: intravenous.

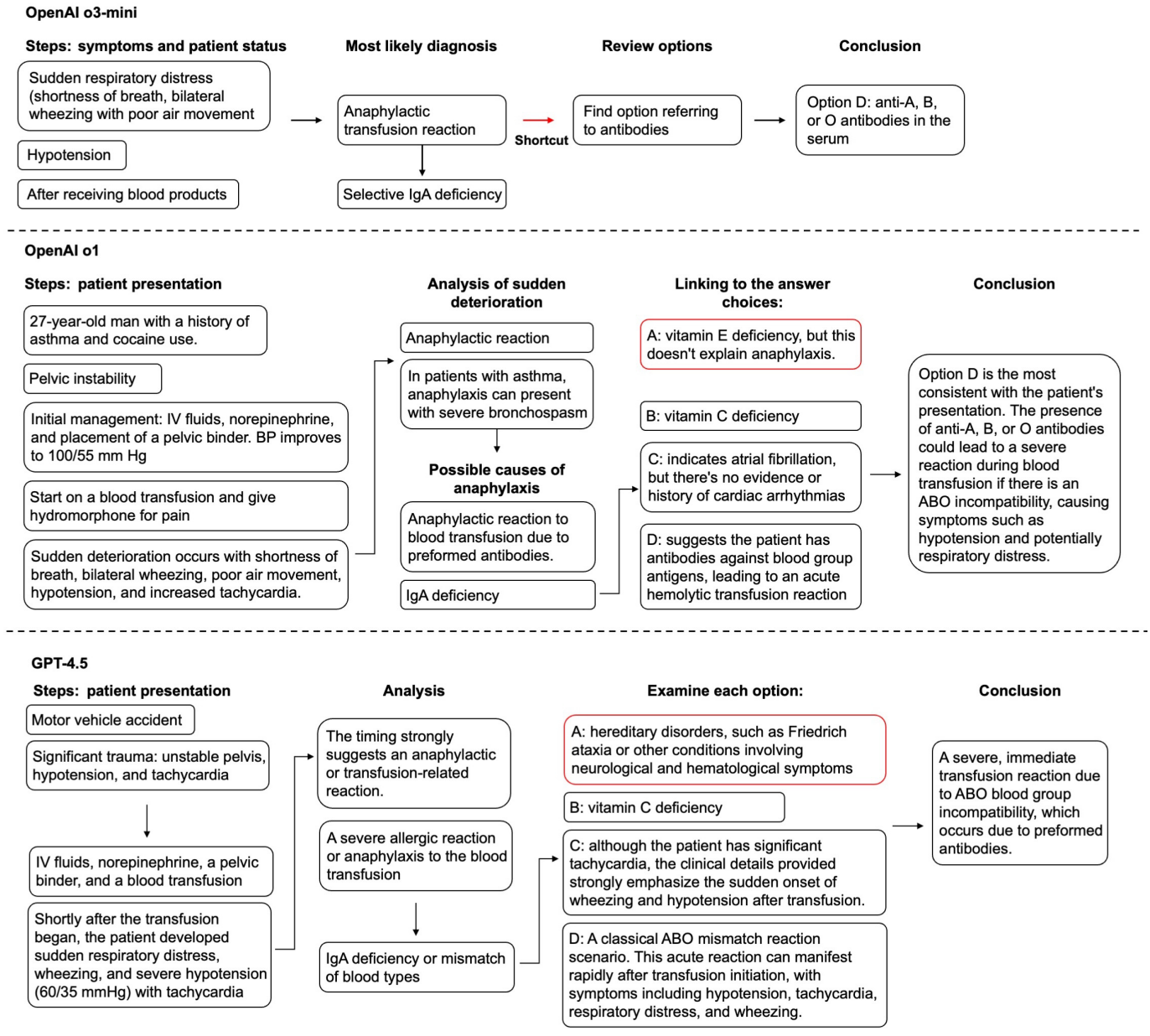


Figure 4. Examples of identified errors in the reasoning traces of DeepSeek-R1. DIC: disseminated intravascular coagulation; GCS: Glasgow Coma Scale; IgA: immunoglobulin A; IV: intravenous; TRALI: transfusion-related acute lung injury.

