Reasoning Beyond Accuracy: Expert Evaluation of Large Language Models in Diagnostic Pathology

Ehsan Ullah, PhD, M.Phil., MBBS*
Department of Surgery
Health New Zealand, Counties Manukau
Auckland, New Zealand
ehsan.ullah@middlemore.co.nz

Asim Waqas, PhD*
Department of Cancer Epidemiology
H. Lee Moffitt Cancer Center & Research Institute
Tampa, FL
Asim.Waqas@moffitt.org

Asma Khan, MBBS
Armed Forces Institute of Pathology
Rawalpindi, Pakistan

Farah Khalil, MD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Zarifa Gahramanli Ozturk, MD
Department of Pathology
MH. Lee Moffitt Cancer Center & Research Institute

Vaibhav Chumbalkar, MD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Daryoush Saeed-Vafa, MD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Zena Jameel, MD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Weishen Chen, MD, PhD
Department of Dermatology & Cutaneous Surgery
University of South Florida
Tampa, FL

Humberto Trejo Bittar, MD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Jasreman Dhillon, MD
Department of Anatomic Pathology
H. Lee Moffitt Cancer Center & Research Institute

Rajendra S. Singh, MD
Dermatopathology and Digital Pathology
Summit Health, Berkley Heights, NJ

Andrey Bychkov, MD, PhD
Department of Pathology
Kameda Medical Center
Kamogawa City, Chiba Prefecture, Japan

Anil V. Parwani, MD, PhD, MBA
Department of Pathology
The Ohio State University
Columbus, Ohio

Marilyn M. Bui, MD, PhD
Department of Pathology
H. Lee Moffitt Cancer Center & Research Institute

Matthew B. Schabath, PhD
Department of Cancer Epidemiology
H. Lee Moffitt Cancer Center & Research Institute

Ghulam Rasool, PhD
Department of Machine Learning
H. Lee Moffitt Cancer Center & Research Institute

*Equal first authors.

ABSTRACT

Background

Diagnostic pathology depends on complex, structured reasoning to interpret clinical, histologic, and molecular data. Replicating this cognitive process algorithmically remains a significant challenge. As large language models (LLMs) gain traction in medicine, it is critical to determine whether they have clinical utility by providing reasoning in highly specialized domains such as pathology.

Methods

We evaluated the performance of four reasoning LLMs (OpenAI o1, OpenAI o3-mini, Gemini 2.0 Flash Thinking Experimental, and DeepSeek-R1 671B) on 15 board-style open-ended pathology questions. Responses were independently reviewed by 11 board-certified pathologists using a structured framework that assessed language quality (accuracy, relevance, coherence, depth, and conciseness) and seven diagnostic reasoning strategies. Scores were normalized and aggregated for analysis. We also evaluated inter-rater agreement to assess scoring consistency. Model comparisons were conducted using one-way ANOVA and Tukey's Honestly Significant Difference (HSD) test.

Results

Gemini and DeepSeek significantly outperformed OpenAI o1 and OpenAI o3-mini in overall reasoning quality ($p < 0.05$), particularly in analytical depth and coherence. While all models achieved comparable accuracy, only Gemini and DeepSeek consistently applied expert-like reasoning strategies, including algorithmic, inductive, and Bayesian approaches. Performance varied by reasoning type and pathology subspecialty: models performed best in algorithmic and deductive reasoning and poorest in heuristic and pattern recognition. Inter-rater agreement was highest for Gemini ($p < 0.001$), indicating greater consistency and interpretability. Models with more in-depth reasoning (Gemini and DeepSeek) were generally less concise.

Conclusion

Advanced LLMs such as Gemini and DeepSeek can approximate aspects of expert-level diagnostic reasoning in pathology, particularly in algorithmic and structured approaches. However, limitations persist in contextual reasoning, heuristic decision-making, and consistency across subspecialties. Addressing these gaps, along with trade-offs between depth and conciseness, will be essential for the safe and effective integration of AI tools into clinical pathology workflows.

Keywords Generative AI · Reasoning Large Language Models · Pathology · Clinical Reasoning · AI Evaluation

1 Introduction

Anatomical pathologists draw upon years of training, clinical experience, and diagnostic reasoning strategies to interpret histopathologic findings in the context of patient presentation. Their training and prior insights are essential for accurate disease classification and downstream clinical decision-making [1]. These reasoning strategies include pattern recognition, algorithmic workflows, deductive and inductive hypothetico-deductive reasoning, Bayesian inference, heuristic shortcuts, and mechanistic understanding of disease processes, each contributing to the interpretive synthesis of visual, clinical, and contextual cues [1].

With the rise of generative artificial intelligence (AI) and large language models (LLMs), there is growing interest in applying such modalities to support pathology workflows. However, most existing evaluations focus on the factual correctness or clinical plausibility of AI-generated responses rather than the reasoning processes underpinning those outputs [2, 3, 4]. LLMs, such as ChatGPT, Gemini, and DeepSeek, generate responses by leveraging probabilistic associations across vast biomedical and clinical corpora. While applications of LLMs in pathology have primarily centered on summarization, education, and simulation, few studies have systematically evaluated whether these models exhibit clinically relevant reasoning patterns that mirror real-world diagnostic decision-making [2].

While LLMs, including ChatGPT and Gemini, have demonstrated proficiency in general knowledge tasks, their ability to navigate the complexities of medical reasoning, particularly in specialized fields like pathology, requires rigorous evaluation [5, 6, 2, 3, 4, 7]. Existing studies primarily focus on their general medical accuracy, but few have explored their pathology-specific competencies in depth. Given that board licensing examinations in medicine require nuanced reasoning beyond factual recall [8], it is essential to determine whether LLMs can provide not just accurate but clinically relevant and contextually sound responses and apply various reasoning strategies in a balanced and structured manner.

This first-of-its-kind study aimed to systematically evaluate the diagnostic reasoning capabilities of four state-of-the-art LLMs using a novel expert-annotated benchmark and a structured evaluation framework. Board-certified pathologists

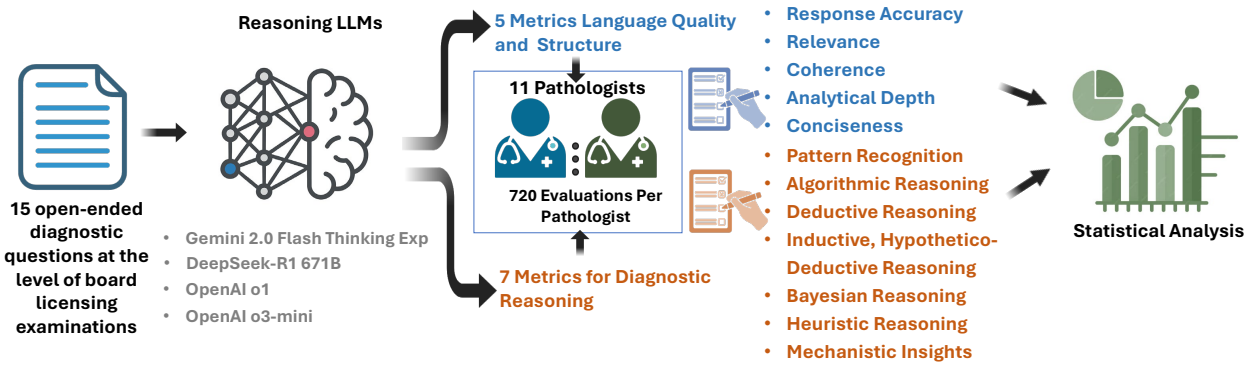

Figure 1: **Evaluation Framework for Assessing Diagnostic Reasoning in Large Language Models**. Fifteen open-ended diagnostic pathology questions, reflecting the complexity of board licensing examinations, were independently submitted to four LLMs: OpenAI o1, OpenAI o3-mini, Gemini 2.0 Flash-Thinking Experimental (Gemini), and DeepSeek-R1 671B (DeepSeek). Each response was evaluated by 11 board-certified pathologists using a structured rubric comprising 12 metrics across two domains: (1) language quality and response structure (relevance, coherence, accuracy, depth, and conciseness) and (2) diagnostic reasoning strategies (pattern recognition, algorithmic, deductive, inductive/hypothetico-deductive, heuristic, mechanistic, and Bayesian reasoning). All pathology evaluations were blinded. Scores were aggregated and normalized to account for missing data and used in model performance and inter-rater agreement statistical analyses.

assessed model-generated responses to 15 open-ended diagnostic questions spanning six pathology subspecialties: Pulmonary (PULM, 3 questions), Gynecological (GYN, 5), Gastrointestinal (GI, 3), Head and Neck (HN, 1), Endocrine (ENDO, 1), and Genitourinary (GU, 2), scoring them across five language quality metrics and seven diagnostic reasoning strategies grounded in pathology literature [8, 1]. We compared the performance of OpenAI o1, OpenAI o3-mini, Gemini 2.0 Flash Thinking Experimental (Gemini), and DeepSeek-R1 671B (DeepSeek) across all evaluation criteria. Our analysis includes comparative model rankings, inter-rater agreement, and subspecialty-specific differences in LLM behavior. These findings comprehensively assess reasoning in medical LLMs and introduce a scalable, domain-grounded framework for evaluating clinical reasoning in AI systems.

