LIVECLIN: A *Live* CLINICAL BENCHMARK WITHOUT LEAKAGE

Anonymous authors
Paper under double-blind review

ABSTRACT

The reliability of medical LLM evaluation is critically undermined by data contamination and knowledge obsolescence, leading to inflated scores on static benchmarks. To address these challenges, we introduce *LiveClin*, a live benchmark designed for the faithful replication of clinical practice. Built from contemporary, peer-reviewed case reports and updated biannually, LiveClin ensures clinical currency and resists data contamination. Using a verified AI-human workflow involving 239 physicians, we transform authentic patient cases into complex, multimodal evaluation scenarios that span the entire clinical pathway. The benchmark currently comprises 1,407 case reports and 6,605 questions. Our evaluation of 26 models on LiveClin reveals the profound difficulty of these real-world scenarios, with the top-performing model achieving a Case Accuracy of just 35.7%. We find that the era of "free lunch" improvements from simple model scaling is over, as newer models do not consistently outperform their predecessors. Furthermore, our analysis uncovers distinct reasoning weaknesses across model classes. LiveClin thus provides a continuously evolving, clinically-grounded framework to steer the development of medical LLMs towards greater reliability and real-world utility.

1 INTRODUCTION

Large language models (LLMs) hold immense promise for transforming healthcare, from aiding in complex diagnostics to personalizing patient care. However, the safe and effective integration of these powerful tools into clinical practice is entirely dependent on our ability to rigorously evaluate their true capabilities. As the gap between general knowledge and expert-level clinical reasoning widens, the development of sophisticated, clinically-grounded benchmarks becomes not just a matter of academic progress, but a prerequisite for building trustworthy medical AI.

However, the prevailing evaluation landscape fails to mirror clinical practice, suffering from two limitations. First, the static design of benchmarks like MedQA (Jin et al., 2020) not only makes them inherently vulnerable to data contamination and knowledge obsolescence (Oren et al., 2024; White et al., 2025) but also risks creating an illusion of capability through inflated scores. Second, their single-turn assessments are misaligned with the longitudinal nature of patient care. By evaluating reasoning in isolated, synthetic snapshots, even advanced systems like MedXpertQA (Zuo et al., 2025) and AgentClinic (Schmidgall et al., 2025) reduce patient management to a series of disconnected tasks. This approach fails to assess the integrated reasoning required to navigate a patient's entire clinical pathway, from initial presentation to long-term management.

To overcome these limitations, we introduce a clinically aligned benchmark named LiveClin, which is biannually updated and contains 1,407 clinical cases and 6,605 questions to date. Concretely, the benchmark is sourced from contemporary, peer-reviewed case reports from PubMed Central (PMC) Open Access subset to mitigate both data contamination and knowledge obsolescence. As illustrated in Figure 1, to simulate the entire clinical pathway, each case is transformed into a multi-stage exam to assess whether a model can sequentially integrate diverse modalities that reflect the patient's evolving condition. The ablation study also indicates that the overall AI-human construction workflow is superior to physician-only curation for generating challenging and high-quality content. To verify the rigor of the benchmark, we implemented a 239-physician screening pipeline, guided by the conservative principle of **rejecting any potentially flawed question**.

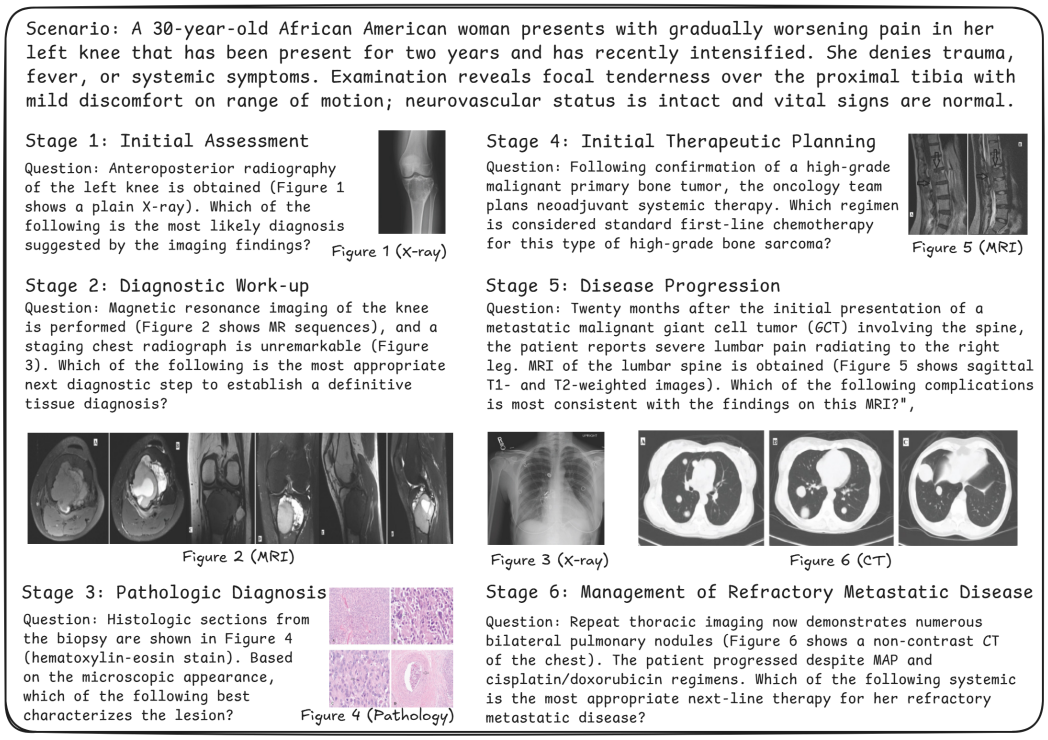

Figure 1: An example from LiveClin simulating the entire clinical pathway. The case progresses from initial assessment to long-term management, with new clinical information and diverse imaging modalities (e.g., X-ray, MRI, pathology, CT) progressively introduced at each key decision point to challenge the model's reasoning in an evolving scenario.

Our evaluation of 26 models on LiveClin yields several key insights into the current state of medical AI. We find that the benchmark is exceptionally challenging, with even top proprietary models like *o3* achieving a Case Accuracy of only 35.7%. Our analysis reveals that the era of "free lunch" improvements from simple model scaling is over, as newer models do not guarantee better performance on these expert-level tasks. Furthermore, our in-depth analysis uncovers that different classes of models exhibit distinct reasoning weaknesses. We observe that top proprietary models often falter when integrating complex diagnostic data, while open-source medical models struggle with long-context retention over the full patient journey.

Contributions The main contributions of this work are threefold: (1) **LiveClin**, a novel, dynamic, and multimodal benchmark that evaluates the full clinical pathway, designed to be contamination-resistant and continuously updated; (2) A scalable and verified **AI-human workflow** for generating and maintaining high-quality evaluations that mirror the clinical practice, proven to be more cost-effective and to produce more challenging questions than human-only authoring; and (3) A comprehensive **evaluation of 26 leading LLMs**, providing a new baseline for state-of-the-art clinical reasoning and uncovering critical, distinct failure modes that inform future model development.

2 Motivation

Data contamination poses a fundamental threat to medical LLM evaluation by eroding benchmark reliability. As models are trained on ever-expanding, web-scale corpora, the questions and answers from popular static benchmarks are inevitably absorbed into their training sets (Deng et al., 2024). This widespread contamination means that models are increasingly being tested on data they have already seen, leading to inflated performance scores (Balunović et al., 2025). This is not a minor flaw but a fundamental threat to evaluation integrity, as it erodes community's ability to distinguish genuine progress from mere benchmark hacking.

While common, decontamination efforts are merely reactive and often incomplete (Zhu et al., 2024). A truly robust solution must be proactive, designed with inherent resistance to contamination. This

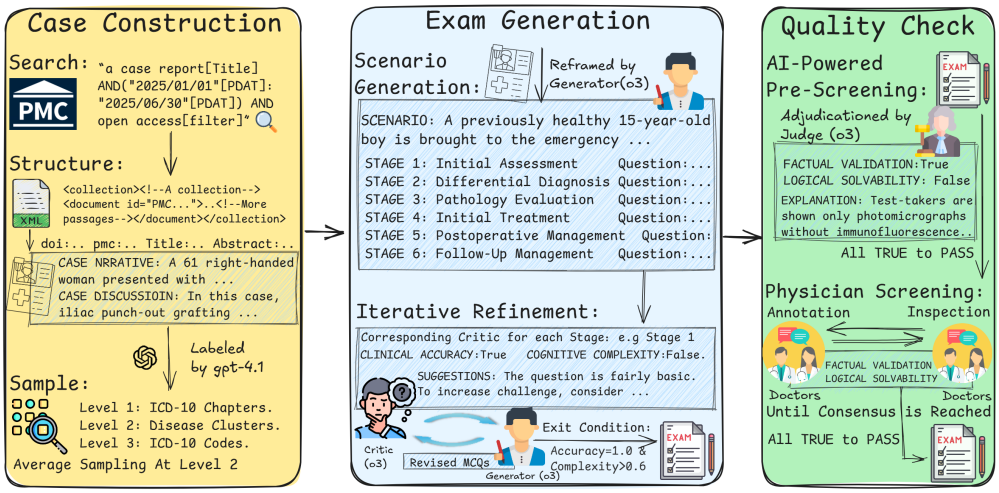

Figure 3: Overview of the LiveClin Construction Pipeline. Our three-stage pipeline creates a dynamic and clinically authentic benchmark. (1) Case Construction: We source and sample recent, peer-reviewed case reports to build a contemporary data foundation. (2) Exam Generation: An iterative Generator-Critic architecture transforms static reports into sequential reasoning problems. (3) Quality Check: A synergistic AI-human workflow, featuring AI pre-screening and multi-physician verification, ensures factual and logical integrity.

challenge is compounded by a second, related issue: knowledge obsolescence. Since clinical medicine evolves constantly (Cullen et al., 2019; Mitchell et al., 2023), any static test not only becomes a predictable target for contamination but also inevitably loses its clinical relevance over time.