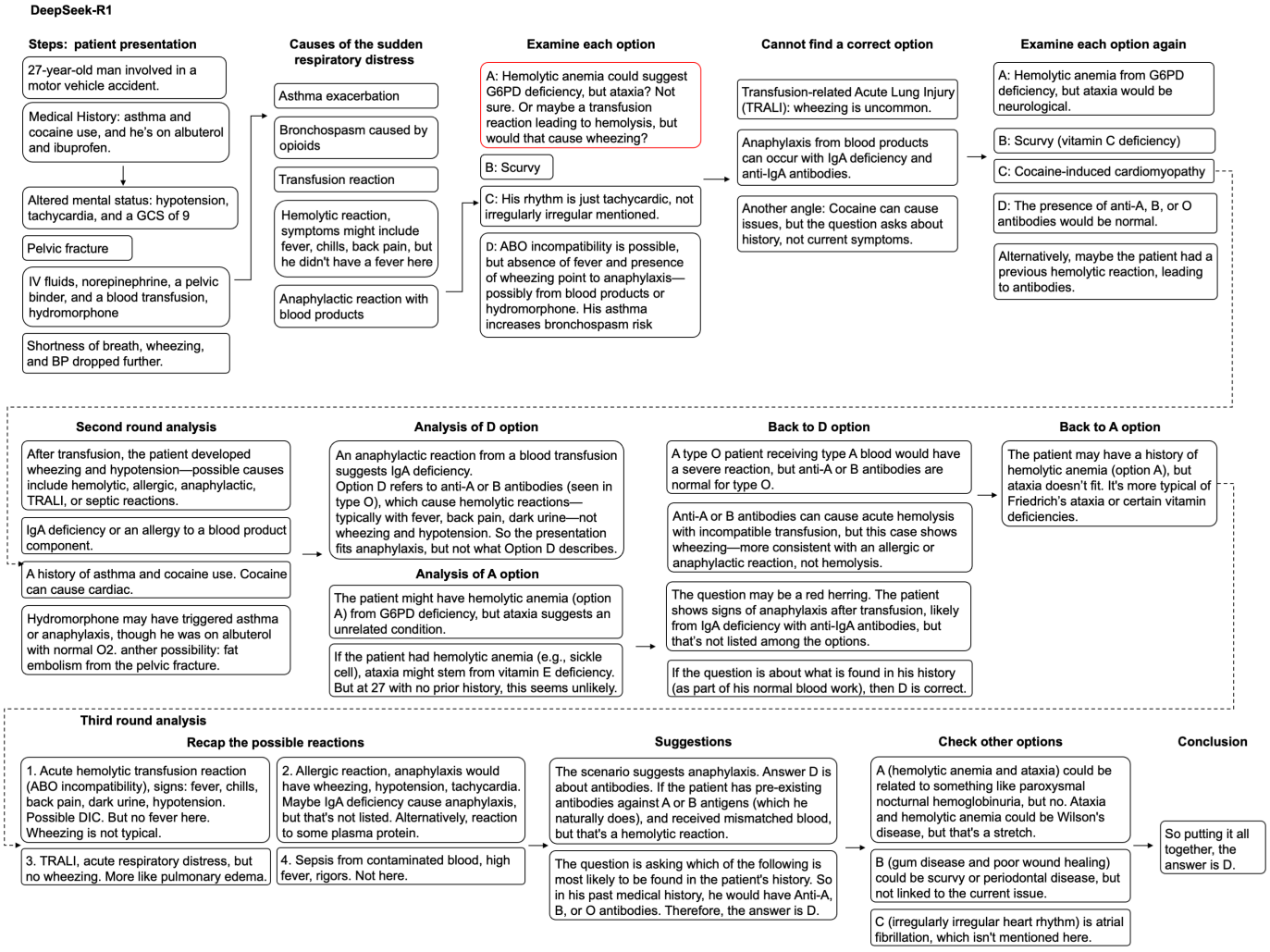


Table 1. Distribution of error types across 4 large language models for 37 challenging questions.