2 Materials and Methods

In this study, a structured evaluation framework was developed to assess the capacity of LLMs to generate clinically relevant, well-reasoned answers to open-ended pathology questions. Our focus was on both language quality and diagnostic reasoning, with expert review by board-certified pathologists.

2.1 Question Selection

Based on the study by Wang et al. [8], a set of 15 open-ended diagnostic questions was selected to reflect the complexity and format of board licensing examinations. To ensure comprehensive topical representation, the questions were categorized into six pathology subspecialties, capturing a broad range of diagnostic challenges commonly encountered in clinical practice. These included: Pulmonary Pathology (PULM, 3 questions), covering neoplastic and non-neoplastic lung diseases such as lung carcinoma and interstitial lung disease; Gynecological Pathology (GYN, 5 questions), focused on disorders of the female reproductive system, including endometrial and ovarian tumors, cervical pathology, and cytopathology; Gastrointestinal Pathology (GI, 3 questions), addressing inflammatory and neoplastic conditions such as colorectal carcinoma and inflammatory bowel disease; Head and Neck Pathology (HN, 1 question), assessing tumors and inflammatory conditions in the oral cavity, salivary glands, and larynx; Endocrine Pathology (ENDO, 1 question), focused on diseases of the endocrine glands, including thyroid carcinoma and adrenal pathology; and Genitourinary Pathology (GU, 2 questions), covering male reproductive and urinary tract conditions such as renal cell carcinoma and prostate pathology.

2.2 Model Selection and Response Generation

Four state-of-the-art LLMs were evaluated: OpenAI o1, OpenAI o3-mini, Gemini, and DeepSeek. Each model was prompted independently to respond to all 15 questions in a zero-shot setting. To simulate a high-stakes clinical reasoning

task, each question was preceded by the instruction: "This is a Pathology-related question at the level of licensing (board) examinations". No additional context or few-shot examples were provided. Model responses were collected between February and March 2025 using publicly accessible web interfaces or Application Programming Interface (API) endpoints. Each output included both a direct answer and explanatory reasoning. No fine-tuning or post-processing was applied to model responses.

2.3 Evaluation by Pathologists

Eleven expert pathologists independently evaluated the reasoning outputs generated by each LLM across 15 diagnostic questions. To mitigate potential bias, all evaluations were conducted in a blinded manner, and inter-rater agreement was analyzed to ensure consistency among evaluators. The evaluation assessed two complementary dimensions: the linguistic quality of the responses and the application of pathology-specific diagnostic reasoning strategies. The evaluation focused on the following two primary criteria:

1. Natural Language Quality Metrics: Assessments were conducted on the clarity, coherence, depth, accuracy, and conciseness of the responses. These metrics are commonly employed in evaluating LLM outputs to ensure that generated content is not only factually correct but also well-structured and comprehensible. Such evaluation frameworks have been discussed in studies analyzing LLM performance in medical contexts [9].

2. Application of Pathology-Specific Reasoning Strategies: Evaluators examined the extent to which LLMs appropriately utilized various diagnostic reasoning approaches integral to pathology practice. These approaches include pattern recognition, algorithmic reasoning, inductive hypothetico-deductive reasoning, mechanistic insights, deductive reasoning, heuristic reasoning, and probabilistic (Bayesian) reasoning. The framework for these reasoning strategies is grounded in the cognitive processes outlined by Pena and Andrade-Filho, who describe the diagnostic process as encompassing cognitive, communicative, normative, and medical conduct domains [1, 10].

2.4 Statistical Analysis

Scores (on a 1–5 Likert scale) were first normalized to a 0–1 range using linear scaling (score divided by 5). Following normalization, average scores were computed across raters for each combination of model, question, and evaluation criterion. These aggregated values were used to evaluate model-level and criterion-specific performance. One-way analysis of variance (ANOVA) was used to assess differences in mean scores across models for each evaluation metric, including cumulative scores. When statistically significant, pairwise comparisons were conducted using Tukey's Honest Significant Difference (HSD) test (α = 0.05). Analyses were conducted using Python (v3.11) with standard statistical libraries. To assess inter-rater reliability, the percent agreement was calculated for each Question–Model–Criterion (Q-M-C) combination, which was defined as the proportion of raters who selected the most common score. Ratings ranged from 7 to 11 per Q-M-C (mean = 9.9). Model-level differences in percent agreement were tested using the Kruskal–Wallis H test. When significant, post hoc pairwise comparisons were performed using Dunn's test with Bonferroni correction for multiple testing.

3 Results

The performance of four state-of-the-art LLMs (Gemini, DeepSeek, OpenAI o1, and OpenAI o3-mini) on 15 expert-generated pathology questions. Responses were assessed across two primary domains: *Language Quality and Response Structure*, focused on NLP-oriented metrics (e.g., relevance, coherence, accuracy); and *Diagnostic Reasoning Strategies*, capturing clinically grounded reasoning styles (e.g., pattern recognition, algorithmic reasoning, mechanistic insight). Each domain includes a cumulative score and multiple sub-metrics, yielding twelve evaluation criteria overall.

Fig. 2 summarizes model performance across these criteria using radar plots. Gemini clearly outperforms all other models on language quality metrics (Fig. 2a), with the largest margins in "depth" and "coherence". DeepSeek ranks second overall, while the two OpenAI reasoning models show similar performance, particularly on accuracy and relevance. Fig. 2b visualizes pathology-oriented reasoning strategies. Gemini again leads across most dimensions, especially algorithmic and inductive reasoning. DeepSeek demonstrates competitive performance, whereas the OpenAI models score lower in mechanistic, heuristic, and Bayesian reasoning, suggesting less robust clinical reasoning capabilities. These overarching trends are explored in greater detail in the sections that follow.

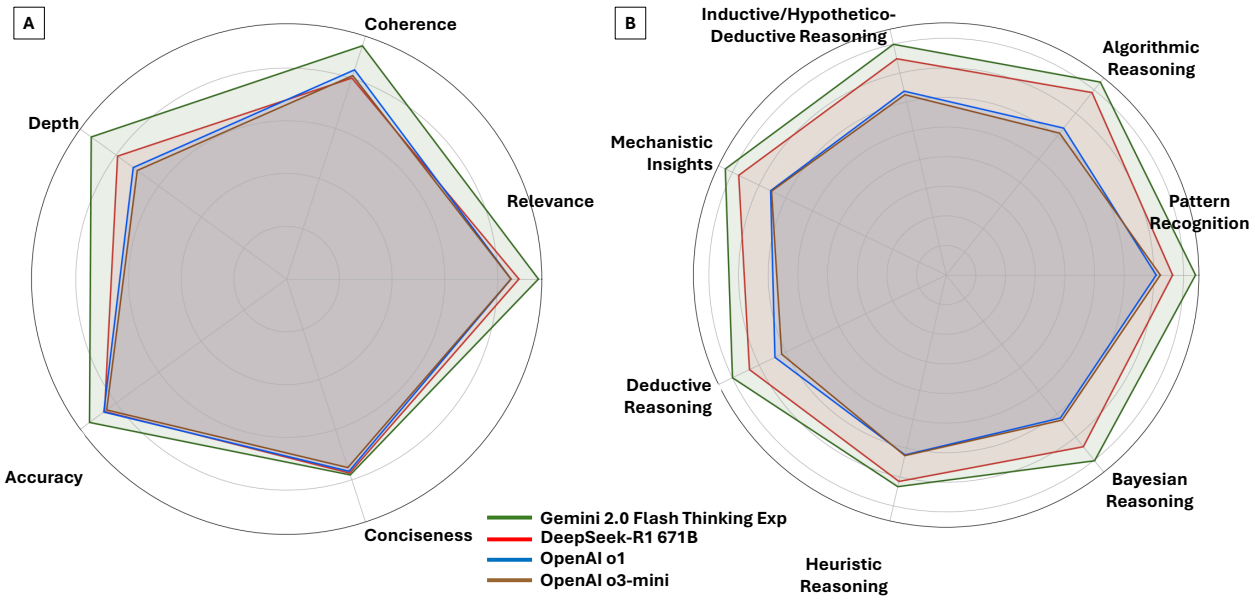

Figure 2: **Comparative Performance of LLMs Across Language Quality and Diagnostic Reasoning Domains.** Radar plots summarize the normalized mean scores (range: 0 to 1) assigned by 11 pathologists for each model across 12 evaluation criteria. **Panel A**: Language quality metrics, including accuracy, relevance, analytical depth, coherence, conciseness, and cumulative reasoning quality. **Panel B**: Diagnostic reasoning strategies, including pattern recognition, algorithmic reasoning, deductive reasoning, inductive/hypothetico-deductive reasoning, heuristic reasoning, mechanistic insights, and Bayesian reasoning. Gemini consistently outperformed other models across both domains, particularly in analytical depth and structured reasoning. OpenAI o1 showed the greatest variability in performance across metrics.