Pilot Study: Quantifying the Impact of Data Recency. To empirically quantify the dual impacts of data contamination and knowledge obsolescence, we conducted a longitudinal pilot study using our main pipeline, with full methodological details available in Sections 3,4 and Appendix C. As shown in Figure 2, the results demonstrate a significant performance gap between a model's performance on older, potentially contaminated data versus novel, contemporary data. This is starkly illustrated by *GPT-5*, which scores as high as 45.0% on data within its knowledge base but drops by nearly 10 percentage points on cases published after its cutoff. This pattern, consistent across models, quantifies the distorting effects of data contamination, which inflates scores on seen data, and knowledge obsolescence, which causes failures on unseen, newer knowledge. These findings suggest that static benchmarks are an unreliable proxy for true clinical reasoning, highlighting the necessity of a live evaluation paradigm.

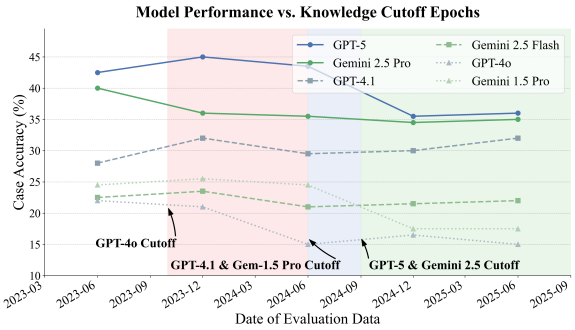

Figure 2: LLM Case Accuracy across datasets with different publication dates. Shaded regions indicate the knowledge cutoff for different models

3 LIVECLIN

To address the critical need for a dynamic, contamination-resistant benchmark established in Section 2, we developed LiveClin through a rigorous methodology, overviewed in Figure 3. This section first introduces the Taxonomy (§3.1) we designed to structure the benchmark for fine-grained analysis. We then detail the three core pipeline stages: (1) Case Construction (§3.2), a process of case curation and sampling to establish a contemporary data foundation; (2) Exam Generation (§3.3), which transforms static reports into multi-step problems simulating the entire clinical pathway; and (3) Quality Check (§3.4), a rigorous, multi-layered assurance process that guarantees the benchmark's medical validity. Finally, we present the Benchmark Statistics and Composition (§3.5).

3.1 CLINICAL TAXONOMY

LiveClin's Taxonomy is a foundational framework for multi-resolution performance analysis, designed to overcome the single-score, narrow-scope limitations of existing benchmarks. It features a three-level hierarchy that evaluates model capabilities across a spectrum of granularity. The full taxonomy is detailed in Appendix D, and the analytical purpose of each level is outlined below:

- **Level 1: ICD-10 Chapters.** The highest level of our taxonomy is adapted from the authoritative ICD-10 framework to ensure its clinical and scientific validity. It consists of 16 clinically coherent chapters designed to provide a macro-level view of model capabilities across major medical specialties. To enhance focus and relevance, we merge closely related domains (e.g., Pregnancy and Perinatal Conditions) and excluded non-specific chapters (e.g., "Symptoms & Laboratory Signs") whose cases are more appropriately classified under a primary disease.
- **Level 2: Disease Clusters.** This intermediate level is adapted from the authoritative China National Healthcare Security Administration (NHSA) standard[1] to ensure a scientifically grounded framework for sub-specialty analysis. It defines 72 distinct disease clusters, balancing the need for specificity with statistical reliability. To resolve significant data sparsity encountered with the original standard, this final number was reached through an expert-guided consolidation process: merging analogous sparse groups while retaining medically unique ones to preserve specificity.
- **Level 3: ICD-10 Codes.** This most granular tier, defined by individual ICD-10 codes, enables fine-grained, diagnostic-level assessment. It is critical for identifying a model's specific strengths and weaknesses across numerous conditions. This level of detail, obscured in broader categories, gives developers the specific feedback required to improve their models and datasets.

3.2 STAGE 1: CASE CONSTRUCTION

With the taxonomic framework established, the first stage of our pipeline focuses on building a high-quality, structured corpus of contemporary clinical cases. This stage directly counters the challenges of **knowledge obsolescence** and **narrow disease scope** identified in existing benchmarks.

Case Curation The process begins by programmatically retrieving all XML-formatted case reports published in the first half of 2025 from the PubMed Central (PMC) **Open Access** Subset. A custom-built pipeline then parses each file, extracting key metadata before analyzing the article's structure. Sections describing the patient's journey (e.g., *Case Presentation*) are aggregated to form the core *case narrative*, while sections containing author analysis (e.g., *Discussion*) are consolidated into the *case discussion*. To enable multimodal proficiency assessment, this process also converts all tabular data into Markdown and extracts persistent URLs for all associated figures with their captions.

Sampling To construct a balanced and representative corpus, we first classify each case report against all three tiers of our taxonomy using `gpt-4.1-2025-04-14`, with detailed methodology and validation presented in Appendix E. With cases fully classified, we then implement a stratified sampling protocol guided by the 72 Level-2 disease clusters. Our protocol aims to sample 30 unique cases per cluster, while prioritizing the diversity of unique Level-3 diseases within each sample to mitigate the overrepresentation of common conditions. This rigorous procedure yielded a final corpus of 2,150 high-quality case reports, which served as the foundation for the subsequent stages.

3.3 STAGE 2: EXAM GENERATION

With the curated cases as our foundation, this stage focuses on generating questions that mirror the **entire clinical pathway**, moving beyond static, single-point assessments. While the challenge is achieving this at scale without sacrificing clinical nuance, we resolve the tension between manual quality and automated scalability by employing a Generator-Critic architecture, whose effectiveness is validated in Section 5. This `o3`-powered, two-agent system is guided by distinct prompts detailed in Appendix F and transforms each case report into a high-quality simulation of the patient journey.

Scenario Generation This process is initiated by the *Generator* Agent, which reframes each case into a progressive clinical challenge. It begins by crafting a concise initial clinical scenario capturing only the information available upon patient arrival. Building on this, the agent generates a sequence

[1] `https://code.nhsa.gov.cn/search.html?sysflag=8`

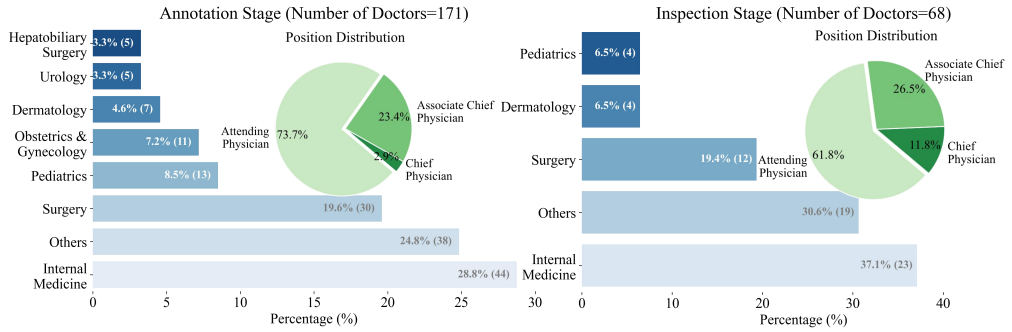

Figure 4: Distribution of the physician reviewers involved in the quality assurance process. The charts detail the team's composition by clinical specialty and professional rank (Chief Physician, Associate Chief Physician, and Attending Physician) for both the Annotation and Inspection stages.

of 3–6 progressive, 10-option MCQs. To make the clinical progression explicit, the agent dynamically assigns each question a *clinical stage* label (e.g., "Initial Assessment"), ensuring a logical flow from diagnosis to long-term management. Each question's context strategically introduces new clinical details at the appropriate workflow step, probing the model's ability to integrate evolving information.

Iterative Refinement The centerpiece of our methodology is the closed-loop quality control orchestrated by the *Critic* Agent. Once the *Generator* produces a question set, it enters an automated "peer-review" cycle where the *Critic* evaluates it on two key dimensions: ***Clinical Accuracy*** and ***Cognitive Complexity***. If a question is flagged, the *Critic* provides actionable feedback, prompting the *Generator* to revise the set. This refinement loop persists until the question set achieves two criteria: 100% Clinical Accuracy, ensuring all content is factually correct, and high Cognitive Complexity for over 60% of its questions. To ensure efficiency, any set failing to converge within 10 cycles is discarded. Applying this process to the 2,150 curated cases yielded 2,092 high-quality question sets.

3.4 Stage 3: Quality Check

Following AI-powered generation, we implement a multi-layered quality assurance protocol engineered to meet the uncompromising standards of medicine. This stage is governed by a principle of conservatism: **any question with a potential flaw is rejected**. The protocol combines AI pre-screening with multi-tiered physician verification. All evaluators apply two strict criteria: ***Factual Validation***, ensuring perfect alignment with the source case, and ***Logical Solvability***, confirming the answer is deducible from the available information. The specific prompts are detailed in Appendix G.

AI-Powered Pre-Screening Each generated question set first undergoes adjudication by a *Judge* Agent, implemented using o3, which acts as a highly conservative pre-filter. Systematically applying both checks by differentiating privileged from test-taker-visible information, its primary objective is to autonomously reject fundamentally flawed questions. As validated in Section 5, this AI-provided audit also serves to elevate the rigor of the subsequent physician review. This process streamlines the expert review, ultimately narrowing the pool from 2,092 to 1,869 high-potential candidates.

Physician Screening The centerpiece of our quality assurance is a rigorous, two-phase verification process conducted by **239** licensed physicians. Question sets that pass AI pre-screening first enter the *Annotation* phase, where attending physicians from top-tier hospitals evaluate each question against the aforementioned criteria. Subsequently, in the *Inspection* phase, more senior physicians conduct a full review of these annotations. Any identified discrepancy triggers a revision loop with the original annotator until consensus is reached, ensuring exceptional quality.