Error category and subcategory	OpenAI o1	GPT-4.5	OpenAI o3-mini	Deepseek-R1
Information Synthesis Errors	11	5	4	3
Misjudgment of Clinical Feature Importance	11	5	4	3
Therapeutic Decision Errors	11	6	7	8
Misapplication of Evidence-Based Guidelines	4	0	0	2
Inadequate Dynamic Risk-Benefit Assessment	3	1	1	1
Misinterpretation of Pharmacologic Mechanisms	3	4	3	2
Premature Cognitive Closure	1	1	3	3
Diagnostic Reasoning Errors	9	8	4	8
Failure to Integrate Pathophysiological Mechanisms	5	6	3	7
Deviation from Prioritized Diagnostic Protocols	4	2	1	1
Foundational Principle Errors	6	3	4	4
Inappropriate Ethical Decision-Making	4	2	3	2
Misinterpretation of Statistical Concepts	2	1	1	2
Total	37	22	19	23

Feature Steering via SAE

Steering with reasoning-specific features enhanced model accuracy, with significant gains on MedQA (Table 2; χ^2_1=10.9; P=.002) and PubMedQA (χ^2_1=18.6; P<.001) and a consistent positive trend on MedMCQA (χ^2_1=3.9; P=.15). The most substantial gains were observed at a moderate steering strength of 2, which increased MedQA accuracy from 0.568 to 0.597 (95% CI 0.584-0.610) and PubMedQA accuracy from 0.708 to 0.739 (95% CI 0.722-0.756). This performance improvement, however, coincided with a

significant increase in the verbosity of the model's reasoning traces. At a steering strength of 4, the average reasoning token count nearly doubled for MedQA (Wilcoxon W=6,010,329; $P<.001$) and tripled for PubMedQA (Wilcoxon W=257,830; $P<.001$). The intervention had a limited effect on hallucination frequency; a significant increase was observed for MedMCQA at strength 2 (Mann-Whitney U=1,364,912; $P=.008$), with no significant changes for MedQA (U=481,828; $P=.99$) or PubMedQA (U=33,695; $P=.99$). Notably, the type of hallucination remained consistent within each dataset regardless of steering. Chi-square analysis revealed a significant shift in the distribution of hallucination types for MedMCQA ($\chi^2_2=8.1$; $P=.02$), but not for MedQA ($\chi^2_2=1.4$; $P=.49$) or PubMedQA ($\chi^2_2=1.1$; $P=.58$). Across all conditions, factual hallucinations were the most common type in MedMCQA and MedQA, whereas input hallucinations predominated in PubMedQA (Table 4). Finally, no significant correlations were identified between reasoning length and key performance metrics such as accuracy, hallucination count, or hallucination severity. These correlations were computed at the level of individual questions within each steering condition. Expert validation achieved 96% precision, with the 4 disagreement cases representing conservative overflagging of borderline medical claims rather than clear errors. The claim-level Pearson correlation between the LLM judge and human annotator severity ratings was 0.86.

Table 2. Model performance across steering strengths on 3 medical question-answering (QA) benchmarks. Values are reported as point estimates with 95% CIs computed from question-level paired data.

Dataset and strength	Accuracy (95% CI)	Reasoning tokens (95% CI)	Hallucinations[a] (95% CI)	Factual hallucinations[b] (95% CI)	Severity[c] (0-3) (95% CI)
MedMCQA					
None	0.505 (0.503-0.507)	1287 (1237-1338)	2.2 (2.154-2.247)	2.15 (2.104-2.197)	2.199 (2.186-2.212)
2	0.522 (0.508-0.536)	1500 (1294-1706)	2.378 (2.284-2.472)	2.284 (2.191-2.377)	2.194 (2.168-2.221)
4	0.525 (0.511-0.538)	2457 (2065-2850)	2.163 (2.066-2.260)	2.107 (2.011-2.203)	2.203 (2.176-2.230)
MedQA					
None	0.568 (0.565-0.570)	1682 (1587-1777)	2.605 (2.515-2.695)	2.488 (2.398-2.577)	2.265 (2.241-2.289)
2	0.597 (0.584-0.610)	1886 (1671-2101)	2.557 (2.456-2.658)	2.456 (2.356-2.556)	2.250 (2.223-2.277)
4	0.589 (0.575-0.602)	2912 (2520-3303)	2.658 (2.554-2.762)	2.531 (2.429-2.633)	2.225 (2.198-2.252)
PubMedQA					
None	0.708 (0.704-0.712)	670 (652-687)	1.064 (0.954-1.174)	0.384 (0.309-0.459)	1.837 (1.785-1.888)
2	0.739 (0.722-0.756)	861 (731-991)	1.08 (0.970-1.190)	0.398 (0.319-0.477)	1.787 (1.735-1.839)
4	0.737 (0.719-0.754)	2187 (1696-2677)	0.962 (0.855-1.069)	0.336 (0.265-0.407)	1.734 (1.677-1.790)

[a]Average count of hallucination claims per reasoning trace.
[b]Average count of factual hallucination claims per reasoning trace.
[c]0-3 rating by a large language model judge (GPT-5-mini) and higher = worse.