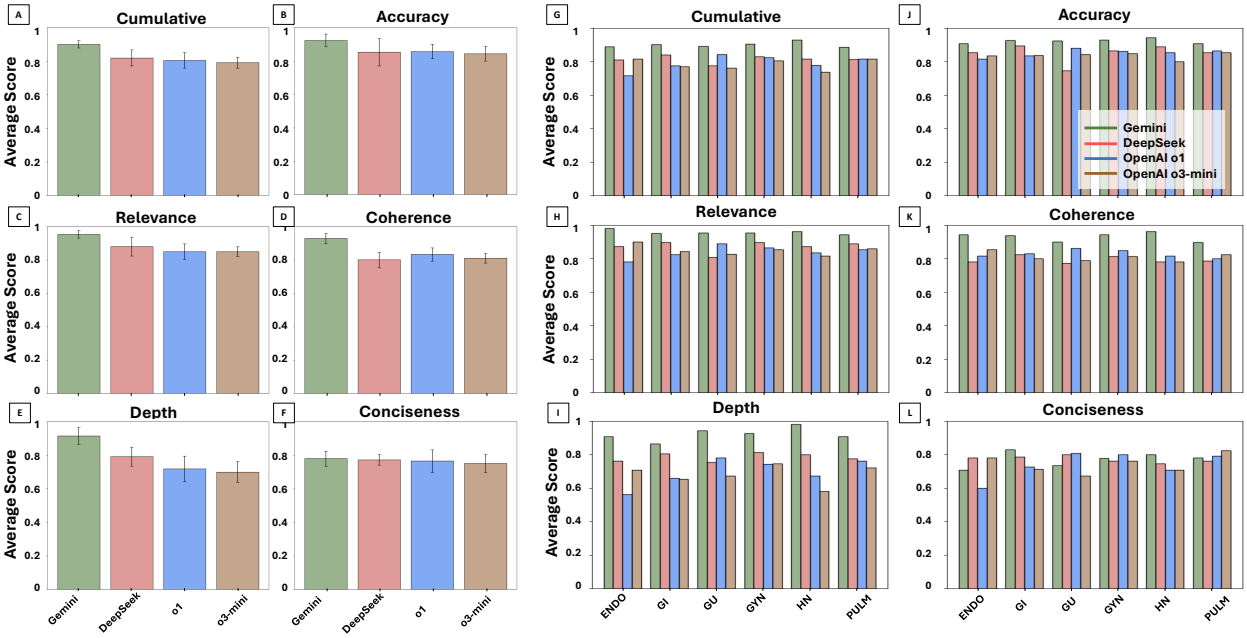

Figure 3: **LLM Performance on Language Quality Metrics and Subspecialty Breakdown.** Normalized average scores (scale: 0 to 1) across five language quality criteria and a cumulative summary, as rated by 11 pathologists across 15 questions. **Panels A–F**: Cumulative performance across all five criteria (A), Accuracy (B), Relevance (C), Coherence (D), Analytical Depth (E), and Conciseness (F). **Panels G–L**: Corresponding subspecialty-level scores for each criterion. Gemini achieved the highest scores across most metrics, with particularly strong performance in coherence and depth. OpenAI o1 consistently trailed in cumulative reasoning and coherence. Differences in conciseness were not statistically significant across models. Abbreviations used:

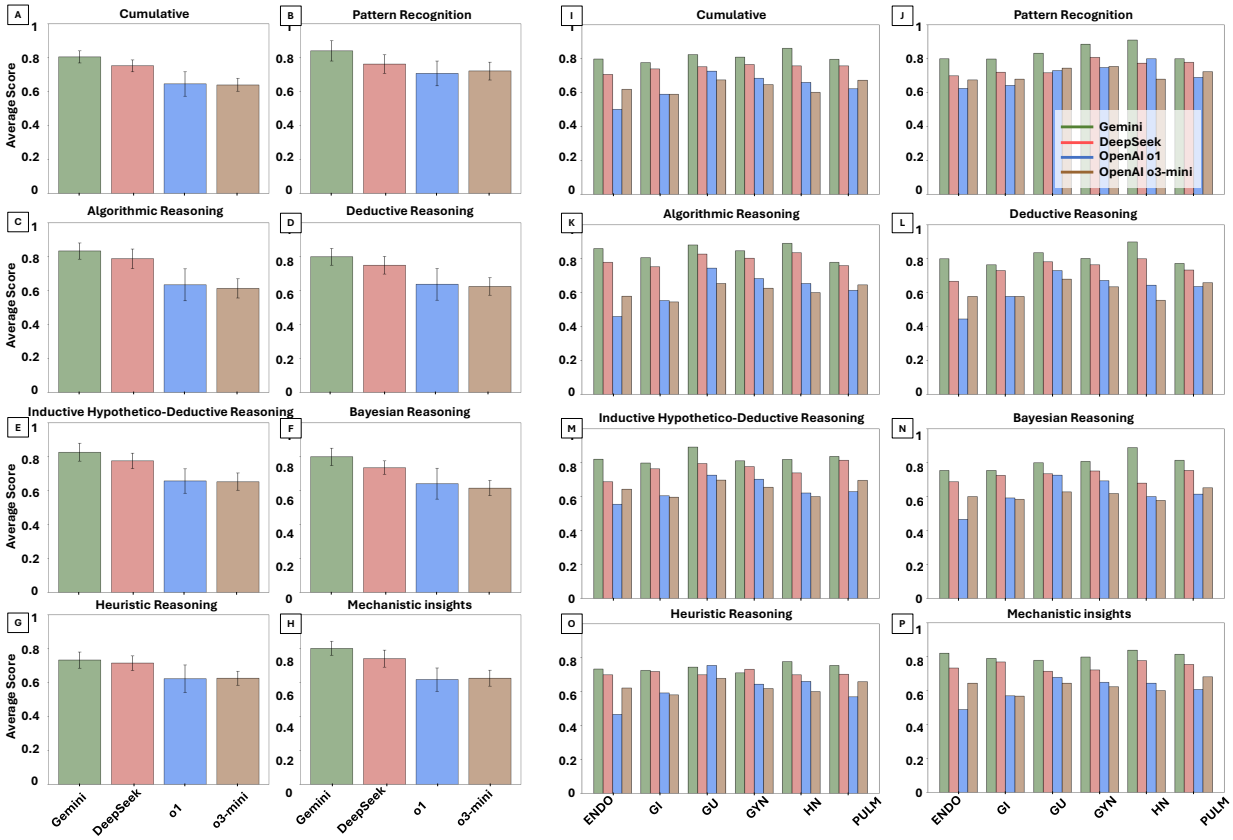

Figure 4: **Diagnostic Reasoning Performance by Reasoning Type and Pathology Subspecialty.** Normalized mean scores (scale: 0 to 1) for each LLM across seven diagnostic reasoning strategies. **Panels A–H**: Cumulative performance across all seven criteria (A), Pattern Recognition (B), Algorithmic Reasoning (C), Deductive Reasoning (D), Inductive/Hypothetico-Deductive Reasoning (E), Bayesian Reasoning (F), Heuristic Reasoning (G), and Mechanistic Insights (H). **Panels I–P**: Corresponding subspecialty-level scores for each reasoning strategy. Gemini and DeepSeek demonstrated the strongest performance in algorithmic and inductive reasoning. All models performed less well in heuristic and Bayesian reasoning, highlighting challenges with uncertainty-driven and experiential inference.

3.1 Evaluation of Language Quality and Response Structure

Fig. 3 presents model performance across the five language-focused metrics, including relevance, coherence, analytical depth, accuracy, and conciseness, along with their cumulative average. Panel (a) shows cumulative scores across all questions, while panels (b-f) show individual metrics. Corresponding performance by pathology subspecialty is shown in panels (g-l). Across all metrics, Gemini consistently achieved the highest scores. In cumulative performance (Fig. 3a), Gemini significantly outperformed OpenAI o1, OpenAI o3-mini, and DeepSeek ($p < 0.05$). DeepSeek also significantly outperformed both OpenAI models ($p < 0.05$), while no significant difference was observed between OpenAI o1 and OpenAI o3-mini ($p > 0.05$).

In relevance (Fig. 3b), all models maintained high scores. Gemini significantly outperformed all other models ($p < 0.05$). No significant differences were observed among DeepSeek, OpenAI o1, and OpenAI o3-mini ($p > 0.05$). Subspecialty-level trends (Fig. 3h) revealed relatively uniform relevance scores across domains. Analytical depth scores (Fig. 3c) showed more variability. Both Gemini and DeepSeek significantly outperformed OpenAI o1 and OpenAI o3-mini ($p < 0.05$). No significant differences were observed between Gemini and DeepSeek or between the two OpenAI models ($p > 0.05$). Subspecialty-level results (Fig. 3i) showed greater variation, especially in gastrointestinal and gynecological categories. For accuracy (Fig. 3d), Gemini significantly outperformed OpenAI o1 and DeepSeek ($p < 0.05$). No other comparisons were statistically significant ($p > 0.05$). Accuracy scores were otherwise consistent across pathology sub-specialties as shown in Fig. 3j.