As detailed in Figure 4, the physician team is structured to ensure both broad specialty coverage and deep expertise, with senior experts (Chief and Associate Chief Physicians) constituting 26.3% of the Annotation team and 38.2% of the Inspection team. This process requires all materials to be translated for native-chinese experts experts via o3, whose accuracy is validated in Appendix H. The entire effort amounted to **1,772.18 person-hours** at a cost of $24 per hour, for a total expenditure of **$42,304.39**. This final layer of scrutiny yielded 1,822 definitively validated question sets. From this pool, we constructed our final benchmark of 1,407 sets through a stratified sampling protocol, selecting 20 cases per Level-2 cluster while prioritizing Level-3 disease diversity.

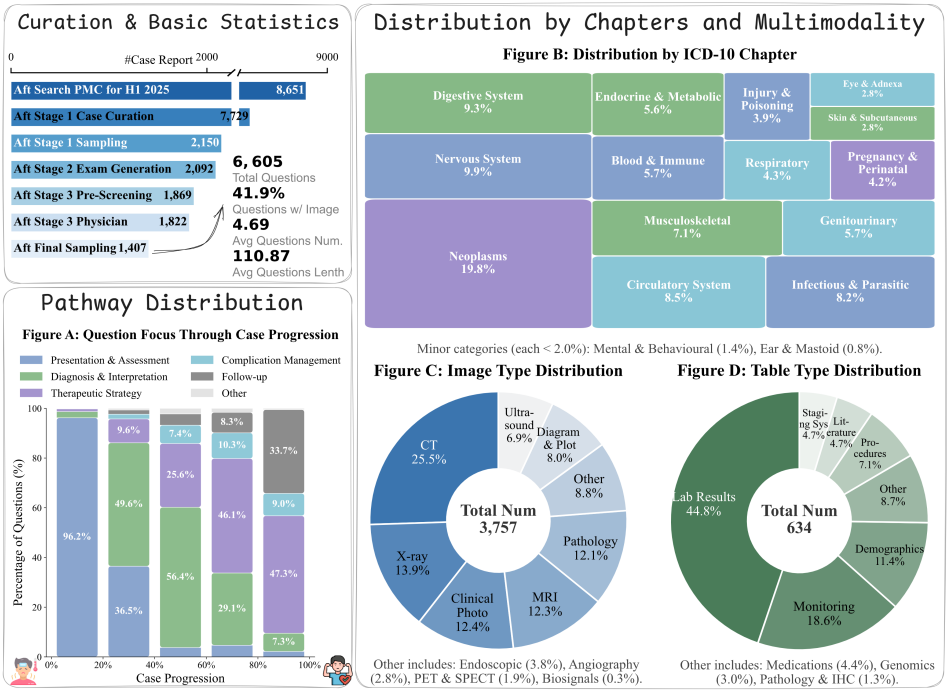

Figure 5: Overview of LiveClin's composition and statistics. The **top-left panel** details the case attrition funnel from construction pipeline and presents summary statistics of the final benchmark. **Figure A** shows the distribution of question focus across the entire clinical pathway. **Figure B** displays the distribution by ICD-10 chapters. **Figures C and D** detail the distribution of image and table modalities, respectively, highlighting the benchmark's multimodal nature.

3.5 Benchmark Statistics and Composition

The final LiveClin benchmark comprises 6,605 questions from 1,407 unique clinical cases, averaging 4.69 questions per case to form a coherent narrative. This section provides a statistical overview of the benchmark's key properties, summarized in Figure 5,

Distribution by Clinical Pathway A key feature of LiveClin is its simulation of the entire clinical pathway. As illustrated in Figure 5A, the question focus dynamically shifts as a case progresses to test a model's ability to handle the entire patient journey. Questions begin overwhelmingly centered on *Presentation & Assessment* (96.2% in the first quintile of a case). As more information becomes available, the focus transitions to *Diagnosis & Interpretation* and *Therapeutic Strategy* in the mid-stages. Towards the conclusion, questions concerning *Follow-up* and *Complication Management* become prominent, completing the simulation of a realistic clinical pathway.

Distribution by Chapters and Multimodality LiveClin provides extensive coverage across specialties and data modalities. The benchmark spans 16 distinct ICD-10 chapters (Figure 5B), led by complex areas like *Neoplasms* (19.8%) and other major specialties such as the *Nervous System* (9.9%) and *Digestive System* (9.3%). The benchmark also heavily emphasizes multimodal proficiency, with 41.9% of all questions requiring direct data interpretation. The dataset incorporates 3,757 images and 634 tables. As shown in Figures 5C and 5D, this includes a diverse array of medical images (e.g., *CT*, *X-ray*, *MRI*) and structured tabular data (e.g., *Lab result*, *Monitoring*, *Demographics*).

4 Experiments

In this section, we present a comprehensive evaluation of leading large language models (LLMs) on the LiveClin benchmark to assess their clinical reasoning capabilities. We first detail the Experimental Setup (§4.1), including the models tested and our evaluation protocol. We then report the Overall Performance Evaluation (§4.2), providing a high-level comparison of model capabilities. Following this, an In-depth Analysis (§4.3) explores model performance across different clinical domains and modalities, leveraging the fine-grained structure of LiveClin.

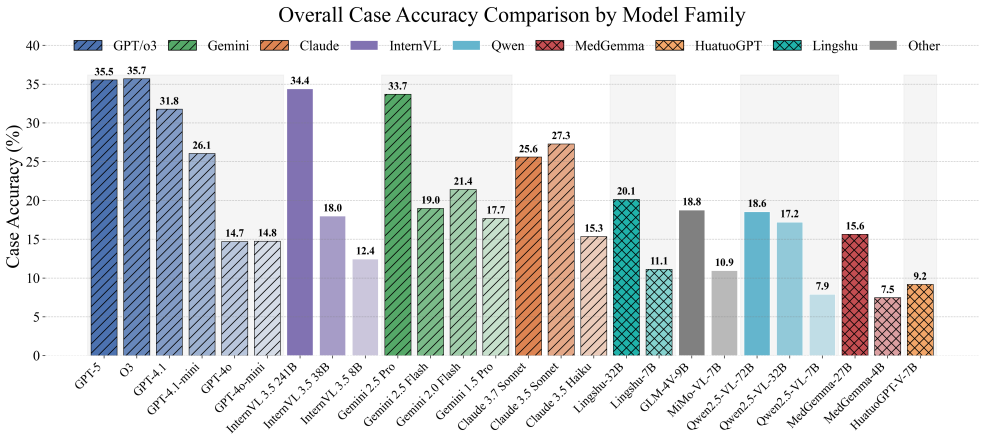

Figure 6: Overall Case Accuracy on the LiveClin benchmark, with models grouped by family and sorted in reverse chronological order (most recent to the left). The bar texture encodes the model type: diagonal hashing for proprietary models, solid colors for general open-source models, and cross-hatching for open-source medical models.

4.1 Experimental Setup

Evaluated Models We conduct evaluation on a comprehensive set of 26 models. Our selection spans three key categories: proprietary models like *GPT-5* (OpenAI et al., 2024), powerful open-source general LLMs like *Qwen2.5-VL* (Bai et al., 2025), and medical-specific LLMs like *MedGemma* (Sellergren et al., 2025). The complete list of all model versions is available in Appendix K.

Evaluation Protocol and Metrics To faithfully simulate sequential clinical encounters, we employed a conversational, zero-shot evaluation protocol. The full conversation history is maintained as context for each subsequent question, forcing the model to continuously integrate new information. For reproducibility, we set `temperature` to 0 for most models, adopting official recommended configurations for those with specific reasoning modes. Our primary metric is Case Accuracy, a stringent measure where a case is deemed correct only if *all* of its sequential questions are answered correctly. Further details on our prompting strategy are provided in Appendix K.

4.2 Overall Performance Evaluation

As presented in Figure 6, the overall Case Accuracy of a diverse suite of leading LLMs provides key insights into the current landscape of clinical AI. These results highlight both state-of-the-art capabilities and emerging challenges in model development.

Present Hierarchy: A Gap Between Giants and Rising Stars. The results show a clear performance hierarchy. Proprietary models currently lead the field, with *o3* (35.7%) and *GPT-5* (35.5%) setting the top scores. However, the benchmark's profound difficulty is evident, as even the best models fail on nearly two-thirds of cases, and a significant performance gap remains between the top tier and smaller models like *GPT-4o-mini* (14.8%). Meanwhile, open-source models are rapidly closing this gap. The large-scale *InternVL-3.5-241B* (34.4%) achieves near-parity with the top commercial models, while more efficient models like *GLM-4V-9B* (18.8%) demonstrate striking performance for their size, significantly outperforming weaker proprietary offerings like *GPT-4o* (14.7%).

Future Trajectory: An End to the "Free Lunch" Scaling. Our findings challenge the assumption that simple model scaling or releasing a newer version guarantees better performance on specialized clinical reasoning. For instance, *Claude 3.5 Sonnet* (27.3%) clearly outperforms its successor, *Claude 3.7 Sonnet* (25.6%). This trend is also visible within the Gemini family, where the older *Gemini 2.0 Flash* (21.4%) surpasses the newer *Gemini 2.5 Flash* (19.0%). This signals an end to the "free lunch" era of general upgrades, highlighting the need for targeted, domain-specific optimizations. This is particularly evident in the medical-specific models. While recent entries like *Lingshu-32B* (20.1%) show notable progress, they still trail far behind top generalist models, with a performance gap of over 14%. This underscores the need for the medical AI community to build future specialized models on more capable foundations to truly advance clinical reasoning capabilities.