Functional Analysis of Reasoning Features

Feature overlap analysis revealed distinct generalization patterns across medical domains. At strength 4, four features (IDs: 10602, 29040, 37660, and 56984) consistently ranked in the top 20 across all 3 datasets, suggesting robust utility. MedQA and MedMCQA showed the strongest feature similarity, sharing 13 features in their top 20 at strength 4, while PubMedQA demonstrated more domain-specific feature requirements. Higher SAE strength (4 vs 2) consistently improved cross-dataset feature generalization. Feature overlap heatmaps are shown in Multimedia Appendix 5. Manual analysis of the top 15 features allowed us to group them into 5 functional categories that aligned with our error taxonomy (Multimedia Appendix 6). These categories included (1) cue-weighting calibration and distractor suppression, which focuses the model on pivotal clinical data; (2) protocol alignment, which steers outputs toward guideline-consistent procedures; (3) mechanistic grounding, which

connects decisions to core pathophysiological or pharmacological principles; (4) rule/criteria enforcement, which ensures the precise application of formal definitions such as Light's criteria; and (5) evidence synthesis and question reframing, which helps the model correctly interpret the core question being asked. Several features were multifunctional, contributing to more than one reasoning category (eg, feature 56,984).

Discussion

Principal Findings

In this study, we conducted a mixed methods analysis to audit the integrity of a widely used medical benchmark, characterize LLM reasoning errors, and test a mechanistic intervention using SAEs. Our findings reveal that a significant portion of apparent model failures on MedQA are attributable to intrinsic data flaws, true reasoning errors can be categorized into a clinically relevant taxonomy, and steering reasoning-specific features can effectively improve LLM accuracy on medical QA benchmarks.

Rethinking Benchmarks: The Problem Lies in the Questions

An important finding of our work is that 41% of the initial incorrect answers generated by OpenAI o1 were due to flawed benchmark questions, including 14 missing necessary figures and 12 containing ambiguities that have since been updated in their source question banks. These integrity findings are specific to MedQA; we did not perform the same source reconciliation for MedMCQA or PubMedQA.

This audit indicates that MedQA, as instantiated in widely used benchmark distributions, contains a nontrivial fraction of items whose fidelity to the original sources is compromised (eg, missing figures or subsequently corrected ambiguity). These issues can confound model evaluation by attributing errors to models that may instead reflect dataset drift or modality loss introduced during benchmark construction.

More broadly, these findings underscore a structural risk in deriving medical benchmarks from examination question banks without preserving all clinically relevant modalities. In clinical problem-solving, visual data (eg, imaging, waveforms, or figures) are often integral to reasoning, and their omission fundamentally alters the task being evaluated. Under these conditions, a model's apparent success on a flawed question may plausibly arise from pattern matching rather than robust clinical reasoning. This concern is consistent with prior work showing substantial performance drops when multiple-choice shortcuts are disrupted, such as replacing the correct option with "none of the above," which reduced accuracy by 8%-38% [14].

As the field relies heavily on such benchmarks to gauge progress, our findings serve as a critical call to action for improved validation and management of medical benchmarks. Unlike static domains such as mathematics or coding, medical knowledge is constantly evolving (eg, clinical practice guidelines are updated and new drugs emerge) [26]. To remain relevant, medical benchmarks must be treated as dynamic resources, continuously aligned with the current state of practice. This necessitates a move toward more rigorously curated, version-controlled, and multimodally complete datasets, supported by automated methods for ongoing validation. In addition, when maintainers of source question banks identify and correct errors, a corresponding process must be in place to update the benchmark promptly. Without such reliable evaluation tools, we risk overestimating the capabilities of current models and misdirecting development efforts.

