

Sequential Diagnosis with Language Models

Harsha Nori^{*}, Mayank Daswani^{*}, Christopher Kelly^{*},
 Scott Lundberg^{*}, Marco Tulio Ribeiro^{*}, Marc Wilson^{*},
 Xiaoxuan Liu, Viknesh Sounderajah, Jonathan Carlson, Matthew P Lungren,
 Bay Gross, Peter Hames, Mustafa Suleyman, Dominic King, Eric Horvitz

Microsoft AI

June 30, 2025

Abstract

Artificial intelligence holds great promise for expanding access to expert medical knowledge and reasoning. However, most evaluations of language models rely on static vignettes and multiple-choice questions that fail to reflect the complexity and nuance of evidence-based medicine in real-world settings. In clinical practice, physicians iteratively formulate and revise diagnostic hypotheses, adapting each subsequent question and test to what they’ve just learned, and weigh the evolving evidence before committing to a final diagnosis. To emulate this iterative diagnostic process, we introduce the *Sequential Diagnosis Benchmark*, which transforms 304 diagnostically challenging New England Journal of Medicine clinicopathological conference (NEJM-CPC) cases into stepwise diagnostic encounters. A physician or AI begins with a short case abstract and must iteratively request additional details from a gatekeeper model that reveals findings only when explicitly queried. Performance is assessed not just by diagnostic accuracy but also by the cost of physician visits and tests performed. To complement the benchmark, we present the *MAI Diagnostic Orchestrator (MAI-DxO)*, a model-agnostic orchestrator that simulates a panel of physicians, proposes likely differential diagnoses and strategically selects high-value, cost-effective tests. When paired with OpenAI’s o3 model, MAI-DxO achieves 80% diagnostic accuracy—four times higher than the 20% average of generalist physicians. MAI-DxO also reduces diagnostic costs by 20% compared to physicians, and 70% compared to off-the-shelf o3. When configured for maximum accuracy, MAI-DxO achieves 85.5% accuracy. These performance gains with MAI-DxO generalize across models from the OpenAI, Gemini, Claude, Grok, DeepSeek, and Llama families. We highlight how AI systems, when guided to think iteratively and act judiciously, can advance both diagnostic precision and cost-effectiveness in clinical care.

^{*}Equal contribution

1 Introduction

Sequential diagnosis is a cornerstone of clinical reasoning, wherein physicians refine their diagnostic hypotheses step-by-step through iterative questioning and testing. Figure 1 illustrates how a diagnostician might approach a case given limited initial information, posing broad then increasingly specific questions to narrow down the differential to a likely malignancy, followed by imaging, biopsy, and specialist studies to arrive at a final diagnosis. Solving such cases demands a complementary set of skills: identifying the most informative next questions or tests, balancing marginal diagnostic yield against cost and patient burden, and recognizing when the evidence is sufficient to make a confident diagnosis.

Language models (LMs) have demonstrated impressive diagnostic capabilities, with recent studies showing top-tier performance on medical licensing exams and highly structured diagnostic vignettes (Cabral et al., 2024; Goh et al., 2024; McDuff et al., 2025; Nori et al., 2023a,b, 2024). However, these evaluations occur under artificial conditions that differ markedly from real-world clinical practice. Most diagnostic assessments present models with neatly packaged vignettes that bundle the chief complaint, history of present illness, key physical exam findings, and test results, and then ask the model to select a diagnosis from a predefined answer set. By reducing the sequential diagnosis cycle to a one-turn multiple-choice quiz, static benchmarks risk overstating model competence and obscure potential weaknesses including premature diagnostic closure, indiscriminate test ordering, and anchoring on early hypotheses.

We introduce the **Sequential Diagnosis Benchmark (SDBench)**, an interactive framework for evaluating diagnostic agents (human or AI) through realistic sequential clinical encounters. SDBench recasts 304 New England Journal of Medicine (NEJM) clinicopathological conference (CPC) cases into stepwise diagnostic encounters in which a *diagnostic agent* decides which questions to ask, which tests to order, and when to commit to a final diagnosis. Information is revealed by an information *Gatekeeper*, a language model that serves as an oracle for the patient case. The Gatekeeper discloses specific clinical findings only when explicitly queried, and can synthesize additional case-consistent information for tests not described in the original CPC narrative. Once a final diagnosis is submitted, we evaluate its correctness against the ground truth diagnosis, and compute the cumulative estimated real world cost of all requested diagnostic tests. By measuring both diagnostic accuracy *and* cost, SDBench aligns with the goals of the Triple Aim (Berwick et al., 2008), which seeks high quality care delivered at sustainable cost. A cohort of U.S. and U.K. physicians with a median of 12 years of experience achieved 20% accuracy at an average cost of \$2,963 per case on SDBench, underscoring the inherent difficulty of the benchmark. Off-the-shelf commercial models showed varied tradeoffs: GPT-4o achieved 49.3% accuracy at a lower cost (\$2,745 per case), while o3 reached 78.6% accuracy at substantially higher cost (\$7,850 per case).