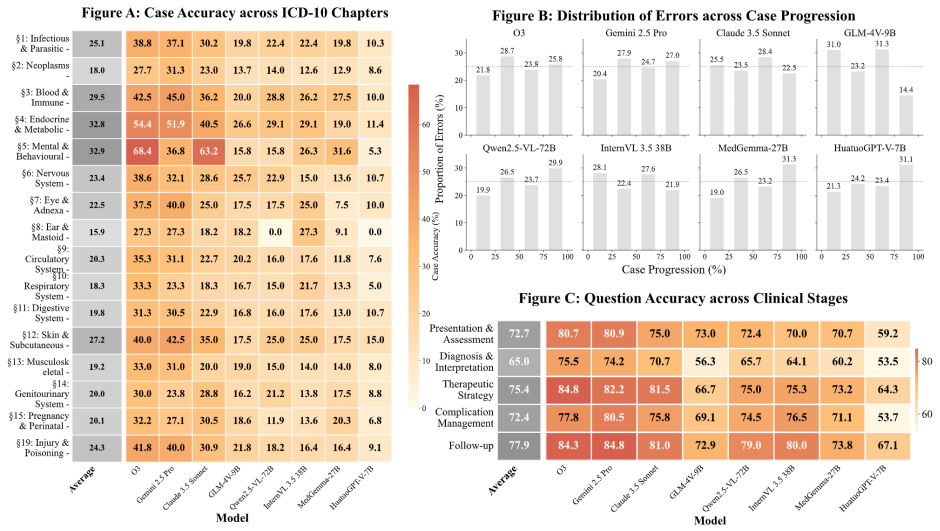

Figure 7: **Fine-grained Performance Analysis of Representative Models. Figure A** shows Case accuracy (%) across 16 major ICD-10 chapters. **Figure B** details Distribution of error proportion (%) across case progression. **Figure C** displays Question accuracy (%) across five clinical stages.

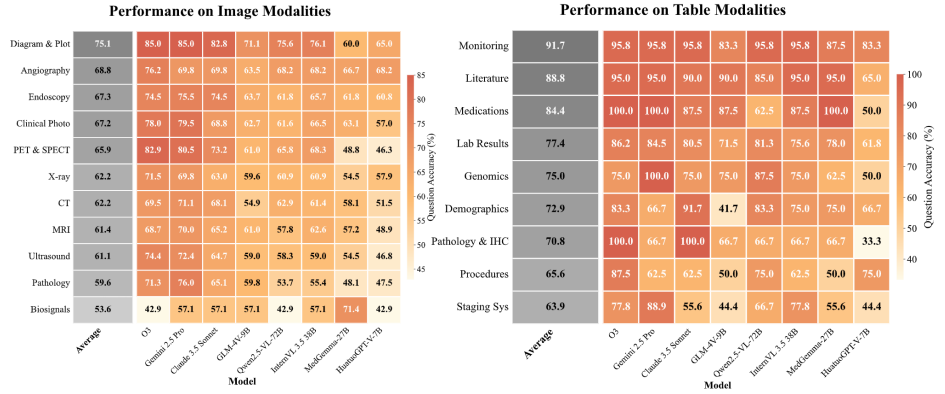

Figure 8: **Performance Analysis across Multimodal Sub-types.** Question accuracy (%) on distinct image and table modalities. Modality sub-types are sorted by the average accuracy across all models.

4.3 In-depth Analysis

To dissect model capabilities beyond aggregate scores, we conducted a fine-grained analysis of eight representative models across clinical domains, the cilinical pathway, and data modalities. This reveals not just what models get wrong, but how and when their reasoning fails.

Errors in Motion: Failure Modes Along the Pathway. An analysis of error patterns across the clinical pathway reveals distinct failure modes characteristic of different model classes (Figure 7B-C). Top proprietary models like *o3* tend to fail mid-pathway, with errors peaking during the cognitively demanding Diagnosis & Interpretation phase. In contrast, open-source medical models exhibit the late-stage failure pattern, where errors cluster in the final quartile during the less complex Follow-up stage, suggesting a critical breakdown in long-context retention. Finally, general-purpose models such as *GLM-4V-9B* exhibit a front-loaded error profile, stumbling early in the process. This highlights an urgent need to improve their ability to reason effectively from the initial clinical presentation.

Errors in Place: Strengths and Weaknesses in Domains. Our analysis of ICD-10 chapters reveals that model performance is highly variable, uncovering distinct specializations alongside universal weaknesses (Figure 7A). For instance, models excel in areas governed by clear systemic logic, such as Endocrine Diseases, yet falter universally in domains that demand nuanced synthesis, like Neoplasms. Interestingly, this specialization transcends scale: both the top-tier *o3* (68.4%) and the compact *GLM-4V-9B* (63.2%) achieve exceptional accuracy in Mental & Behavioural Disorders.

Method	Accuracy (%)	Trivial Ratio (%)	Time (hrs)†	Costa ($)
Physicians	92.5	38.5	188.9	4534.30
Generator	84.5	16.5	0.13	35.34
Generator-Critic	93.0	5.5	0.45	221.69
Generator-Critic-Judge	89.5*	5.5	0.55	244.19

Table 1: Ablation study of exam generation methods. Accuracy is the physician-validated pass rate. Costs are normalized per 100 case reports. †Time denotes person-hours for humans vs. API run-time for AI. *The accuracy on the Judge-approved subset is 98.4%.

This pattern of domain-specific proficiency extends to multimodal reasoning, where a critical gap separates simple data extraction from complex inference (Figure 8). Models confidently interpret structured data like Diagrams (75.1%) but struggle when expert-level reasoning is required, evidenced by poor performance on modalities such as Pathology (59.6%) and Biosignals (53.6%). Although specialized training shows promise, with *MedGemma-27B* displaying a surprising aptitude for Biosignals (71.4%), foundational robustness remains a key challenge. Even the most capable models can falter on seemingly simple inputs like Demographics tables, underscoring this core issue.

5 Ablation Study of the Agent Workflow

To validate the contribution of each component in our agent-based pipeline, we conducted an ablation study on randomly sampled 200 case reports. We benchmarked our full *Generator-Critic-Judge* pipeline against three alternatives: a physician-authored baseline, a Generator-only model, and a Generator-Critic pipeline. Each configuration was evaluated for accuracy, cost, and complexity. We quantified complexity as the proportion of 'trivial' questions, which are defined as questions correctly solved by all three baseline models (gpt-4o-mini, claude-3-5-haiku, and gemini-1.5-pro). The detailed comparison across all evaluation metrics is presented in Table 1.

Automated Exam Creation: Scaling with *Generator* and *Critic*. The results first demonstrate the profound impact of LLM on achieving scalability and complexity. The baseline *Generator* agent, operating alone, reduces both time and financial costs by nearly two orders of magnitude compared to human authoring, making a continuously updated benchmark feasible. It also produces inherently more challenging questions, more than halving the trivial question ratio from 38.5% to 16.5%. The introduction of the *Critic* agent is then indispensable for ensuring factual integrity. This *Generator-Critic* workflow boosts physician-validated accuracy from 84.5% to 93.0%, surpassing the human-authored baseline while further shrinking the trivial ratio to just 5.5%. This iterative refinement loop is thus essential for producing reliable and challenging content at scale.

Augmented Quality Check: Enhancing Verification with *Judge*. The final *Judge* agent functions not merely as a filter but as a crucial enhancement to the human review process. Although its inclusion nominally lowers the pass rate to 89.5%, this decrease signifies a positive outcome: a more rigorous quality standard. By providing physicians with a structured audit trail and direct evidence from the source case, the *Judge* enables them to identify subtle flaws that might otherwise be overlooked. By providing a structured audit trail and highlighting potential issues, *Judge* empowers physicians to conduct their validation with efficiency and rigor.

6 Conclusion

To combat the threats of data contamination and knowledge obsolescence in medical LLM evaluation, we introduced **LiveClin**, a dynamic benchmark built from a constantly refreshed stream of contemporary, peer-reviewed case reports. Our AI-powered workflow transforms these cases into multimodal challenges that span the entire clinical pathway. Evaluation on LiveClin reveals a stark performance gap, with a top Case Accuracy of just 35.7%, and uncovers distinct failure modes across the clinical journey, such as mid-case synthesis struggles in top models and late-stage context loss in specialized ones. LiveClin marks a paradigm shift from static knowledge testing to the dynamic assessment of applied clinical reasoning. By providing a continuously evolving and clinically-grounded challenge, we aim to guide the development of medical LLMs towards greater real-world reliability and safety.

REFERENCES

Rahul K. Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, Johannes Heidecke, and Karan Singhal. Healthbench: Evaluating large language models towards improved human health, 2025. URL https://arxiv.org/abs/2505.08775.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions, 2025. URL https://arxiv.org/abs/2505.23281.

Junying Chen, Chi Gui, Ruyi Ouyang, Anningzhe Gao, Shunian Chen, Guiming Hardy Chen, Xidong Wang, Ruifei Zhang, Zhenyang Cai, Ke Ji, Guangjun Yu, Xiang Wan, and Benyou Wang. Huatuogpt-vision, towards injecting medical visual knowledge into multimodal llms at scale, 2024. URL https://arxiv.org/abs/2406.19280.

Michael W Cullen, Jeffrey B Geske, Nandan S Anavekar, Julie A McAdams, Mary Ellen Beliveau, Steve R Ommen, and Rick A Nishimura. Reinvigorating continuing medical education: meeting the challenges of the digital age. In *Mayo Clinic Proceedings*, volume 94, pp. 2501–2509. Elsevier, 2019.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. Investigating data contamination in modern benchmarks for large language models, 2024. URL https://arxiv.org/abs/2311.09783.

Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering, 2020. URL https://arxiv.org/abs/2003.10286.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. Livecodebench: Holistic and contamination free evaluation of large language models for code, 2024. URL https://arxiv.org/abs/2403.07974.

Yixing Jiang, Kameron C. Black, Gloria Geng, Danny Park, James Zou, Andrew Y. Ng, and Jonathan H. Chen. Medagentbench: A virtual ehr environment to benchmark medical llm agents. *NEJM AI*, 2(9):AIdbp2500144, 2025. doi: 10.1056/AIdbp2500144. URL https://ai.nejm.org/doi/full/10.1056/AIdbp2500144.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. What disease does this patient have? a large-scale open domain question answering dataset from medical exams, 2020. URL https://arxiv.org/abs/2009.13081.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, and Victor Tseng. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLOS Digital Health*, 2(2):e0000198, 2023. doi: 10.1371/journal.pdig.0000198.