A Taxonomy of Reasoning Failures: From "What" to "Why"

By isolating the 37 model errors, we developed a 4-category taxonomy that shifts the focus from whether a model is wrong to why it is wrong. The prevalence of Information Synthesis Errors (particularly for OpenAI o1) and Therapeutic Decision Errors across all tested models highlights vulnerabilities in processes central to clinical practice: weighing evidence, applying guidelines, and performing dynamic risk-benefit assessments. The observation that different state-of-the-art models exhibit distinct error profiles (Table 1) suggests that their underlying architectures and training data instill unique cognitive biases. For instance, some models may be prone to premature closure, while others struggle to integrate complex pathophysiological mechanisms. This granular, qualitative understanding of failure modes is essential for identifying high-risk scenarios and is a necessary prerequisite for building safer, more reliable systems.

A Mechanistic Path Toward Safer AI: From Observation to Intervention

Another contribution of this study is the bridge between the qualitative error taxonomy and the quantitative SAE intervention. By identifying and steering reasoning-specific features, we not only improved accuracy on multiple medical benchmarks but also uncovered the functional roles of these features. Some clear alignments were identified: feature groups such as "prioritizes critical information and filters out irrelevant data" and "protocol alignment" mechanistically address the error categories of "Misjudgment of Clinical Feature Importance" and "Deviation from Prioritized Diagnostic Protocols," respectively (Multimedia Appendix 6). These results provide an initial proof of concept that interpretable, reasoning-specific SAE features can be modulated to shift overall accuracy and that the resulting feature categories conceptually correspond to several of the error types in our taxonomy. They do not, however, demonstrate that steering individual feature subgroups causally corrects the specific error categories with which they conceptually align; such a claim would require single-feature or feature-group interventions on a larger error-annotated corpus drawn from the same model, which we identify as a key direction for future work.

This intervention also revealed a trade-off. Steering with strengthened reasoning-specific features increased accuracy,

particularly on MedQA and PubMedQA, but also significantly lengthened the reasoning traces, more than doubling the token count in some cases. Notably, however, the relationship between steering strength and accuracy was nonmonotonic: increasing strength from 2 to 4 more than doubled token counts across all 3 benchmarks, yet accuracy slightly declined in some cases (eg, MedQA: 0.597-0.589). This pattern, combined with the absence of significant correlations between reasoning trace length and any performance metric, indicates that improved accuracy is not a simple byproduct of increased verbosity or test-time computation. Rather, it likely results from the targeted activation of specific, high-value reasoning pathways. The distinction between group-level and individual-level effects is important here: while steering shifts the mean accuracy and mean trace length upward relative to baseline, within a given condition, longer traces do not predict correct answers for individual questions, further suggesting that the benefit is feature-specific rather than length-dependent. We note that prompt-based interventions designed to elicit longer reasoning operate at the input level rather than directly modifying internal representations and therefore target a different causal pathway than activation steering; exploring such comparisons remains a valuable direction for future work. More broadly, the increased token cost associated with steering presents a practical design challenge. This work represents a proof of concept; potential strategies for mitigating this trade-off in future implementations could include selective steering for high-uncertainty cases or adaptive strength calibration.

Strengths and Limitations

This study's primary strength lies in its novel mixed methods approach, which combines a qualitative analysis of benchmark integrity and LLM reasoning errors with a quantitative, mechanistic intervention using SAEs. This provides a uniquely holistic view of the challenges and opportunities in improving LLM performance on medical reasoning benchmarks.

However, several limitations remain. First, we used different models across study stages: the error taxonomy was developed on frontier models (eg, OpenAI o1), while SAE training and steering were conducted on a single open-source distilled model (DeepSeek-R1-Distill-Llama-8B). Because distilled student models can develop internal representations distinct from their teachers [27,28], the alignment between our taxonomy and the 8B model's SAE features is suggestive rather than direct mechanistic evidence. Consequently, these identified features are not directly transferable to closed-source models. Second, our error taxonomy is based on 37 incorrect responses from the o1 model, which may not capture the full spectrum of reasoning failures. Specifically, it misses "silent failures," cases where incorrect reasoning coincidentally yields a correct answer. Characterizing these hidden errors through expert review of complete reasoning traces remains an important direction for future work. Third, using an LLM-as-a-judge to evaluate hallucinations is scalable but introduces potential biases. Our human validation revealed conservative overflagging by the LLM judge. Furthermore, comprehensively estimating recall (false negatives) was infeasible because of the massive data volume and specialized clinical expertise required. As a result, our reported hallucination counts should be interpreted as lower-bound estimates. Finally, there are technical and evaluative constraints regarding the SAE intervention. Our evaluation relies on QA benchmarks, which do not fully replicate the dynamic, unstructured nature of real-world clinical decision-making. Technically, the SAE training was limited to a single layer (layer 19) and faced common interpretability challenges, such as incomplete resolution of polysemanticity and the difficulty of completely disentangling reasoning from factual features. Future work should compare this approach against steering random, nonreasoning features and extend evaluations to real patient data.