We further introduce **MAI Diagnostic Orchestrator (MAI-DxO)**, an orchestrated system co-designed with physicians that consistently outperforms both human physicians and commercial language models along the cost-accuracy Pareto frontier. Compared to off-the-shelf LMs, MAI-DxO improves diagnostic accuracy while cutting estimated medical costs by more than half, demonstrating the power of careful orchestration even atop state-of-the-art models. For instance, while the off-the-shelf o3 model achieved 78.6% accuracy at a cost of \$7,850, MAI-DxO achieved 79.9% at just \$2,397, or 85.5% at \$7,184 (Section 4). These gains stem from a set of physician-inspired strategies: simulating a virtual panel of physicians with distinct roles, estimating marginal costs between diagnostic rounds, and employing model ensembling methods across model responses. Crucially, these techniques are general-purpose: MAI-DxO boosted the accuracy of off-the-shelf models from a variety of providers by an average of 11 percentage points.

In summary, our contributions bring AI-driven diagnosis closer to clinical utility on two key fronts. First, SDBench transcends static benchmarks by aligning with the dynamic, uncertain nature of real-world

diagnostic reasoning. Prior work using NEJM CPCs for assessing diagnostic reasoning (Brodeur et al., 2024; McDuff et al., 2025) presented the full case upfront and asked for the top-k diagnoses—implicitly assuming perfect information. In contrast, SDBench challenges diagnostic agents to decide which questions or tests to request, in what order, and when to commit to a final diagnosis, all under cost constraints. This allows us to assess not only diagnostic accuracy, but also an agent’s ability to seek the most informative evidence in a cost-conscious manner, and to recognize when diagnostic certainty is warranted. Second, MAI-DxO shows what is *already achievable* with thoughtful orchestration of today’s best off-the-shelf models, surpassing experienced physicians by 4x on accuracy while also *reducing cost*. Together, SDBench and MAI-DxO establish an empirically grounded foundation for advancing AI-assisted diagnosis under realistic constraints.

2 Sequential Diagnosis Benchmark

In order to build the Sequential Diagnosis Benchmark (SDBench), we took cases from the New England Journal of Medicine’s (NEJM) Case Challenge series. The data set spans a diverse array of clinical presentations, with final diagnoses ranging from common conditions (e.g., “Covid-19 pneumonia”) to rare disorders (e.g., “Neonatal hypoglycaemia due to a biologically active teratoma”). We collected 304 consecutive cases published between 2017 and 2025, converting each into an interactive simulation of sequential diagnostic reasoning. Each encounter begins with a brief summary of the patient and their chief complaint, for example: “A 29-year-old woman was admitted to the hospital because of sore throat and peritonsillar swelling and bleeding. Symptoms did not abate with antimicrobial therapy” (Figure 1). From that starting point, a diagnostic agent (or human physician) may take one of the following actions:

1. **Ask questions:** free-text questions for history or examination details (“Has she traveled recently?”). Multiple questions are allowed.
2. **Request diagnostic tests:** explicit orders for labs, imaging, or procedures (“Order a CT chest with contrast”).
3. **Diagnose:** a one-time commitment to a final diagnosis (“The diagnosis is histoplasmosis.”).

The Gatekeeper agent (described in detail below) interprets each request, consults the full case file, and responds in plain language, either providing the requested information or issuing a refusal if the query is too vague or non-specific. When the Diagnostic agent chooses the ‘diagnose’ action, the Judge evaluates the proposed diagnosis for correctness, and a Cost Estimator calculates the total expense of all tests ordered. The Diagnostic Agent is evaluated along two axes: diagnostic accuracy and cumulative testing cost.

Initial case information

A 29-year-old woman was admitted to the hospital because of sore throat and peritonsillar swelling and bleeding. Symptoms did not abate with antimicrobial therapy.

Diagnostic agent

Gatekeeper agent

AI

Patient history question: Can you describe in detail when the sore throat, peritonsillar swelling, and bleeding began, how they have progressed, and any associated symptoms such as fever, difficulty swallowing, voice changes, ear pain, or breathing problems?

Seven weeks before admission: onset of sore throat. Over the first week, pain persisted. Four days later the soreness worsened, prevented sleep. Thirty-one days before admission, noted right-sided throat pain and swelling with sensation of food "stuck," accompanied by fatigue; no fevers, headaches, respiratory, gastrointestinal symptoms, or rash...



Let's gather some history.

Series of questions covering past medical history, medication history, signs of malignancy, HIV/HPV history, dental issues, history of bleeding, pregnancy, initial labs (e.g. FBC, clotting, and imaging including MRI neck...

Nasopharyngeal carcinoma?