J. Lau, S. Gayen, Asma Ben Abacha, Dina Demner-Fushman, Shyam Visweswaran, and Yanshan Wang. A dataset of clinically generated visual questions and answers about radiology images. *Scientific Data*, 5:180251, 2018. doi: 10.1038/sdata.2018.251. URL https://doi.org/10.1038/sdata.2018.251.

Sharon Mitchell, Julien-Carl Phaneuf, Silvia Matilda Astefanei, Sissel Guttormsen, Amy Wolfe, Esther de Groot, and Carolin Sehlbach. A changing landscape for lifelong learning in health globally. *Journal of CME*, 12(1):2154423, 2023.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL https://arxiv.org/abs/2303.08774.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=KS8mIvetg2.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. Medmcqa : A large-scale multi-subject multi-choice dataset for medical domain question answering, 2022. URL https://arxiv.org/abs/2203.14371.

Samuel Schmidgall, Rojin Ziaei, Carl Harris, Eduardo Reis, Jeffrey Jopling, and Michael Moor. Agentclinic: a multimodal agent benchmark to evaluate ai in simulated clinical environments, 2025. URL https://arxiv.org/abs/2405.07960.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. Medgemma technical report, 2025. URL https://arxiv.org/abs/2507.05201.

Core Team, Zihao Yue, Zhenru Lin, Yifan Song, Weikun Wang, Shuhuai Ren, Shuhao Gu, Shicheng Li, Peidian Li, Liang Zhao, Lei Li, Kainan Bao, Hao Tian, Hailin Zhang, Gang Wang, Dawei Zhu, Cici, Chenhong He, Bowen Ye, Bowen Shen, Zihan Zhang, Zihan Jiang, Zhixian Zheng, Zhichao Song, Zhenbo Luo, Yue Yu, Yudong Wang, Yuanyuan Tian, Yu Tu, Yihan Yan, Yi Huang, Xu Wang, Xinzhe Xu, Xingchen Song, Xing Zhang, Xing Yong, Xin Zhang, Xiangwei Deng, Wenyu Yang, Wenhan Ma, Weiwei Lv, Weiji Zhuang, Wei Liu, Sirui Deng, Shuo Liu, Shimao Chen, Shihua Yu, Shaohui Liu, Shande Wang, Rui Ma, Qiantong Wang, Peng Wang, Nuo Chen, Menghang Zhu, Kangyang Zhou, Kang Zhou, Kai Fang, Jun Shi, Jinhao Dong, Jiebao Xiao, Jiaming Xu, Huaqiu Liu, Hongshen Xu, Heng Qu, Haochen Zhao, Hanglong Lv, Guoan Wang, Duo Zhang, Dong Zhang, Di Zhang, Chong Ma, Chang Liu, Can Cai, and Bingquan Xia. Mimo-vl technical report, 2025a. URL https://arxiv.org/abs/2506.03569.

LASA Team, Weiwen Xu, Hou Pong Chan, Long Li, Mahani Aljunied, Ruifeng Yuan, Jianyu Wang, Chenghao Xiao, Guizhen Chen, Chaoqun Liu, Zhaodonghui Li, Yu Sun, Junao Shen, Chaojun Wang, Jie Tan, Deli Zhao, Tingyang Xu, Hao Zhang, and Yu Rong. Lingshu: A generalist foundation model for unified multimodal medical understanding and reasoning, 2025b. URL https://arxiv.org/abs/2506.07044.

V Team, Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, Shuaiqi Duan, Weihan Wang, Yan Wang, Yean Cheng, Zehai He, Zhe Su, Zhen Yang, Ziyang Pan, Aohan Zeng, Baoxu Wang, Bin Chen, Boyan Shi, Changyu Pang, Chenhui Zhang, Da Yin, Fan Yang, Guoqing Chen, Jiazheng Xu, Jiale Zhu, Jiali Chen, Jing Chen, Jinhao Chen, Jinghao Lin, Jinjiang Wang, Junjie Chen, Leqi Lei, Letian Gong, Leyi Pan, Mingdao Liu, Mingde Xu, Mingzhi Zhang, Qinkai Zheng, Sheng Yang, Shi Zhong, Shiyu Huang, Shuyuan Zhao, Siyan Xue, Shangqin Tu, Shengbiao Meng, Tianshu Zhang, Tianwei Luo, Tianxiang Hao, Tianyu Tong, Wenkai Li, Wei Jia, Xiao Liu, Xiaohan Zhang, Xin Lyu, Xinyue Fan, Xuancheng Huang, Yanling Wang, Yadong Xue, Yanfeng Wang, Yanzi Wang, Yifan An, Yifan Du, Yiming Shi, Yiheng Huang, Yilin Niu, Yuan Wang, Yuanchang Yue, Yuchen Li, Yutao Zhang, Yuting Wang, Yu Wang, Yuxuan Zhang, Zhao Xue, Zhenyu Hou, Zhengxiao Du, Zihan Wang, Peng Zhang, Debing Liu, Bin Xu, Juanzi Li, Minlie Huang, Yuxiao Dong, and Jie Tang. Glm-4.5v and glm-4.1v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning, 2025c. URL https://arxiv.org/abs/2507.01006.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang,

Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Xidong Wang, Guiming Hardy Chen, Dingjie Song, Zhiyi Zhang, Zhihong Chen, Qingying Xiao, Feng Jiang, Jianquan Li, Xiang Wan, Benyou Wang, and Haizhou Li. Cmb: A comprehensive medical benchmark in chinese, 2024. URL https://arxiv.org/abs/2308.08833.

Colin White, Samuel Dooley, Manley Roberts, Arka Pal, Benjamin Feuer, Siddhartha Jain, Ravid Shwartz-Ziv, Neel Jain, Khalid Saifullah, Sreemanti Dey, Shubh-Agrawal, Sandeep Singh Sandha, Siddartha Venkat Naidu, Chinmay Hegde, Yann LeCun, Tom Goldstein, Willie Neiswanger, and Micah Goldblum. Livebench: A challenging, contamination-limited LLM benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=sKYHBTAxVa.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, Jianpeng Jiao, Xiangsheng Li, Jingkai Liu, Shuang Ni, Zhoufutu Wen, Ge Zhang, Kaiyuan Zhang, Xin Zhou, Jose Blanchet, Xipeng Qiu, Mengdi Wang, and Wenhao Huang. Futurex: An advanced live benchmark for llm agents in future prediction, 2025. URL https://arxiv.org/abs/2508.11987.

Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, Oishi Banerjee, Julián N. Acosta, Josh Miller, Ouwen Huang, and Pranav Rajpurkar. Rexrank: A public leaderboard for ai-powered radiology report generation, 2024. URL https://arxiv.org/abs/2411.15122.

Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. Don't make your llm an evaluation benchmark cheater, 2023. URL https://arxiv.org/abs/2311.01964.

Qin Zhu, Qingyuan Cheng, Runyu Peng, Xiaonan Li, Tengxiao Liu, Ru Peng, Xipeng Qiu, and Xuanjing Huang. Inference-time decontamination: Reusing leaked benchmarks for large language model evaluation, 2024. URL https://arxiv.org/abs/2406.13990.

Yuxin Zuo, Shang Qu, Yifei Li, Zhangren Chen, Xuekai Zhu, Ermo Hua, Kaiyan Zhang, Ning Ding, and Bowen Zhou. Medxpertqa: Benchmarking expert-level medical reasoning and understanding, 2025. URL https://arxiv.org/abs/2501.18362.

A USE OF LARGE LANGUAGE MODELS (LLMS)

To ensure the linguistic accuracy and fluency of this manuscript, a Large Language Model (LLM) was employed as a writing-enhancement tool. The use of the LLM was primarily focused on two areas:

- **Grammar and Spell Checking:** Identifying and correcting grammatical errors, typos, and punctuation mistakes.
- **Wording and Phrasing Refinement:** Optimizing word choice and sentence structure to enhance clarity, flow, and academic tone.
- **Data Construction:** Constructing Benchmark as Agent.

It is important to state that the LLM was used solely for improving the language and presentation. All the core ideas and conclusions presented in this paper are the original work of the author.

B RELATED WORK

Medical Datasets The evaluation of medical LLMs has rapidly evolved to address the complexities of clinical practice. Initial efforts focused on static, text-only question-answering datasets like MedQA (Jin et al., 2020), MedMCQA (Pal et al., 2022), and medical subsets of MMLU (Hendrycks et al., 2021). While foundational, these benchmarks primarily test static knowledge recall. To incorporate visual data, multimodal benchmarks such as VQA-RAD (Lau et al., 2018) and PathVQA (He et al., 2020) were developed, with recent efforts like MedXpertQA (Zuo et al., 2025) significantly raising the bar by introducing expert-level difficulty and rich clinical context. However, these approaches largely retain a single-turn, static format. A parallel line of work has explored multi-step evaluation. Agent-based benchmarks like AgentClinic (Schmidgall et al., 2025) and MedAgentBench (Jiang et al., 2025) simulate clinical workflows, though often with synthetic scenarios. A distinct approach is presented by HealthBench (Arora et al., 2025), which evaluates open-ended, multi-turn conversations using thousands of physician-designed rubric criteria to assess not just accuracy but also safety and communication quality. Finally, benchmarks like CMB (Wang et al., 2024) have highlighted the importance of localization. LiveClin complements these diverse efforts by tackling two critical, unaddressed dimensions: the **temporal decay of medical knowledge**, which makes static benchmarks quickly obsolete, and the evaluation of **sequential reasoning**, grounded in authentic, contemporary patient cases with explicit long-tail disease coverage.