Future Work

Our findings point toward several key directions for future research. First, there is an urgent need for the community to develop validation and management methods for maintaining medical benchmarks, ensuring fair and reproducible evaluations. Second, the reasoning-specific features identified by SAEs could be integrated into more advanced training paradigms. For example, they could inform process-based reward models in reinforcement learning to explicitly penalize flawed reasoning pathways identified in our taxonomy, potentially training models that are both accurate and efficient. In addition, the error taxonomy developed here could inform the design of targeted medical test cases that intentionally trigger specific failure modes, serving both as a validation tool for the taxonomy and as a stress-testing framework for evaluating model robustness across known error categories. For clinical translation, our work underscores that accuracy is an insufficient metric for safety and reliability. The ability to audit, understand, and even steer a model's reasoning process, as demonstrated here, represents a critical step toward building the "glass-box" systems necessary to earn clinician trust and ensure patient safety [29].

Conclusion

This study identified that 41% of initial model errors on MedQA reflected benchmark integrity issues, including missing figures and subsequently corrected ambiguities, rather than true model failures. Among the 37 confirmed reasoning errors, inductive analysis yielded a 4-category taxonomy (Information Synthesis, Therapeutic Decision, Diagnostic Reasoning, and Foundational Principle Errors) that revealed distinct failure profiles across 4 frontier LLMs. Steering reasoning-specific SAE features significantly improved accuracy on MedQA and PubMedQA, with a consistent positive trend on MedMCQA, while also increasing reasoning trace length, with no significant correlation between verbosity and performance. These findings demonstrate that medical LLM evaluation is constrained by flawed benchmarks and that reasoning failures follow identifiable, model-specific patterns with the potential for mechanistic correction via feature steering.

Acknowledgments

The authors used generative artificial intelligence solely to identify key citations in the literature review. These systems were not involved in data collection, data analysis, study design, or paper drafting. The authors have reviewed the paper and accept full responsibility for it.

Funding

This work was supported by the National Institutes of Health (NIH) grant R00LM014097-02. The NIH had no role in the design and conduct of the study; the collection, management, analysis, and interpretation of the data; the preparation, review, or approval of the paper; and the decision to submit the paper for publication.

Data Availability

The prompts are listed in Multimedia Appendix 1. MedQA questions with missing figures and ambiguous MedQA questions updated in source question banks are reported in Multimedia Appendices 2 and 3. The datasets used in this study are publicly available: MedQA (GitHub), MedMCQA (GitHub), and PubMedQA (GitHub). The training data for the sparse autoencoder are publicly available: LMSYS-Chat-1M (Hugging Face) and OpenThoughts-114k (Hugging Face). The code for the official ReasonScore implementation is available on GitHub.

Authors' Contributions

SL contributed to conceptualization, data curation, formal analysis, investigation, methodology, software development, and writing of the original draft. JL contributed to conceptualization, data curation, investigation, methodology, and writing of the original draft. AW contributed to the conceptualization, review, and editing of the paper.

Conflicts of Interest

None declared.