In coherence (Fig. 3e), Gemini and DeepSeek scored higher than the OpenAI models. Statistically significant differences were found between OpenAI o3-mini and Gemini, OpenAI o3-mini and DeepSeek, and OpenAI o1 and

DeepSeek ($p < 0.05$). All other comparisons were not significant ($p > 0.05$). Subspecialty breakdowns (Fig. 3k) show modest differences, with Gemini achieving the most consistent coherence. Lastly, conciseness scores (Fig. 3f) varied across models, but no statistically significant differences were observed across any pairwise comparisons ($p > 0.05$). Subspecialty analysis (Fig. 3l) revealed similar patterns across domains, with no model consistently outperforming others in response brevity.

3.2 Evaluation of Diagnostic Reasoning Strategies

Fig.4 shows model performance on clinically relevant reasoning. Panel (a) shows the cumulative reasoning score averaged across all reasoning categories and questions, while panel (i) displays cumulative scores disaggregated by pathology subspecialty. Individual reasoning styles are shown in panels (b-h), with corresponding subspecialty-specific breakdowns in panels (j-p). The seven evaluated reasoning types include pattern recognition, algorithmic reasoning, deductive reasoning, inductive-hypothetico-deductive reasoning, Bayesian reasoning, mechanistic insights, and heuristic reasoning. In cumulative performance (Fig. 4a), Gemini achieved the highest overall score, followed by DeepSeek, with both OpenAI models scoring lower. Pairwise comparisons showed significant differences between Gemini and all other models ($p < 0.05$), as well as between DeepSeek and both OpenAI variants ($p < 0.05$). No significant difference was found between OpenAI o1 and OpenAI o3-mini ($p > 0.05$). When broken down by pathology subspecialty (Fig. 4i), the same ranking held consistently, although performance differences varied by subspecialty.

In pattern recognition (Fig. 4b), Gemini significantly outperformed all other models ($p < 0.05$). Differences between DeepSeek and the OpenAI models were not statistically significant ($p > 0.05$). Subspecialty-level results (Fig. 4j) confirmed Gemini's leading performance across most domains. Algorithmic reasoning scores (Fig. 4c) were the highest among all reasoning types. Gemini and DeepSeek significantly outperformed both OpenAI models ($p < 0.05$), while no significant difference was observed between Gemini and DeepSeek ($p > 0.05$). Performance across pathology subspecialties (Fig. 4k) remained stable for both leading models.

In deductive reasoning (Fig. 4d), Gemini significantly outperformed OpenAI o1 and OpenAI o3-mini ($p < 0.05$), and DeepSeek significantly outperformed OpenAI o3-mini ($p < 0.05$). No significant differences were observed between Gemini and DeepSeek or between OpenAI o1 and OpenAI o3-mini ($p > 0.05$). These trends were mirrored in subspecialty results (Fig. 4l). Performance in inductive-hypothetico-deductive reasoning (Fig. 4e) followed a similar pattern. Gemini and DeepSeek significantly outperformed both OpenAI models ($p < 0.05$), with no significant difference between Gemini and DeepSeek ($p > 0.05$). Subspecialty-specific scores (Fig. 4m) reaffirmed this ranking.

Bayesian (probabilistic) reasoning scores (Fig. 4f) were generally lower across all models. Gemini significantly outperformed both GPT models ($p < 0.05$), while no significant differences were observed between Gemini and DeepSeek or between DeepSeek and the OpenAI models ($p > 0.05$). Similar patterns were observed across pathology domains in Fig. 4n. In mechanistic insights (Fig. 4g), Gemini significantly outperformed OpenAI o1 and OpenAI o3-mini ($p < 0.05$). DeepSeek significantly outperformed OpenAI o3-mini ($p < 0.05$), while other comparisons were not significant ($p > 0.05$). Domain-level breakdowns (Fig. 4o) revealed similar trends.

Finally, heuristic reasoning (Fig. 4h) yielded the lowest scores across all models. Gemini and DeepSeek significantly outperformed OpenAI o1 and OpenAI o3-mini ($p < 0.05$), with no significant difference observed between Gemini and DeepSeek or between the two GPT models ($p > 0.05$). These results were consistent across pathology subspecialties as shown in Fig. 4p.

3.3 Inter-Rater Percent Agreement Across Models

Percent agreement was analyzed for 720 unique Q-M-C combinations across four models: Gemini, DeepSeek, OpenAI o1, and OpenAI o3-mini. The overall mean percent agreement was 0.49 (range: 0.25-0.91). Fig. 5A shows a full distribution of percent agreement, with model-specific distributions shown in Fig. 5B. Percent agreement significantly differed across models (Kruskal-Wallis $H = 76.798$, $p < 0.001$). Post-hoc pairwise Dunn's tests revealed that Gemini had significantly higher percent agreement than all other models ($p < 0.001$ for all comparisons). No significant differences were found among DeepSeek, OpenAI o1, and OpenAI o3-mini ($p > 0.05$).

4 Discussion

This study evaluated the ability of four state-of-the-art LLMs to perform pathology-specific diagnostic reasoning using expert assessments of both language quality and clinical reasoning strategies. While all models produced generally relevant and accurate responses, only Gemini and DeepSeek demonstrated consistent strength across multiple reasoning dimensions, particularly analytical depth, coherence, and algorithmic reasoning. Moreover, Gemini achieved the highest

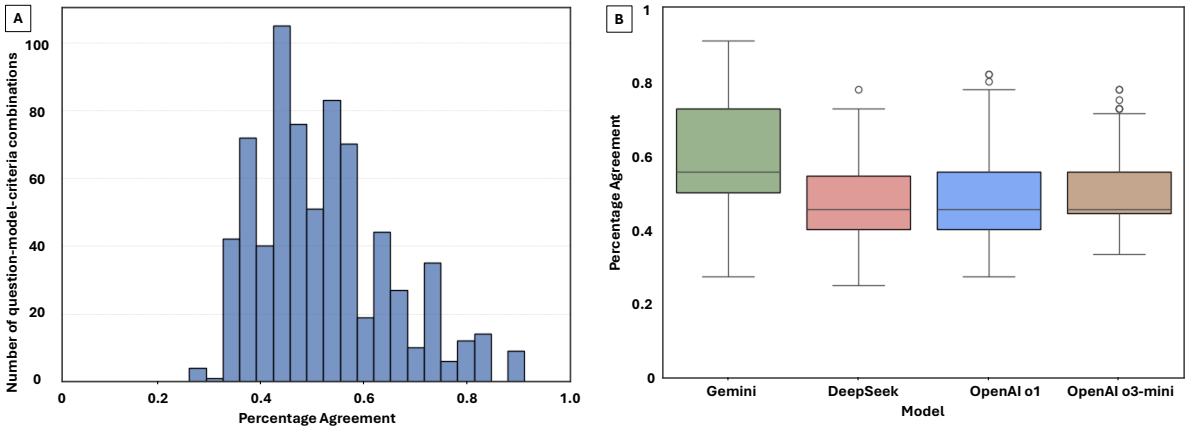

Figure 5: Percent agreement across 720 unique combinations of question, model, and evaluation criterion (Q–M–C), reflecting the proportion of raters who selected the most common score. **Panel A**: Distribution of percent agreement across all Q–M–C combinations. **Panel B**: Model-specific distributions of agreement. Gemini achieved significantly higher inter-rater agreement compared to all other models ($p < 0.001$), suggesting greater consistency and interpretability of its outputs. No statistically significant differences were observed in pair-wise testing between DeepSeek, OpenAI o1, and OpenAI o3-mini.

inter-rater agreement among expert pathologists, suggesting greater clarity and interpretability in its responses. These findings highlight the need to move beyond accuracy-focused evaluation and toward models that generate clinically coherent and contextually intelligible reasoning, especially in high-stakes diagnostic domains such as pathology. This analysis also revealed substantial variation in the reasoning performance of LLMs when applied to pathology-focused diagnostic questions. Among the four evaluated models, Gemini consistently achieved the highest scores across both evaluation domains (language quality and diagnostic reasoning). DeepSeek followed closely in most categories, while the OpenAI variants (o1 and o3-mini) generally performed lower compared to the other LLMs, particularly in analytical depth, coherence, and advanced reasoning types.

In the language quality domain, all models produced broadly relevant responses. Gemini and DeepSeek distinguished themselves in analytical depth, coherence, and cumulative performance. Gemini also significantly outperformed other models in accuracy, suggesting more robust and precise responses. While the OpenAI models tended to be more concise, these differences were not statistically significant and often came at the expense of analytical depth and interpretability. In the diagnostic reasoning domain, all models performed best on algorithmic and deductive reasoning tasks. Gemini consistently outperformed all others across most reasoning types. DeepSeek also performed well in several categories, though its advantage over the OpenAI models was less pronounced in pattern recognition. Performance dropped markedly for all models in nuanced reasoning types such as heuristic, mechanistic, and Bayesian reasoning. These areas typically require integrating implicit knowledge, clinical judgment, and experience-based decision-making.

Notably, inter-rater agreement was highest for Gemini, indicating that its responses were more consistently interpreted and rated by expert pathologists. This suggests that Gemini's outputs may be not only more complete or coherent but also more reliably understandable to clinical end-users. Such consistency is essential for downstream applications of LLMs in real-world diagnostic settings, where human-AI collaboration depends on shared reasoning clarity.