Data Contamination and Live Benchmarks The vulnerability of static benchmarks to data contamination is a well-documented problem, as models can achieve inflated scores by memorizing test items seen during pretraining rather than demonstrating true reasoning capabilities (Oren et al., 2024; Zhou et al., 2023). This risk is particularly acute in high-stakes medical contexts, where the leakage of data from sources like licensing exams can create a false sense of security about a model's real-world clinical utility (Kung et al., 2023; Zhang et al., 2024). To address these challenges, dynamic or "live" evaluation paradigms have emerged as a robust solution, designed to counter contamination by continuously introducing unseen, time-stamped data. This approach has been successfully implemented across diverse domains. LiveBench (White et al., 2025) established a foundational model by automatically sourcing new questions from recent public data, which was extended to specialized fields with LiveCodeBench (Jain et al., 2024) using challenges from recent coding contests. Pushing the paradigm further, FutureX (Zeng et al., 2025) introduced a live benchmark for agent-based future prediction, a task requiring real-time information retrieval. While these pioneering efforts effectively mitigate data leakage, they remain primarily text-centric and do not address the tripartite challenge of clinical evaluation: the need to integrate diverse **multimodal** data, assess complex **sequential reasoning** in realistic patient workflows, and ensure **comprehensive disease coverage**. LiveClin bridges this critical gap. It adapts the live paradigm to the clinical domain by grounding evaluation in a constantly refreshed corpus of authentic, peer-reviewed patient cases, and uniquely embeds temporally-staged multimodal evidence within a structured workflow to enable a dynamic assessment of clinical reasoning that spans the full ICD-10 taxonomy.

C Pilot Study Methodology

This appendix provides the detailed methodology for the longitudinal pilot study presented in Section 2. The following sections detail the dataset construction and evaluation protocol, highlighting key distinctions from the main LiveClin pipeline that were made to specifically isolate the impact of data recency in a cost-effective manner.

Dataset Construction The construction process began with the random sampling of an initial pool of case reports from the PubMed Central (PMC) Open Access Subset for five distinct biannual periods (2023-S1 to 2025-S1). This initial sampling strategy differs from the stratified approach used for the main benchmark, as the goal here was to capture a typical temporal cross-section rather than a taxonomically balanced one.

These cases were then processed through the AI-powered generation and refinement pipeline described in Section 3, utilizing both the `Generator-Critic` architecture for question creation and the `Judge` Agent for automated quality screening. In a key departure from our main pipeline and to ensure cost-efficiency for this preliminary study, the multi-tiered physician verification phase was omitted. This AI-only pipeline resulted in a variable number of successfully generated question sets for each temporal slice. To ensure a fair and balanced comparison across periods, we performed a final random sampling step to create five standardized test sets, each containing 200 question sets.

Model Evaluation Protocol For the evaluation phase of the pilot study, we selected a representative set of leading proprietary models, as these are often the subject of contamination discussions and their performance provides a meaningful signal of the state-of-the-art. The evaluation itself was conducted under conditions identical to our main experiments to ensure the comparability of our findings. We employed the exact same sequential, zero-shot evaluation protocol detailed in Section 4 and Appendix K. This consistency in the testing framework ensures that any observed performance differences can be confidently attributed to the changing temporality of the data, not variations in the evaluation method.

D LiveClin Clinical Taxonomy

This appendix provides the complete details of the three-level clinical taxonomy introduced in Section 3.1. This hierarchical framework is a foundational component of LiveClin, designed to enable a multi-resolution performance analysis that moves beyond the single-score, narrow-scope evaluations of traditional benchmarks.

Table 2 details the full structure, mapping the 16 Level-1 Clinical Chapters to their constituent 72 Level-2 Disease Clusters. Level 3, the most granular tier, corresponds to the specific ICD-10 codes used for fine-grained analysis of individual conditions. The design of this taxonomy, particularly the expert-guided consolidation at Level 2, was crucial for balancing analytical specificity with the statistical reliability needed for robust model evaluation.

E Details for Case Report Classification

To automatically classify each case report into our three-level taxonomy, we employed `gpt-4.1-2025-04-14`. We designed a hierarchical, two-stage classification pipeline to enhance accuracy and consistency by breaking down the complex task into simpler, sequential steps. The process first identifies the broad clinical chapter (Level 1) and then, based on that result, determines the more specific disease cluster (Level 2) and ICD-10 code (Level 3).

Stage 1: Level-1 Chapter Classification In the first stage, the model is provided with the case report's title and abstract, along with the complete list of 16 Level-1 chapters. Its task is to select the single most relevant chapter for the case. The prompt used for this stage is summarized in Figure 9.

Stage 2: Level-2 Cluster and Level-3 ICD-10 Classification Once the Level-1 chapter is determined, the pipeline proceeds to the second stage. The model is again given the case's title and abstract, but this time it is provided with a constrained list of options containing only the Level-2 disease clusters and Level-3 ICD-10 codes that fall under the previously identified chapter. This significantly reduces the search space and improves precision. The model's task is to select the most fitting Level-2 cluster

ICD-10 Chapters (Level 1)	Disease Clusters (Level 2)
§1: Infectious & Parasitic	Intestinal Inf. & Tuberculosis (A00-19); Zoonotic bacterial (A20-28); Bacterial, Spirochetal, Chlamydial Inf., etc. (A30-79); Viral Inf. (CNS, HIV, etc.) (A80-B34); Fungal & Protozoal Inf. (B35-64); Parasitic & Helminthic (B65-98)
§2: Neoplasms	Malignant Neoplasms of Lip, Oral, Pharynx (C00-14); Digestive Organs (C15-26); Respiratory & Intrathoracic (C30-39); Bone & Cartilage (C40-41); Skin (C43-44); Mesothelial & Soft Tissue (C45-49); Breast (C50); Female Genital Organs (C51-58); Male Genital Organs (C60-63); Urinary Tract (C64-68); Eye, Brain & CNS (C69-72); Thyroid & Endocrine (C73-75); Lymphoid Tissues (C76-96); Benign, In Situ. (C97, D00-48)
§3: Blood, Blood-forming Organs & Immune Mechanism	Nutr, Hemol& Aplast Anemias (D50-64); Coagulation Defects, Purpura & Hemorrhagic Cond. (D65-69); Blood-Forming Organ (D70-77); Immune Mechanism Disord. (D80-89)
§4: Endocrine, Nutritional & Metabolic	Thyroid Disorders. (E00-07); Diabetes mellitus (E10-14); Endocrine, Glucose Reg. & Nutritional Deficiencies (E15-64); Obesity & Metabolic Disorders (E65-90)
§5: Mental & Behavioural Disord.	Comprehensive Mental & Behavioural Disord. (F00-98)
§6: Nervous System	Inflammatory & systemic atrophic of CNS (G00-14); Extrapyramidal, Movement, & Degen. Disord. (G20-32); Demyelinating of CNS (G35-37); Episodic & paroxysmal disord. (G40-47); Nerve root & plexus disord. (G50-59); Polyneuropathies & other PNS disord. (G60-64); Muscle, Myoneural, & Paralytic Disord. (G70-99)
§7: Eye & Adnexa	Disord. of eyelid, lacrimal system, orbit, conjunctiva, sclera, etc. (H00-22); Disord. of Lens, Retina, Glaucoma, Globe, vitreous body, optic nerve, visual pathways, etc. (H25-59)
§8: Ear & Mastoid Process	Comprehensive Ear & Mastoid Disord. (H60-95)
§9: Circulatory System	Rheumatic, Hypertensive, & Ischemic Heart (I00-25); Pulmonary heart & circulation (I26-28); Other forms of heart (I30-52); Cerebrovascular (I60-69); Arteries, arterioles & capillaries (I70-79); Venous, Lymphatic (I80-99)
§10: Respiratory System	Acute & Chronic Respiratory Infections and Disorders (J00-39); Chronic, Environmental & Pleural (J40-94); Other diseases of the respiratory system (J95-99)
§11: Digestive System	Oral Cav, Saliv Glands & Jaws (K00-14); Esoph, Stomach, Appendix (K20-38); Liver (K70-77); Gallbldr, Biliary Tract, Pancreas (K80-93); Hernia (K40-46); Noninfective enteritis & colitis (K50-52); Intestinal & Peritoneal Disorders (K55-67)
§12: Skin & Subcutaneous Tissue	Infectious, Bullous & Eczematous Skin. (L00-30); Papulosquamous, Urticarial, Radiation & Miscellaneous Skin. (L40-99)
§13: Musculoskeletal & Connective Tissue	Arthropathies (M00-25); Systemic connective tissue disord. (M30-36); Dorsopathies (M40-54); Soft tissue disord. (M60-79); Osteopathies, Chondropathies, etc. (M80-99)
§14: Genitourinary System	Glomerular, Tubulo-Interstitial & Renal Failure Disord. (N00-19); Urolithiasis and Other Kidney & Ureter Disord. (N20-29); Urinary System & Male Genital Organ Disord. (N30-51); Breast & Female Reproductive System. (N60-99)
§15: Pregnancy, Childbirth & Perinatal Period	Pregnancy with Abortive Outcome & Maternal Disord. (O00-48); Labor, Delivery & Puerperal Complications (O60-92); Fetus Affected by Maternal, Perinatal Factors (P00-83)
§19: Injury, Poisoning & External Causes	Injuries & Foreign Bodies by Body Region (S00-T19); Burns, Frostbite, & Drug Poisoning (T20-50); Toxic Effects of Substances Chiefly Nonmedicinal as to Source (T51-65)
Total	72

Table 2: Hierarchical Taxonomy of the LiveClin Benchmark. The table displays the first two levels of the taxonomy. Level 3, the most granular tier, corresponds to the specific ICD-10 codes.

Prompt for Level-1 Classification

Based on the following medical case report title and abstract, please identify the most relevant chapter from the medical classification system provided below.
[List of all 16 Level-1 Chapters is inserted here]
Case Report Title: *"title"* Case Report Abstract: *"abstract"*
Instructions: Provide your classification after a '—' separator line. Format the response exactly as follows:
Level1: Chapter *[Number]*: *[Chapter Name]*

Figure 9: The prompt designed to classify a case report into one of the 16 high-level clinical chapters.

and the most specific Level-3 ICD-10 code from these filtered lists. The core structure of this prompt is shown in Figure 10.