Multimedia Appendix 1

Large language model prompt for answer and reasoning generation.
[DOCX File (Microsoft Word File), 17 KB-Multimedia Appendix 1]

Multimedia Appendix 2

MedQA questions with missing figures.
[DOCX File (Microsoft Word File), 87 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Ambiguous MedQA questions updated in source banks.
[DOCX File (Microsoft Word File), 35 KB-Multimedia Appendix 3]

Multimedia Appendix 4

Example of a complex question failed by all large language models (OpenAI o1, OpenAI o3-mini, and OpenAI GPT-4.5, and DeepSeek-R1).
[DOCX File (Microsoft Word File), 28 KB-Multimedia Appendix 4]

Multimedia Appendix 5

Feature overlap heatmaps across medical datasets. For each dataset and steering strength, features were ranked by their poststeering accuracy (exact match) on that dataset. The heatmaps show the set intersection count of top-K feature lists between each dataset pair, that is, |Top-K(A) ∩ Top-K(B)|, for K=5, 10, 15, and 20 at steering strengths of 2 and 4. Diagonal entries equal K.
[DOCX File (Microsoft Word File), 238 KB-Multimedia Appendix 5]

Multimedia Appendix 6

Functional analysis of top 15 reasoning-specific sparse autoencoder features.
[DOCX File (Microsoft Word File), 20 KB-Multimedia Appendix 6]

References

1. Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. arXiv. Preprint posted online on Mar 4, 2022. [doi: 10.48550/arXiv.2203.02155]
2. Liu S, Wright AP, Patterson BL, et al. Using AI-generated suggestions from ChatGPT to optimize clinical decision support. J Am Med Inform Assoc. Jun 20, 2023;30(7):1237-1245. [doi: 10.1093/jamia/ocad072] [Medline: 37087108]

3. Tai-Seale M, Baxter SL, Vaida F, et al. AI-generated draft replies integrated into health records and physicians' electronic communication. JAMA Netw Open. Apr 1, 2024;7(4):e246565. [doi: 10.1001/jamanetworkopen.2024.6565] [Medline: 38619840]
4. Liu S, McCoy AB, Wright AP, et al. Leveraging large language models for generating responses to patient messages—a subjective analysis. J Am Med Inform Assoc. May 20, 2024;31(6):1367-1379. [doi: 10.1093/jamia/ocae052]
5. Van Veen D, Van Uden C, Blankemeier L, et al. Clinical text summarization: adapting large language models can outperform human experts. arXiv. Preprint posted online on Sep 14, 2023. [doi: 10.48550/arXiv.2309.07430]
6. Liu S, McCoy AB, Wright AP, et al. Why do users override alerts? Utilizing large language model to summarize comments and optimize clinical decision support. J Am Med Inform Assoc. May 20, 2024;31(6):1388-1396. [doi: 10.1093/jamia/ocae041] [Medline: 38452289]
7. DeepSeek-AI, Guo D, Yang D, et al. DeepSeek-R1: incentivizing reasoning capability in LLMs via reinforcement learning. arXiv. Preprint posted online on Jan 22, 2025. [doi: 10.48550/arXiv.2501.12948]
8. Xie Y, Wu J, Tu H, et al. A preliminary study of o1 in medicine: are we closer to an AI doctor? arXiv. Preprint posted online on Sep 23, 2024. [doi: 10.48550/arXiv.2409.15277]
9. Bedi S, Liu Y, Orr-Ewing L, et al. Testing and evaluation of health care applications of large language models: a systematic review. JAMA. Jan 28, 2025;333(4):319-328. [doi: 10.1001/jama.2024.21700] [Medline: 39405325]
10. Liu S, McCoy AB, Wright A. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. J Am Med Inform Assoc. Apr 1, 2025;32(4):605-615. [doi: 10.1093/jamia/ocaf008] [Medline: 39812777]
11. Jin D, Pan E, Oufattole N, Weng WH, Fang H, Szolovits P. What disease does this patient have? A large-scale open domain question answering dataset from medical exams. Appl Sci (Basel). 2020;11(14):6421. [doi: 10.3390/app11146421]
12. Raji ID, Daneshjou R, Alsentzer E. It's time to bench the medical exam benchmark. NEJM AI. Jan 23, 2025;2(2). [doi: 10.1056/AIe2401235]
13. Liu J, Wang C, Liu S. Utility of ChatGPT in clinical practice. J Med Internet Res. Jun 28, 2023;25:e48568. [doi: 10.2196/48568] [Medline: 37379067]
14. Bedi S, Jiang Y, Chung P, Koyejo S, Shah N. Fidelity of medical reasoning in large language models. JAMA Netw Open. Aug 1, 2025;8(8):e2526021. [doi: 10.1001/jamanetworkopen.2025.26021] [Medline: 40779272]
15. Yu F, Gao A, Wang B. OVM, outcome-supervised value models for planning in mathematical reasoning. arXiv. Preprint posted online on Nov 16, 2023. [doi: 10.48550/arXiv.2311.09724]
16. Lightman H, Kosaraju V, Burda Y, et al. Let's verify step by step. arXiv. Preprint posted online on May 31, 2023. [doi: 10.48550/arXiv.2305.20050]
17. Cunningham H, Ewart A, Riggs L, Huben R, Sharkey L. Sparse autoencoders find highly interpretable features in language models. arXiv. Preprint posted online on Sep 15, 2023. [doi: 10.48550/arXiv.2309.08600]
18. Galichin A, Dontsov A, Druzhinina P, et al. I have covered all the bases here: interpreting reasoning features in large language models via sparse autoencoders. arXiv. Preprint posted online on Mar 24, 2025. [doi: 10.48550/arXiv.2503.18878]
19. Dejong G, Horn SD, Gassaway JA, Slavin MD, Dijkers MP. Toward a taxonomy of rehabilitation interventions: using an inductive approach to examine the "black box" of rehabilitation. Arch Phys Med Rehabil. Apr 2004;85(4):678-686. [doi: 10.1016/j.apmr.2003.06.033] [Medline: 15083447]
20. Zheng L, Chiang WL, Sheng Y, et al. LMSYS-chat-1M: a large-scale real-world LLM conversation dataset. arXiv. Preprint posted online on Sep 21, 2023. [doi: 10.48550/arXiv.2309.11998]
21. Guha E, Marten R, Keh S, et al. OpenThoughts: data recipes for reasoning models. arXiv. Preprint posted online on Jun 4, 2025. [doi: 10.48550/arXiv.2506.04178]
22. GitHub. SAEDashboard. URL: https://github.com/jbloomAus/SAEDashboard [Accessed 2025-07-27]
23. Stolfo A, Balachandran V, Yousefi S, Horvitz E, Nushi B. Improving instruction-following in language models through activation steering. arXiv. Preprint posted online on Oct 15, 2024. [doi: 10.48550/arXiv.2410.12877]
24. Turner AM, Thiergart L, Leech G, et al. Steering language models with activation engineering. arXiv. Preprint posted online on Aug 20, 2023. [doi: 10.48550/ARXIV.2308.10248]
25. SAELens. GitHub. URL: https://github.com/jbloomAus/SAELens [Accessed 2026-06-03]
26. Vernooij RWM, Sanabria AJ, Solà I, Alonso-Coello P, Martínez García L. Guidance for updating clinical practice guidelines: a systematic review of methodological handbooks. Implement Sci. Jan 2, 2014;9(1):3. [doi: 10.1186/1748-5908-9-3] [Medline: 24383701]
27. Haskins R, Adams B. Distilled circuits: a mechanistic study of internal restructuring in knowledge distillation. arXiv. Preprint posted online on May 16, 2025. [doi: 10.48550/arXiv.2505.10822]