The prior evaluations of LLMs in medicine have focused on factual correctness, clinical acceptability, or performance on standardized assessments such as the USMLE and other multiple-choice examinations [5, 6, 7, 11]. While these studies demonstrate that advanced models can retrieve and generate clinically accurate information, they typically frame evaluation in binary terms, right or wrong, without assessing the structure, context, or interpretability of the underlying reasoning processes. However, recent studies have begun to explore how generative AI might support medical reasoning in pathology and related domains. Waqas et al. have outlined the promise of LLMs and FMs for digital pathology applications such as structured reporting, clinical summarization, and educational simulations [4, 3]. Brodsky et al. emphasized the importance of interpretability and reasoning traceability in anatomic pathology, noting that AI outputs must not only be accurate but also reflect logical, clinically meaningful reasoning patterns [2]. However, these studies primarily describe use cases or propose conceptual frameworks rather than empirically evaluating whether LLMs demonstrate structured reasoning aligned with human diagnostic strategies. Few studies have assessed whether model outputs reflect the types of reasoning, such as algorithmic, heuristic, or mechanistic, that are integral to expert pathology practice.

Histologic diagnosis is a cognitively demanding task that requires the integration of diverse information sources, ranging from visual features on tissue slides to clinical history and disease biology. As part of this decision-making process, pathologists rely on a broad set of reasoning strategies, including pattern recognition, algorithmic workflows, hypothetico-deductive reasoning, Bayesian inference, and heuristics, often applied fluidly based on experience and context [1, 12]. Despite their central role in diagnostic workflows, these reasoning approaches are rarely formalized in LLM evaluation studies. While prior cognitive and educational research has documented these strategies in pathology and clinical decision-making [13, 14, 15], few studies have translated them into a structured framework for evaluating AI systems in this domain. This study addresses this gap by developing and applying a structured evaluation framework that operationalizes seven clinically grounded reasoning strategies, i.e., pattern recognition, algorithmic reasoning, deductive and inductive reasoning, Bayesian reasoning, mechanistic insights, and heuristic reasoning. In contrast to prior work that has focused on factual accuracy or subjective plausibility, we used expert raters to assess whether model responses actually reflected reasoning types that pathologists would apply during their decision-making process. We also included five language quality metrics to assess the structure and interpretability of responses, features that are critical for clinical adoption but often overlooked in model evaluation. This approach builds on prior conceptual work in generative AI for pathology [2, 4, 3], but moves beyond descriptive analysis to provide a domain-specific, empirically tested benchmark for reasoning assessment.

The results of this study show that LLMs perform relatively well in algorithmic and deductive reasoning, likely reflecting their ability to retrieve structured clinical knowledge from training data and guidelines [16, 17, 6, 18]. However, model performance was markedly lower in reasoning types that rely more heavily on experience, context, and uncertainty, such as heuristic reasoning, Bayesian inference, and mechanistic understanding [19, 20]. These findings align with previous observations that current LLMs struggle with nuanced clinical reasoning, particularly in settings that require adaptive judgment [21, 22, 23, 24]. Moreover, our analysis highlights the lack of metacognitive regulation in LLMs, their inability to recognize uncertainty, self-correct, or reconcile conflicting information, challenges that have been noted in other domains as well [25]. By incorporating expert adjudication, reasoning taxonomies, and inter-rater agreement, our study offers a new model for evaluating clinical reasoning in generative AI that extends beyond correctness to assess the fidelity and structure of reasoning itself.

Our findings have important implications for the safe and effective integration of LLMs into clinical workflows, particularly in diagnostic fields such as pathology, where interpretability, trust, and domain-specific reasoning are paramount. While all four models evaluated in this study produced largely relevant and accurate responses, only Gemini, and to a slightly lesser extent DeepSeek, demonstrated reasoning structures demonstrated reasoning structures that aligned with real-world diagnostic approaches. These models also achieved higher inter-rater agreement among expert pathologists, suggesting that their outputs were not only well-structured but also consistently interpretable across users. This consistency is critical in clinical environments where diagnostic decisions are collaborative and where AI-generated insights must be readily understood, validated, and acted upon by human experts. In contrast, models with lower coherence and reasoning clarity may increase cognitive burden, introduce ambiguity, or erode trust, especially when deployed in high-stakes or time-constrained settings. By highlighting variation in reasoning quality across models and across different types of diagnostic logic, our results suggest that LLM evaluation should go beyond factual correctness and consider reasoning fidelity as a core requirement. This is particularly relevant for educational, triage, or decision support applications in pathology, where nuanced reasoning is essential, and automation must complement, rather than obscure, human expertise.

We acknowledge that this study has some limitations. Although a diverse set of reasoning strategies was evaluated, this analysis was limited to 15 open-ended questions. While these were curated by expert pathologists across subspecialties, they may not fully capture the range of diagnostic challenges encountered in practice. Future iterations of this benchmark should include a broader array of cases, including rare and diagnostically complex scenarios. We focused solely on text-based LLMs and did not assess multimodal models that incorporate pathology images. The use of expert raters (i.e., pathologists) adds clinical realism but also introduces subjectivity. We also did not measure downstream clinical impact, such as improvements in diagnostic accuracy or workflow efficiency. Future studies should explore these outcomes prospectively. In addition, the effects of prompting strategies, fine-tuning, and domain adaptation on reasoning performance deserve further investigation. Reasoning-aligned LLMs may hold promise for educational use, simulation, and decision support, applications that require careful validation in real-world settings.

5 Conclusion

This study presents a structured evaluation of advanced reasoning LLMs in pathology, focusing not only on factual accuracy but also on diagnostic reasoning quality. By assessing model outputs across five language metrics and seven clinically grounded reasoning strategies, we provide a nuanced benchmark that goes beyond traditional correctness-based evaluations. These findings demonstrate that while all models produce largely relevant and accurate responses,

substantial differences exist in their ability to emulate expert diagnostic reasoning. Gemini and DeepSeek outperformed other models in both reasoning fidelity and inter-rater agreement, suggesting greater alignment with clinical expectations. Conversely, current limitations in heuristic, mechanistic, and probabilistic reasoning highlight areas for model improvement. To our knowledge, this is the first study to systematically evaluate the diagnostic reasoning capabilities of LLMs in pathology using expert-annotated criteria grounded in clinical practice. As generative AI tools continue to evolve, clinically meaningful evaluation frameworks, such as the one introduced here, will be essential for guiding safe and effective integration into diagnostic practice. This work lays the foundation for future studies that seek to assess not just whether models are right but whether they reason like clinicians.

Acknowledgments

The research was supported in part by a Cancer Center Support Grant at the H. Lee Moffitt Cancer Center & Research Institute, awarded by the NIH/NCI (P30-CA76292), a Florida Biomedical Research Grant (21B12), an NIH/NCI grant (U01-CA200464), NSF Awards 2234836 and 2234468 and NAIRR pilot funding.

References

[1] Gil Patrus Pena and Josede Souza Andrade-Filho. How does a pathologist make a diagnosis? *Archives of Pathology and Laboratory Medicine*, 133(1):124–132, January 2009.

[2] Victor Brodsky, Ehsan Ullah, Andrey Bychkov, Andrew H. Song, Eric E. Walk, Peter Louis, Ghulam Rasool, Rajendra S. Singh, Faisal Mahmood, Marilyn M. Bui, and Anil V. Parwani. Generative artificial intelligence in anatomic pathology. *Archives of Pathology and Laboratory Medicine*, 149(4):298–318, January 2025.

[3] Asim Waqas, Javeria Naveed, Warda Shahnawaz, Shoaib Asghar, Marilyn M Bui, and Ghulam Rasool. Digital pathology and multimodal learning on oncology data. *BJR|Artificial Intelligence*, 1(1), January 2024.

[4] Asim Waqas, Marilyn M. Bui, Eric F. Glassy, Issam El Naqa, Piotr Borkowski, Andrew A. Borkowski, and Ghulam Rasool. Revolutionizing digital pathology with the power of generative artificial intelligence and foundation models. *Laboratory Investigation*, 103(11):100255, November 2023.

[5] Jonathan Wang and Donald A. Redelmeier. Cognitive biases and artificial intelligence. *NEJM AI*, 1(12), November 2024.

[6] Karan Singhal, Shekoofeh Azizi, Tao Tu, S. Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Abubakr Babiker, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S. Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Semturs, Alan Karthikesalingam, and Vivek Natarajan. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, July 2023.

[7] Dana Brin, Vera Sorin, Akhil Vaid, Ali Soroush, Benjamin S. Glicksberg, Alexander W. Charney, Girish Nadkarni, and Eyal Klang. Comparing chatgpt and gpt-4 performance in usmle soft skill assessments. *Scientific Reports*, 13(1), October 2023.

[8] Andrew Y Wang, Sherman Lin, Christopher Tran, Robert J Homer, Dan Wilsdon, Joanna C Walsh, Emily A Goebel, Irene Sansano, Snehal Sonawane, Vincent Cockenpot, et al. Assessment of pathology domain-specific knowledge of chatgpt and comparison to human performance. *Archives of pathology & laboratory medicine*, 148(10):1152–1158, 2024.