Validation To validate the reliability of this automated pipeline, we randomly sampled 200 case reports from our corpus. We then invited clinical experts to manually assign the appropriate Level-1, Level-2, and Level-3 labels to these cases. The model's classifications were compared against these expert annotations, achieving accuracies of **98.5%** for Level-1 Chapters, **97.5%** for Level-2 Disease Clusters, and 97.0% for Level-3 ICD-10 codes. These high accuracy rates confirmed the method's suitability for constructing our benchmark.

F PROMPTS FOR *Generator* AND *Critic* AGENT

As described in Section 3.3, our iterative refinement pipeline is powered by a *Generator-Critic* architecture. Both agents are instances of the o3 large language model, guided by distinct, highly-

Prompt for Level-2 and Level-3 Classification

Based on the medical case report, which belongs to *[Pre-identified Level-1 Chapter]*, please provide:
The most specific sub-category (Level 2) from the list below.
The most specific ICD-10 code WITH its disease name that matches this case.
Available Level-2 categories for this chapter (YOU MUST CHOOSE ONE FROM THIS LIST):
[List of relevant Level-2 Disease Clusters is inserted here]
Allowed ICD-10 codes in this chapter (you MUST copy exactly one line from below):
[List of relevant Level-3 ICD-10 Codes is inserted here]
Case Report Title: *"title"* Case Report Abstract: *"abstract"*
Instructions: Provide your classifications after a '—' separator line. Format EXACTLY as follows:
Level2: *[Sub-category Name] ([Code Range])* ICD-10: *[Code] [Disease Name]*

Figure 10: The prompt for the second classification stage. It uses the Level-1 result to provide a constrained set of options for Level-2 and Level-3 classification.

Prompt for `Generator` Agent

Goal: Generate an initial clinical scenario and 3-6 progressive MCQs (10 options each) based *only* on the provided case report sections.
CRITICAL CONSTRAINTS:

- **FORBIDDEN CONTENT REFERENCES:** NEVER mention "case report," "case description," or "discussion." Test-takers only see the scenario, progressive question info, and Figures/Tables.
- **MULTIMODAL ASSESSMENT:** NEVER describe image findings in the question text. Only state the figure number and modality (e.g., "Figure 1 shows an ultrasound..."). Questions MUST test image interpretation ability.
- **DISTRACTOR QUALITY:** ALL 9 incorrect options MUST be clinically plausible distractors representing realistic differential diagnoses. They must require clinical reasoning to eliminate.

Instructions:

1. **Create an Initial Scenario:** Describe *only* the patient's initial presentation.
2. **Generate 3-6 Progressive MCQs:**
   - Frame questions as queries a clinician might ask an AI assistant.
   - Arrange questions along a logical clinical timeline (e.g., assessment -> diagnosis -> treatment).
   - Incrementally introduce new clinical findings from the case report within each question's stem.
3. **Output Format:** MUST output ONLY a valid JSON object containing the scenario and a list of MCQs.

Figure 11: The prompt for the `Generator` agent, which initiates the creation of the clinical examination by converting a case report into a progressive set of MCQs.

structured prompts to fulfill their specific roles. This section details the core instructions provided to each agent.

***Generator* Agent** The *Generator*'s primary role is to transform a static, unstructured case report into a dynamic, multi-step clinical reasoning challenge. It receives the full case report text, including figure/table captions and the discussion section, as its ground-truth context. Its prompt instructs it to synthesize this information into a progressive series of Multiple-Choice Questions (MCQs) that simulate a clinical workflow. Key constraints emphasize creating plausible distractors, ensuring questions test image interpretation rather than caption recall, and strictly forbidding any reference to the source "case report," which is invisible to the test-taker. The core of this initial generation prompt is outlined in Figure 11.

***Critic* Agent** The `Critic` agent orchestrates the quality control loop. After the `Generator` produces an initial set of MCQs, the `Critic` receives both this generated set and the original ground-truth case report. Its prompt (Figure 12) instructs it to perform a meticulous "peer review" of each MCQ. The evaluation is structured around two key dimensions: *Clinical Accuracy* (verifying factual correctness against the source report and ensuring sufficient information is present to answer)

Prompt for `Critic` Agent

Your Task: You are an expert medical exam question critic. Evaluate **each of the MCQs** provided below. For each one, consider the initial scenario, all information revealed in preceding MCQs, and the ground-truth Case Report Context.
Evaluation Criteria for EACH MCQ:

1. **Correctness & Clarity:**
   - **Information Sufficiency:** Based *strictly* on the information available to the test-taker at this stage, is there enough detail to unambiguously arrive at the correct answer?
   - **Reference Violations:** Does the question improperly reference invisible content like the "case description"?
   - **Distractor Logic:** Can each incorrect option be eliminated with clear clinical reasoning based on the available information?
2. **Difficulty & Cognitive Level:**
   - **Challenge Level:** Is the question challenging enough for a clinician-level AI assistant?
   - **Distractor Quality:** Are the incorrect options strong enough to require complex reasoning to eliminate?

Output Format: For each MCQ, provide a JSON object containing two evaluations: *correctness_evaluation* and *difficulty_evaluation*. Each must include a boolean flag (*is_correct_and_clear*, *is_sufficiently_difficult*) and a detailed text field for *critique_and_suggestions*.

Figure 12: The prompt for the *Critic* agent, designed to systematically evaluate the generated MCQs for clinical accuracy and cognitive complexity.

Prompt for *ReGenerator* Phase

Goal: Revise and improve a set of medical exam questions based on critiques.
Context:

1. **Original Patient Case Report:** [Full text of the source case report]
2. **Previously Generated Exam Content:** [The full scenario and list of MCQs from the previous attempt]
3. **Critique and Suggestions for Improvement:** [The structured feedback from the *Critic* agent]

Your Task: Based on all the information above, **regenerate the entire exam content (initial scenario and all MCQs)**, addressing ALL feedback provided in the "Critique and Suggestions." Your primary focus is to fix the identified issues.

Figure 13: The prompt used during the iterative refinement loop. It provides the *Generator* with the *Critic*'s feedback to guide the revision process.

and *Cognitive Complexity* (assessing whether questions demand high-level reasoning and feature challenging distractors). The agent must output its structured feedback in a JSON format, providing specific critiques and a binary judgment for each dimension.

***ReGenerator* Phase** If the *Critic* flags any issues, the process enters a revision phase. The *Generator* agent is invoked again, but with an expanded prompt. This "ReGenerator" prompt (Figure 13) includes the original case report, the previously generated (flawed) MCQs, and, crucially, the structured feedback from the *Critic*. The agent's task is to regenerate the entire examination set, specifically addressing every point of critique to improve the quality of the questions. This loop continues until the *Critic*'s criteria are fully met.

G Prompts for *Judge* Agent and Doctors

The multi-layered quality assurance protocol described in Section 3.4 is guided by a unified set of evaluation criteria applied to both AI and human reviewers. The core of this protocol is a prompt designed to enforce a strict, two-part validation for every question. This ensures that each question is not only factually correct according to the source material but also logically solvable from the perspective of an examinee who lacks access to that privileged information.

Prompt for Quality Assurance *Judge* Agent

Your Task: You are a QA Verifier for Medical Case Simulations. Perform a rigorous, two-part validation on the provided question set by adopting two distinct personas.
Core Directive: Information Segregation

- **Privileged Information (For Factual Validation Only):** The "Source Case Report" and "Image Captions". This is 100% invisible to the test-taker.
- **Test-Taker Visible Information (For Logical Solvability Only):** The initial scenario, question stems, images (without captions), and answers from previous steps.

Your Two-Part Validation Task for Each Question:
1. Factual Validation (as a "Backend Auditor"):

- Using the **Privileged Information**, is the question's premise, data, and correct answer factually correct and perfectly consistent with the source case?

2. Logical Solvability (as an "Examinee"):

- Using **ONLY the Test-Taker Visible Information**, is there a clear, logical path to the single correct answer? Erase all memory of the privileged information for this check.

Output Format Requirement: You MUST output your assessment as a JSON object. For each MCQ, provide a "factual_validation" and a "logical_solvability" block, each containing a boolean verdict ("is_factually_correct" / "is_logically_solvable") and a detailed "explanation".

Figure 14: The prompt for the *Judge* agent, which enforces a strict two-part validation focused on factual correctness and logical solvability.

***Judge* Agent Prompt** For the AI pre-screening step, a *Judge* agent (an instance of *o3*) is provided with the full context: the ground-truth case report and the finalized set of questions. The agent's prompt, summarized in Figure 14, explicitly instructs it to adopt two distinct personas for evaluation. First, as a "backend auditor," it performs *Factual Validation* by comparing the question against the source case. Second, as an "examinee," it performs *Logical Solvability* by assessing if the answer can be deduced using only the information presented up to that point. The agent must output its findings in a structured JSON format, giving a clear verdict for each criterion.

Instructions for Physician Reviewers Questions that pass the AI pre-screening are then subjected to final verification by human clinical experts. The physicians are provided with the same materials and instructed to follow the exact same two-part validation protocol as the *Judge* agent. They first act as auditors, using the source case to perform ***Factual Validation***. Then, they switch to an examinee's perspective to confirm ***Logical Solvability***. This parallel process ensures our principle of conservatism is upheld: any question identified with a potential flaw by either the AI or the human experts is rigorously scrutinized and ultimately rejected if the issue cannot be resolved.

H Translation Quality Validation

To facilitate the rigorous expert review process with our native Chinese-speaking physicians, all English-language question sets were systematically translated. Ensuring the fidelity of this translation was paramount. To this end, we conducted a comparative validation study to quantify the consistency between translations generated by our *o3*-powered API and those produced by professional human translators.

Validation Methodology and Results We randomly sampled 100 complete question sets for this study. For each set, we generated two parallel Chinese versions: the first was produced by our *o3*-powered API, while the second, serving as a gold-standard reference, was created by two professional biomedical translators. These two translated versions were then presented side-by-side to a panel of three native Chinese-speaking physicians for a comparative review.