28. Baek DD, Tegmark M. Towards understanding distilled reasoning models: a representational approach. arXiv. Preprint posted online on Mar 5, 2025. [doi: 10.48550/arXiv.2503.03730]
29. Liu S, McCoy AB, Peterson JF, et al. Leveraging explainable artificial intelligence to optimize clinical decision support. J Am Med Inform Assoc. Apr 3, 2024;31(4):968-974. [doi: 10.1093/jamia/ocae019] [Medline: 38383050]

Abbreviations

AI: artificial intelligence
API: application programming interface
IgA: immunoglobulin A
LLM: large language model
QA: question-answering
SAE: sparse autoencoder
USMLE: United States Medical Licensing Examination

Edited by Andrew Coristine; peer-reviewed by Chaochen Wu, Chunwei Ma, Guru Lakshmi Priyanka Bodagala, Ruslan Kurmashev; submitted 20.Dec.2025; final revised version received 15.May.2026; accepted 18.May.2026; published 12.Jun.2026

Please cite as:
Liu J, Liu S, Wright A
Benchmark Integrity and Reasoning-Trace Errors in Medical Question Answering With Large Language Models: Mixed Methods Study With Sparse Autoencoders
J Med Internet Res 2026;28:e90061
URL: https://www.jmir.org/2026/1/e90061
doi: 10.2196/90061