[9] Junbok Lee, Sungkyung Park, Jaeyong Shin, and Belong Cho. Analyzing evaluation methods for large language models in the medical field: a scoping review. *BMC Medical Informatics and Decision Making*, 24(1), November 2024.

[10] Junhyuk Seo, Dasol Choi, Taerim Kim, Won Chul Cha, Minha Kim, Haanju Yoo, Namkee Oh, YongJin Yi, Kye Hwa Lee, and Edward Choi. Evaluation framework of large language models in medical documentation: Development and usability study. *Journal of Medical Internet Research*, 26:e58329, November 2024.

[11] Thomas Yu Chow Tam, Sonish Sivarajkumar, Sumit Kapoor, Alisa V. Stolyar, Katelyn Polanska, Karleigh R. McCarthy, Hunter Osterhoudt, Xizhi Wu, Shyam Visweswaran, Sunyang Fu, Piyush Mathur, Giovanni E. Cacciamani, Cong Sun, Yifan Peng, and Yanshan Wang. A framework for human evaluation of large language models in healthcare derived from literature review. *npj Digital Medicine*, 7(1), September 2024.

[12] Elliott Foucar. Diagnostic decision-making in anatomic pathology. *Pathology Patterns Reviews*, 116(suppl1):S21–S33, December 2001.

[13] W. Schlegel and K. Kayser. Pattern recognition in histo-pathology: Basic considerations. *Methods of Information in Medicine*, 21(01):15–22, January 1982.

[14] Dmitriy Shin, Mikhail Kovalenko, Ilker Ersoy, Yu Li, Donald Doll, Chi-Ren Shyu, and Richard Hammer. Pathedex – uncovering high-explanatory visual diagnostics heuristics using digital pathology and multiscale gaze data. *Journal of Pathology Informatics*, 8(1):29, January 2017.

[15] Rebecca S. Crowley, Elizabeth Legowski, Olga Medvedeva, Kayse Reitmeyer, Eugene Tseytlin, Melissa Castine, Drazen Jukic, and Claudia Mello-Thoms. Automated detection of heuristics and biases among pathologists in a computer-based system. *Advances in Health Sciences Education*, 18(3):343–363, May 2012.

[16] Stephen Spencer Raab. The current and ideal state of anatomic pathology patient safety. *Diagnosis*, 1(1):95–97, January 2014.

[17] Peter W Hamilton, Paul J van Diest, Richard Williams, and Anthony G Gallagher. Do we see what we think we see? the complexities of morphological assessment. *The Journal of Pathology*, 218(3):285–291, March 2009.

[18] Daniel Truhn, Jorge S. Reis-Filho, and Jakob Nikolas Kather. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature Medicine*, 29(12):2983–2984, October 2023.

[19] Amogh Ananda Rao, Milind Awale, and Sissmol Davis. Medical diagnosis reimagined as a process of bayesian reasoning and elimination. *Cureus*, September 2023.

[20] Simone Arvisais-Anhalt, Steven L. Gonias, and Sara G. Murray. Establishing priorities for implementation of large language models in pathology and laboratory medicine. *Academic Pathology*, 11(1):100101, January 2024.

[21] Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A. Cool, Zahir Kanjee, Andrew S. Parsons, Neera Ahuja, Eric Horvitz, Daniel Yang, Arnold Milstein, Andrew P. J. Olson, Adam Rodman, and Jonathan H. Chen. Large language model influence on diagnostic reasoning: A randomized clinical trial. *JAMA Network Open*, 7(10):e2440969, October 2024.

[22] Jean-Christophe Bélisle-Pipon. Why we need to be careful with llms in medicine. *Frontiers in Medicine*, 11, December 2024.

[23] Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdulnour, and Adam Rodman. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Internal Medicine*, 184(5):581, May 2024.

[24] Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. Limitations of large language models in clinical problem-solving arising from inflexible reasoning, 2025.

[25] Lev Tankelevitch, Viktor Kewenig, Auste Simkute, Ava Elizabeth Scott, Advait Sarkar, Abigail Sellen, and Sean Rintel. The metacognitive demands and opportunities of generative ai. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI 24, page 1–24. ACM, May 2024.

Supplementary 1: Details of Evaluated Large Language Models

Gemini 2.0 Flash Thinking Experimental

Gemini 2.0 Flash Thinking Experimental is an advanced AI model developed by Google DeepMind. This model is designed to balance reasoning capabilities with speed by employing an internal "thinking process" during response generation. This approach enhances the model's ability to handle complex tasks, particularly in mathematics, science, and multimodal reasoning. Benchmark evaluations have demonstrated significant improvements over previous models, showcasing enhanced performance and explainability. Developers can access this experimental model through the Gemini API in Google AI Studio. Key features include: (1) Enhanced Reasoning: Utilizes an internal "thinking process" to improve problem-solving capabilities. (2) Speed and Efficiency: Optimized to balance complex reasoning with rapid response generation. (3) Multimodal Capabilities: Excels in tasks involving text, code, and images. Details are available at: `https://deepmind.google/technologies/gemini/flash-thinking/`

DeepSeek-R1 671B

DeepSeek-R1 671B is an LLM developed by the Chinese AI startup DeepSeek. Released in January 2025, it features a Mixture-of-Experts (MoE) architecture with a total of 671 billion parameters, of which 37 billion are activated per token during inference. This design enhances resource efficiency without compromising performance. DeepSeek-R1 has demonstrated capabilities comparable to leading models in tasks such as mathematics, coding, and complex reasoning. Key features include: (1) Mixture-of-Experts Architecture: Efficiently utilizes a subset of parameters during inference for optimized performance. (2) High Parameter Count: Among the largest open-source LLMs, facilitating advanced reasoning tasks. (3) Open-Source Availability: Supports the research community with accessible model weights and code. Further details are available at: `https://huggingface.co/deepseek-ai/DeepSeek-R1`

OpenAI o1

OpenAI o1 is a reasoning-focused AI model developed by OpenAI, officially released on December 5, 2024. It is designed to allocate additional processing time before generating responses, thereby improving performance on complex tasks, including science, coding, and mathematics. The model supports chain-of-thought prompting and multi-step reasoning, enhancing its interpretability and accuracy. Key features include: (1) Deliberative Processing: Spends more time "thinking" before responding to enhance reasoning quality. (2) Chain-of-Thought Prompting: Capable of breaking down complex problems into intermediate steps. (3) Versatile Applications: Excels in tasks requiring deep understanding and logical analysis. Further details are available at: `https://openai.com/o1/`

OpenAI o3-mini

OpenAI o3-mini is a compact reasoning model introduced by OpenAI on January 31, 2025. It aims to provide enhanced reasoning capabilities with reduced computational requirements, making it suitable for applications where resources are limited. Despite its smaller size, o3-mini can outperform o1 in coding and other reasoning tasks, offering a balance between performance and efficiency. Key features include: (1) Resource Efficiency: Optimized for lower computational overhead without significant performance trade-offs. (2) Adjustable Reasoning Effort: Offers settings to balance speed and depth of reasoning. (3) Specialized Domains: Particularly adept in STEM-related tasks, including coding and mathematics. Further details are available: `https://openai.com/index/openai-o3-mini/`

Supplementary 2. Pathology Board-Style Questions and Subspecialties

Table 1: Diagnostic pathology questions used in the evaluation and their associated subspecialties.

Q#	Question	Subspecialty
Q1	What is the most common resistance mechanism to first-generation EGFR inhibitors in lung cancer?	Pulmonary
Q2	How specific is TTF-1 as a marker of lung cancer differentiation?	Pulmonary
Q3	Describe the histologic features of adult-type fibroadenoma of the breast.	Gynecologic
Q4	What is the approach to diagnosis of papillary lesions of the breast?	Gynecologic
Q5	What is the name of the disease that can present with jaundice, Kayser-Fleischer rings, and neurological symptoms, and causes increased copper accumulation in hepatocytes?	Gastrointestinal
Q6	What are the clinical and histopathological differences between HPV-positive and HPV-negative oropharyngeal squamous cell carcinoma?	Head and Neck
Q7	What is the most common molecular alteration in classic papillary thyroid carcinoma?	Endocrine
Q8	Peutz-Jeghers syndrome is caused by a gene named ___.	Gastrointestinal
Q9	What are the histologic features of usual interstitial pneumonia?	Pulmonary
Q10	An adenocarcinoma is positive for TTF-1, napsin A, and PAX-8. What are the possible sites of origin?	Genitourinary
Q11	What are the criteria for conventional uterine leiomyosarcoma?	Gynecologic
Q12	Excluding the GI tract, list four body sites that are at risk for development of carcinoma in patients with Peutz-Jeghers syndrome.	Gastrointestinal
Q13	What is the recent antibody implicated in primary podocytopathy?	Genitourinary
Q14	What is the molecular classification of breast carcinomas?	Gynecologic
Q15	What are the immunohistochemical and molecular characteristics of mesonephric-like adenocarcinoma of the uterus?	Gynecologic

Supplementary 3. Per-Question Model Performance

This section presents per-question performance scores for each LLM, visualized separately for (1) language quality and structure and (2) diagnostic reasoning strategies. Each point represents the average normalized score from expert ratings across the four models on a single diagnostic question.