Crucially, the physicians did not see the original English text. Their task was to determine if the two Chinese versions were clinically and semantically identical. A pair was marked 'consistent' only if the *o3* translation conveyed the exact same medical facts, nuances, and logical relationships as the expert human translation. The final inter-translation consistency rate was **99.2%**, confirming that the API-powered translation is a highly reliable substitute for manual expert translation for this task.

Prompt for Comparative Translation Validation

Goal: Your task is to determine if two Chinese translations of the same medical content are clinically and semantically identical.
Context: You will be presented with pairs of text segments (e.g., a question stem, an option, or a scenario).

- **Translation A:** The text produced by our AI system.
- **Translation B:** The text produced by a professional human medical translator (this is your reference).

CRITICAL: You will NOT see the original English source. Your judgment must be based solely on comparing Translation A and Translation B.
Core Evaluation Question: Would a medical professional reading either translation arrive at the exact same clinical understanding? Consider all aspects:

1. **Factual Equivalence:** Are all medical facts, numbers, units, and terminologies identical?
2. **Nuance Equivalence:** Is the degree of certainty, tone, and clinical implication the same in both versions (e.g., "可能" vs. "确诊")?
3. **Logical Equivalence:** Does the logic of the question and the relationship between premise and conclusion remain unchanged?

Your Task: Provide a Verdict for Each Pair For each pair of translations, please provide one of the following two verdicts:

Verdict: Consistent Choose this if Translation A is clinically and semantically identical to the reference Translation B. A medical professional would interpret both in exactly the same way.

Verdict: Inconsistent Choose this if there is any meaningful difference between the two translations that could potentially alter clinical understanding, even if it's a subtle change in nuance or terminology.

Final Output: For each pair, provide your verdict ("Consistent" or "Inconsistent"). If you choose "Inconsistent," please briefly describe the discrepancy.

Figure 15: The prompt provided to native Chinese-speaking physicians to validate the consistency between AI-generated and human-expert translations.

Instructions for Comparative Review To standardize the assessment, each physician reviewer was guided by the prompt detailed in Figure 15. The prompt instructed them to perform a direct comparison and provide a binary judgment on the semantic equivalence of the two translations.

I Annotation of Clinical Stages

To analyze the sequential nature of clinical reasoning, we categorized each question into one of five primary workflow stages. This was achieved through a detailed, two-step clustering process designed to systematically organize the diverse stage labels produced by the AI into a consistent, analyzable framework.

Initial AI Labeling During the examination generation stage (Stage 2), our *Generator* agent assigned a descriptive, free-text "stage" label to each question. This initial step resulted in hundreds of unique, granular labels that captured the specific context of each question, such as "Post-operative follow-up," "Initial diagnostic imaging," or "Surgical Intervention planning."

Expert-Guided Clustering While granular, these free-text labels were too diverse for a high-level statistical analysis. We, therefore, implemented a two-level clustering protocol under physician supervision. First, we programmatically grouped the AI-generated labels into 8 intermediate clinical categories based on a set of keywords (e.g., all labels containing "therapy" or "surgical" were grouped together). Following this, our physician team reviewed and further consolidated these intermediate groups into the five final, clinically coherent workflow phases shown in Figure 5A. For example, the intermediate categories for "Therapeutic Planning" and "Surgical Intervention" were both mapped to the final *Therapeutic Strategy* stage. This expert-guided logic, which underpins the entire categorization, is summarized in Figure 16.

Clustering Logic for Clinical Stages

Principle: Map granular, AI-generated stage labels to broader categories based on clinical keywords.
Example Mappings:

- Keywords like "initial", "presentation", "assessment" → **Presentation & Assessment**
- Keywords like "imaging", "diagnostic", "pathology", "histology" → **Diagnosis & Interpretation**
- Keywords like "management", "therapy", "surgical", "planning" → **Therapeutic Strategy**
- Keywords like "complication", "deterioration", "adverse" → **Complication Management**
- Keywords like "follow-up", "surveillance", "post-operative" → **Follow-up**

Figure 16: The expert-defined logic used to cluster granular, AI-generated stage labels into the five primary clinical workflow categories.

Prompt for Initial Modality Classification

Your Task: You are a medical data specialist. Your task is to classify the given item (image or table) into its most specific and accurate type.
Input: You will be given the item's original caption and its content (for tables) or the image itself.
Instruction: Provide a concise, specific label describing the item.

- **For Images:** e.g., "CT Scan of the Abdomen," "Chest X-ray (PA view)," "H&E Stained Pathological Slide."
- **For Tables:** e.g., "Table of Baseline Patient Demographics," "Table of Serial Blood Test Results."

Figure 17: The prompt used to instruct the *o3* model to perform initial classification of images and tables.

J Annotation of Multimodal Content

The classification of all images and tables into standardized types followed a three-step, AI-human hybrid protocol. This approach was designed to leverage the scalability of AI while ensuring the uncompromising accuracy required for a medical benchmark.

1. **AI Classification:** We first employed the *o3* model to perform an initial classification of all **3,757 images** and **634 tables**. Guided by the prompt in Figure 17, the model assigned a specific, descriptive label (e.g., "CT Scan of the Abdomen," "Table of Baseline Patient Demographics") to each item based on its content and original caption.
2. **Physician Validation:** The complete set of AI-generated labels was then meticulously reviewed by our physician team. This crucial step served as a 100% audit of the AI's output. The experts verified the correctness of the AI's classification with **100% accuracy**, confirming its reliability for this task.
3. **Expert-Guided Clustering:** Finally, to create clear, high-level categories for analysis, these specific and now-validated labels were grouped based on expert-defined rules, as summarized in Figure 18. This consolidation, conducted under physician supervision, allowed us to analyze broad trends. For example, validated labels like "CT Angiography" and "Helical Tomography" were grouped into the "CT Scans" category.

K Evaluation Protocol Details

This section provides a detailed description of the models, parameters, and procedures used in our evaluation.

Model Suite and Parameters We evaluated a comprehensive suite of 24 models, which can be grouped into three main categories:

- **Proprietary Models (13):** Leading closed-source models from major AI labs.
  The models tested were: *o3*, *gpt-5*, *gemini-2.5-pro*, *gpt-4.1*, *claude-3-5-sonnet-20241022*, *claude-3-7-sonnet-20250219-thinking*, *gemini-1.5-pro*, *claude-3-5-haiku-20241022*, and their high-efficiency

Clustering Logic for Modalities

Principle: Map specific, validated labels into broader analytical categories based on expert-defined rules.

Image Clustering Examples:

- Labels like "CT Angiography", "Helical Tomography" → **CT Scans**
- Labels like "Radiography", "Fluoroscopy" → **Radiography & X-ray**
- Labels like "Fundus Photograph", "Dermoscopy" → **Clinical & Surface Photography**

Table Clustering Examples:

- Labels like "Blood Test Results", "Biochemistry Panel" → **Lab Results**
- Labels like "Patient Demographics", "Vital Signs Table" → **Patient Demographics & Characteristics**
- Labels like "Longitudinal Treatment Response" → **Treatment Outcomes & Monitoring**

Figure 18: The expert-defined logic used to group specific, validated modality labels into the broader categories shown in the final statistics.

Prompt for Evaluation

Scenario:
Initial clinical scenario text, along with any associated figures and tables.
Previous Q&A Turns:
Full text of previous questions, options, and the model's prior answers are inserted here for context.
Current Question:
Text of the current question.
Options:
A. Option A text B. Option B text ... J. Option J text
Please provide the letter of the correct option, formatted as \boxed{LETTER} (e.g., \boxed{A}).

Figure 19: The conversational prompt structure used for evaluation. Context from previous turns is cumulatively added for subsequent questions within the same case.

variants: *gpt-4.1-mini-2025-04-14*, *gemini-2.5-flash*, *gemini-2.0-flash*, *gpt-4o-2024-11-20*, and *gpt-4o-mini-2024-07-18*.

- **Open-Source General-Purpose VLMs (7):** Powerful, publicly available Vision-Language Models. The models tested were: *Qwen2.5-VL-72B-Instruct*, *Qwen2.5-VL-32B*, *Qwen2.5-VL-7B-Instruct* (Bai et al., 2025), *InternVL3.5-38B*, *InternVL3.5-241B-A28B*, *InternVL3.5-8B* (Wang et al., 2025), *GLM-4.1V-9B* (Team et al., 2025c), and *MiMo-VL-7B-RL-2508* (Team et al., 2025a).
- **Open-Source Medical-Specific VLMs (4):** Models specifically fine-tuned on medical data. The models tested were: *medgemma-27b-it*, *medgemma-4b-it* (Sellergren et al., 2025), *HuatuoGPT-Vision-7B* (Chen et al., 2024), *LingShu-7B* and *LingShu-32B* (Team et al., 2025b).

To balance reproducibility with optimal performance, *temperature* was set to 0 for most models. For those featuring specific reasoning modes (e.g., "thinking" variants), we adopted their official recommended inference configurations. If an API call failed due to transient errors like rate limits or timeouts, it was automatically retried up to two times with a brief delay.

Prompting Strategy Our protocol maintains a continuous conversation history for each case. The prompt for each question is structured to build upon all previous turns. Figure 19 illustrates the general format. For the first question in a case, only the *Scenario* and *Current Question* blocks are used. For subsequent questions, the *Previous Q&A Turns* block is populated with the full history of preceding questions and the model's own answers. This conversational context is critical for testing sequential reasoning.

Answer Parsing To robustly extract the model's final choice, we implemented a hierarchical parsing strategy. The system prioritizes answers enclosed in a *\boxed{}* command (e.g., *\boxed{A}*). If this is not found, it searches for explicit keywords (e.g., "The correct answer is: A"). As a final fallback, it looks for single-letter answers at the end of the response. This multi-layered approach ensures high-fidelity extraction of the model's intended answer.