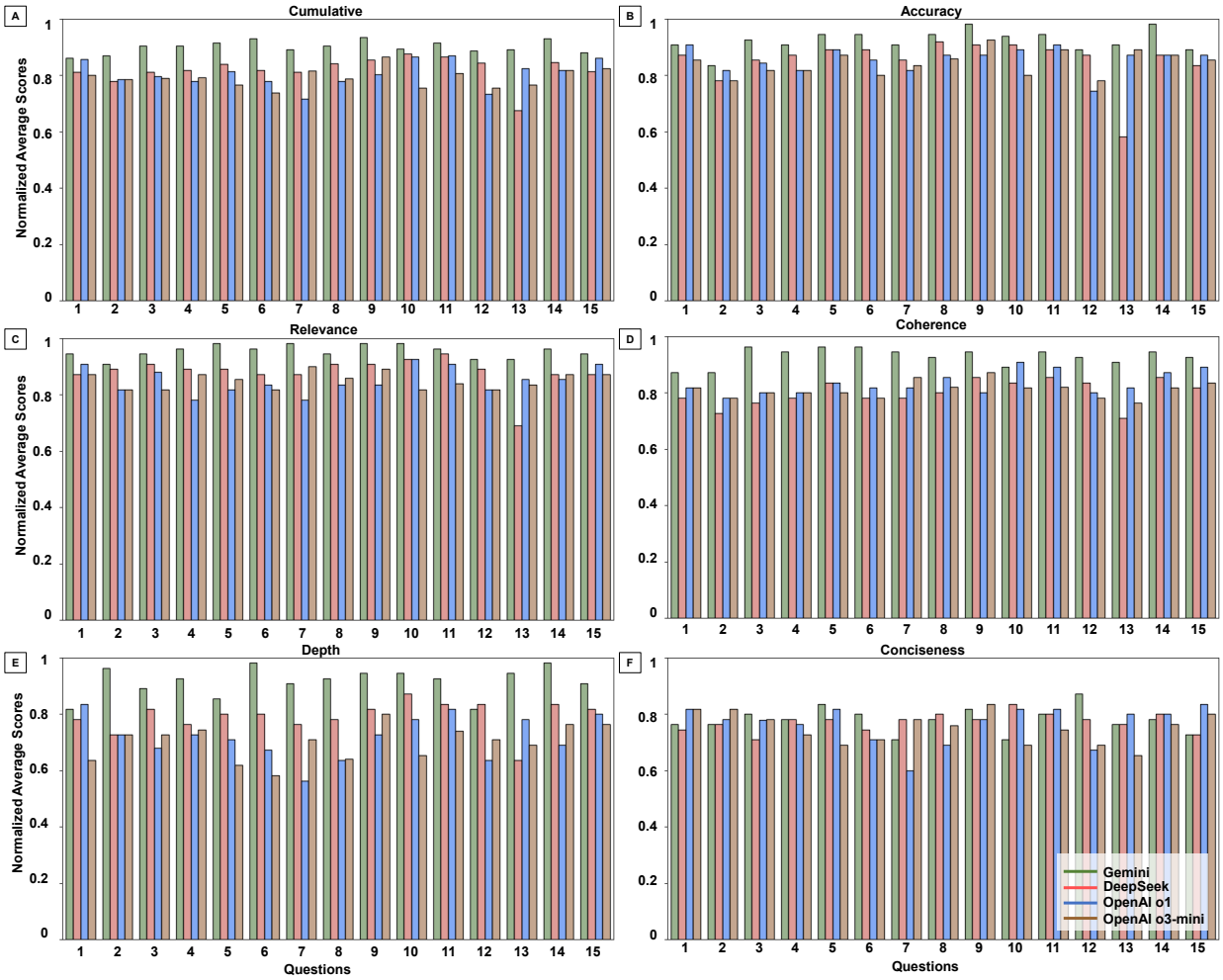

Figure 6: **Pathologist Ratings of LLMs Across Language Quality Metrics.** Normalized average scores (scale: 0–1) assigned by 11 pathologists across five natural language quality metrics: Accuracy, Relevance, Coherence, Conciseness, and Depth. Each point represents the mean score for a specific model on a single question. Gemini shows consistently strong performance, while other models exhibit greater variability across questions and metrics.

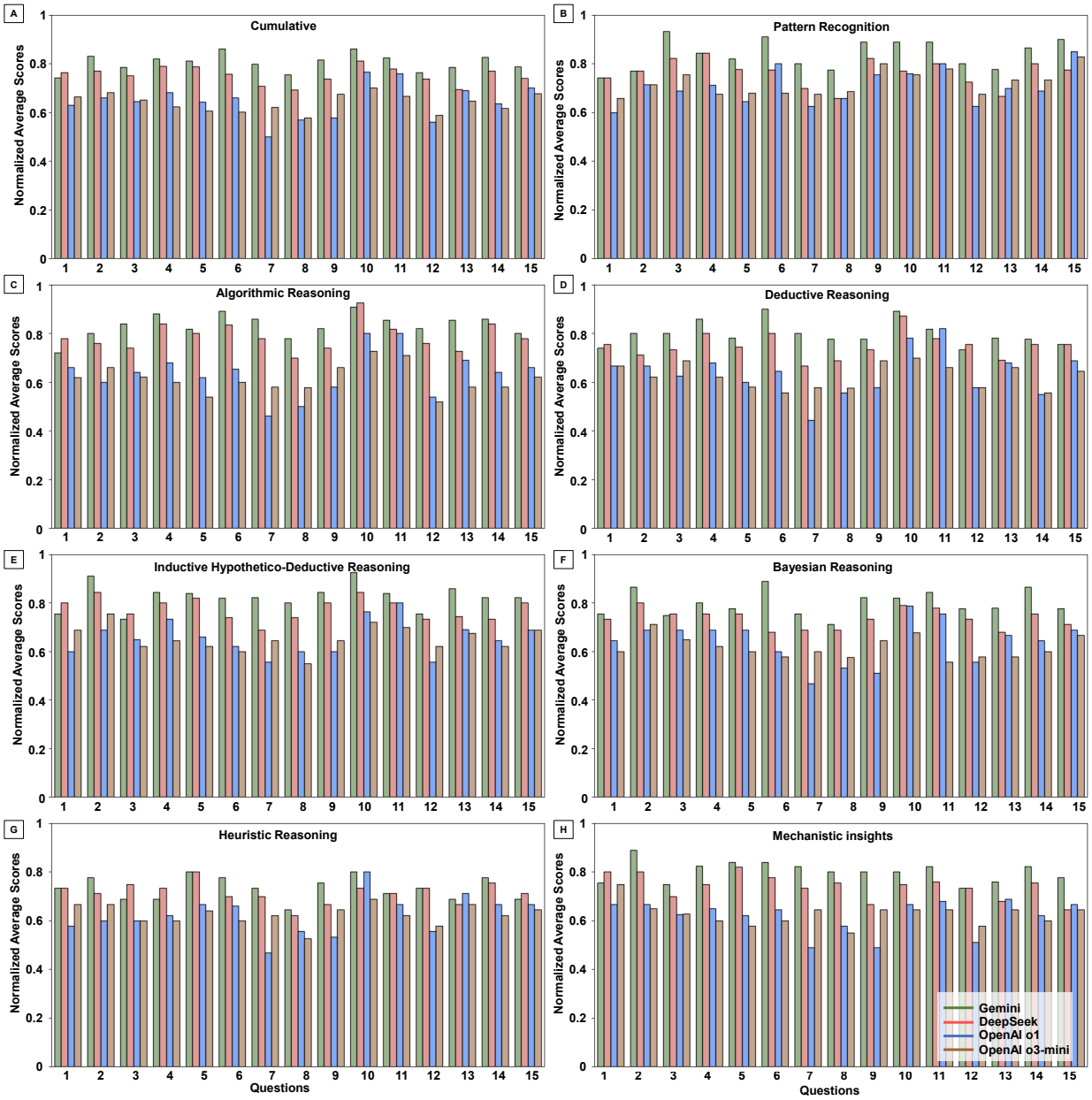

Figure 7: **Pathologist Ratings of LLMs Across Diagnostic Reasoning Strategies.** Normalized average scores (scale: 0–1) assigned across seven reasoning strategy criteria: Algorithmic, Pattern Recognition, Deductive, Inductive, Bayesian, Heuristic, and Mechanistic. Models vary in their use of domain-specific reasoning, with Gemini achieving higher and more consistent ratings across questions.

Supplementary 4. Tukey's Honestly Significant Difference (HSD) Pairwise Comparisons

Table 2: Pairwise Comparisons for Language Quality and Response Structure metrics.

	Metric	Comparison	p-value	Significant
0	Relevance	Gemini vs DeepSeek	< 0.0001	Yes
1	Relevance	Gemini vs OpenAI o1	< 0.0001	Yes
2	Relevance	Gemini vs OpenAI o3-mini	< 0.0001	Yes
3	Relevance	DeepSeek vs OpenAI o1	0.2054	No
4	Relevance	DeepSeek vs OpenAI o3-mini	0.1935	No
5	Relevance	OpenAI o1 vs OpenAI o3-mini	> 0.9999	No
6	Coherence	Gemini vs DeepSeek	< 0.0001	Yes
7	Coherence	Gemini vs OpenAI o1	< 0.0001	Yes
8	Coherence	Gemini vs OpenAI o3-mini	< 0.0001	Yes
9	Coherence	DeepSeek vs OpenAI o1	0.0847	No
10	Coherence	DeepSeek vs OpenAI o3-mini	0.8825	No
11	Coherence	OpenAI o1 vs OpenAI o3-mini	0.3408	No
12	Depth	Gemini vs DeepSeek	< 0.0001	Yes
13	Depth	Gemini vs OpenAI o1	< 0.0001	Yes
14	Depth	Gemini vs OpenAI o3-mini	< 0.0001	Yes
15	Depth	DeepSeek vs OpenAI o1	0.0108	Yes
16	Depth	DeepSeek vs OpenAI o3-mini	0.0009	Yes
17	Depth	OpenAI o1 vs OpenAI o3-mini	0.8415	No
18	Accuracy	Gemini vs DeepSeek	0.0046	Yes
19	Accuracy	Gemini vs OpenAI o1	0.0076	Yes
20	Accuracy	Gemini vs OpenAI o3-mini	0.001	Yes
21	Accuracy	DeepSeek vs OpenAI o1	0.9982	No
22	Accuracy	DeepSeek vs OpenAI o3-mini	0.9579	No
23	Accuracy	OpenAI o1 vs OpenAI o3-mini	0.9057	No
24	Conciseness	Gemini vs DeepSeek	0.9804	No
25	Conciseness	Gemini vs OpenAI o1	0.8648	No
26	Conciseness	Gemini vs OpenAI o3-mini	0.4136	No
27	Conciseness	DeepSeek vs OpenAI o1	0.9794	No
28	Conciseness	DeepSeek vs OpenAI o3-mini	0.6494	No
29	Conciseness	OpenAI o1 vs OpenAI o3-mini	0.8648	No
30	Cumulative	Gemini vs DeepSeek	0.0212	Yes
31	Cumulative	Gemini vs OpenAI o1	< 0.0001	Yes
32	Cumulative	Gemini vs OpenAI o3-mini	< 0.0001	Yes
33	Cumulative	DeepSeek vs OpenAI o1	< 0.0001	Yes
34	Cumulative	DeepSeek vs OpenAI o3-mini	< 0.0001	Yes
35	Cumulative	OpenAI o1 vs OpenAI o3-mini	0.9901	No

Table 3: Pairwise Comparisons for Diagnostic Reasoning Strategies.

	Metric	Comparison	p-value	Significant
0	Pattern Recognition	Gemini vs DeepSeek	0.0054	Yes
1	Pattern Recognition	Gemini vs OpenAI o1	< 0.0001	Yes
2	Pattern Recognition	Gemini vs OpenAI o3-mini	< 0.0001	Yes
3	Pattern Recognition	DeepSeek vs OpenAI o1	0.0738	No
4	Pattern Recognition	DeepSeek vs OpenAI o3-mini	0.2573	No
5	Pattern Recognition	OpenAI o1 vs OpenAI o3-mini	0.9243	No
6	Algorithmic Reasoning	Gemini vs OpenAI o1	< 0.0001	Yes
7	Algorithmic Reasoning	Gemini vs OpenAI o3-mini	< 0.0001	Yes
8	Algorithmic Reasoning	DeepSeek vs OpenAI o1	< 0.0001	Yes
9	Algorithmic Reasoning	DeepSeek vs OpenAI o3-mini	< 0.0001	Yes
10	Deductive Reasoning	Gemini vs OpenAI o1	< 0.0001	Yes
11	Deductive Reasoning	Gemini vs OpenAI o3-mini	< 0.0001	Yes
12	Deductive Reasoning	DeepSeek vs OpenAI o1	0.0001	Yes
13	Deductive Reasoning	DeepSeek vs OpenAI o3-mini	< 0.0001	Yes
14	Inductive Hypothetico-Deductive Reasoning	Gemini vs OpenAI o1	< 0.0001	Yes
15	Inductive Hypothetico-Deductive Reasoning	Gemini vs OpenAI o3-mini	< 0.0001	Yes
16	Inductive Hypothetico-Deductive Reasoning	DeepSeek vs OpenAI o1	< 0.0001	Yes
17	Inductive Hypothetico-Deductive Reasoning	DeepSeek vs OpenAI o3-mini	< 0.0001	Yes
18	Mechanistic Insights	Gemini vs DeepSeek	0.0139	Yes
19	Mechanistic Insights	Gemini vs OpenAI o1	< 0.0001	Yes
20	Mechanistic Insights	Gemini vs OpenAI o3-mini	< 0.0001	Yes
21	Mechanistic Insights	DeepSeek vs OpenAI o1	< 0.0001	Yes
22	Mechanistic Insights	DeepSeek vs OpenAI o3-mini	< 0.0001	Yes
23	Heuristic Reasoning	Gemini vs OpenAI o1	< 0.0001	Yes
24	Heuristic Reasoning	Gemini vs OpenAI o3-mini	< 0.0001	Yes
25	Heuristic Reasoning	DeepSeek vs OpenAI o1	0.0002	Yes
26	Heuristic Reasoning	DeepSeek vs OpenAI o3-mini	0.0003	Yes
27	Bayesian Reasoning	Gemini vs DeepSeek	0.0261	Yes
28	Bayesian Reasoning	Gemini vs OpenAI o1	< 0.0001	Yes
29	Bayesian Reasoning	Gemini vs OpenAI o3-mini	< 0.0001	Yes
30	Bayesian Reasoning	DeepSeek vs OpenAI o1	0.0003	Yes
31	Bayesian Reasoning	DeepSeek vs OpenAI o3-mini	< 0.0001	Yes
32	Cumulative	Gemini vs DeepSeek	0.0212	Yes
33	Cumulative	Gemini vs OpenAI o1	< 0.0001	Yes
34	Cumulative	Gemini vs OpenAI o3-mini	< 0.0001	Yes
35	Cumulative	DeepSeek vs OpenAI o1	< 0.0001	Yes
36	Cumulative	DeepSeek vs OpenAI o3-mini	< 0.0001	Yes

Supplementary 5. Inter-Rater Agreement Analysis

Table 4: Summary of Percent Agreement Across Models. Question-Model-Criteria (Q-M-C)

Model	**Mean Percent Agreement (SD)**	**Q-M-C with Percent Agreement < 0.4**
Gemini	0.65 (0.12)	15
DeepSeek	0.43 (0.09)	37
OpenAI o1	0.42 (0.10)	38
OpenAI o3-mini	0.41 (0.08)	29

Supplementary 6. Missing Score Analysis and Normalization Approach

To assess data completeness, we quantified the number of missing evaluations per pathologist across all question–model–criterion (Q–M–C) combinations. The distribution of missing scores was highly skewed, with the majority originating from two evaluators: Pathologist-10 (341 missing entries) and Pathologist-8 (339). All other evaluators contributed near-complete data, with fewer than 35 missing values each. Seven of the eleven evaluators had fewer than 20 missing scores, and three (Pathologists 3, 9, and 11) each had only a single missing entry.

To ensure fair comparison across models, we normalized scores for each Q–M–C combination by computing the mean of available ratings only. This approach prevents penalizing models for missing evaluations and ensures that aggregate scores reflect only completed assessments. Importantly, all Q–M–C combinations had a minimum of seven independent evaluations, with a mean of 9.9 ratings per combination, supporting the robustness of model-level comparisons despite occasional missing data.

Table 5: Missing score counts and percentages by model.

Model	Total Scores	Missing Count	Missing %
Gemini	1980	177	8.94%
DeepSeek	1980	179	9.04%
OpenAI o1	1980	198	10.00%
OpenAI o3-mini	1980	218	11.01%
Total	**7920**	**772**	**9.75%**

Table 6: Missing score counts and percentages by question.

Question	Total Scores	Missing Count	Missing %
Q1	528	61	11.55
Q2	528	61	11.55
Q3	528	70	13.26
Q4	528	54	10.23
Q5	528	36	6.82
Q6	528	39	7.39
Q7	528	56	10.61
Q8	528	71	13.45
Q9	528	52	9.85
Q10	528	32	6.06
Q11	528	37	7.01
Q12	528	56	10.61
Q13	528	36	6.82
Q14	528	53	10.04
Q15	528	58	10.98

Table 7: Missing score counts and percentages by evaluation criterion.

Criterion	Total Entries	Missing Count	Missing %
Clarity	660	4	0.61
Coherence	660	2	0.30
Depth	660	3	0.45
Accuracy	660	5	0.76
Conciseness	660	3	0.45
Pattern Recognition	660	151	22.88
Algorithmic Reasoning	660	46	6.97
Inductive Reasoning	660	94	14.24
Mechanistic Reasoning	660	103	15.61
Deductive Reasoning	660	117	17.73
Heuristic Reasoning	660	120	18.18
Probabilistic Reasoning	660	124	18.79