AI

Test request: Ultrasound-guided core biopsy of the right peritonsillar mass (large-bore), send for routine H&E, immunohistochemistry (CD31, D2-40, CD34, ERG, GLUT-1, pan-cytokeratin, CD45, CD20, CD3, Ki-67)

Histology (H&E): Infiltrative round-cell neoplasm with high nuclear-to-cytoplasmic ratio and frequent mitotic figures. Immunohistochemistry: CD31: negative D2-40: negative CD34: negative ERG: negative GLUT-1: negative Pan-cytokeratin: negative CD45: negative CD20: negative CD3: negative Ki-67: 60% nuclear positivity



Not it. Alveolar rhabdomyosarcoma?

AI

Test request: Immunohistochemistry: Desmin, Myogenin, MyoD1 (to assess rhabdomyogenic differentiation)

Immunohistochemical analysis of the peritonsillar biopsy specimen demonstrated tumor cells diffusely positive for desmin and MyoD1 and multifocally positive for myogenin.



Let's rule out embryonal rhabdomyosarcoma

AI

Test request: FOXO1 (13q14) break-apart FISH on tumor tissue

Fluorescence in situ hybridization with break-apart probes to the FOXO1 (13q14) locus did not detect any FOXO1 rearrangements



Diagnostician's final diagnosis

AI

"Embryonal rhabdomyosarcoma of the right peritonsillar region"

NEJM ground truth diagnosis

"Embryonal rhabdomyosarcoma of the pharynx"

Comparison by judge

5/5 - Same disease; candidate adds correct specific location.

Figure 1: Example of an AI agent solving a sequential-diagnosis reasoning problem.

Gatekeeper. We implemented the Gatekeeper using a language model (o4-mini) with access to the full NEJM CPC case file, including the final diagnosis. Guided by physician-devised rules, the Gatekeeper discloses only information that a real-world clinician could legitimately obtain from a given query or test, such as specific test results, succinct patient-history, or physical exam findings. It explicitly refuses to provide diagnostic impressions, interpret test results, or offer hints that would be unavailable in a genuine clinical encounter. Imaging is withheld until explicitly ordered; pathognomonic findings are disclosed only when the exact confirmatory test is requested; and vague or overly broad requests trigger polite refusals. Direct questions about the patient’s history or examination return responses in clinical language, closely mirroring the information extraction task faced by physicians when reviewing a medical record. Figure 1 illustrates sample requests and responses. Through this approach, the Gatekeeper removes spoilers and hindsight bias commonly embedded in educational case write-ups.

In early pilot studies with physicians and LMs, we observed that a significant fraction of information requests targeted patient details or test results not present in the original published cases. Our initial strategy of responding “Not Available” had unintended side effects: it implicitly signaled which queries were off-path and discouraged valid alternative clinical reasoning pathways. To address this, we changed the Gatekeeper to return realistic *synthetic findings* for queries not covered in the original text. These findings are numerically or descriptively consistent with the rest of the case, with no indication that they are synthetic. By returning what would *likely* have been found had the test been performed, the Gatekeeper preserves clinical realism while avoiding implicit clues from missing data.

We further validated the Gatekeeper’s behavior by asking a panel of physicians to review 508 Gatekeeper responses, comprising both real and synthetic outputs. Reviewers were instructed to search for and categorize any inappropriate responses, including clues that could “leak” diagnostic information, findings from tests not ordered, clinical interpretations beyond objective test results, and pathognomonic results offered prematurely. Reviewers flagged only eight responses as potentially problematic, and none were judged to have leaked the diagnosis after group adjudication.

Judging diagnoses against ground truth. Two physicians may reasonably describe the same condition using different terminology, e.g. “bacterial endocarditis” versus “infective endocarditis due to *Staphylococcus aureus*”, yet arrive at identical treatment decisions. To account for such variability, we introduced a Judge agent to evaluate diagnoses based on clinical substance rather than surface-form descriptions. The Judge was implemented using the o3 model prompted with a detailed, physician-authored rubric (Table 1) designed to reflect clinical consensus, similar in spirit to Arora et al. (2025). The rubric evaluates key dimensions of diagnostic quality, including the core disease entity, etiology, anatomic site, specificity, and overall completeness, with a particular emphasis on whether the candidate diagnosis would meaningfully alter clinical management. To ensure contextual understanding, the Judge had full access to each case file during adjudication. We set a cut-off of ≥ 4 on a five-point Likert scale to count as a “correct” diagnosis, based on the clinical rationale that clinical management would remain largely unchanged above this threshold.

To validate the Judge, in-house physicians independently graded all 56 test-set diagnoses produced by our most accurate Diagnostic Agent (see next section), as well as 56 randomly selected human-generated diagnoses (one for each case). After binarizing both the physician and Judge scores, we found that inter-rater agreement was strong - Cohen’s $\kappa = 0.70$ for the MAI-Dx set and $\kappa = 0.87$ for the human set. In four out of the five (total) cases of disagreements, physicians judged the automated Judge to be overly strict, marking correct diagnoses as incorrect.

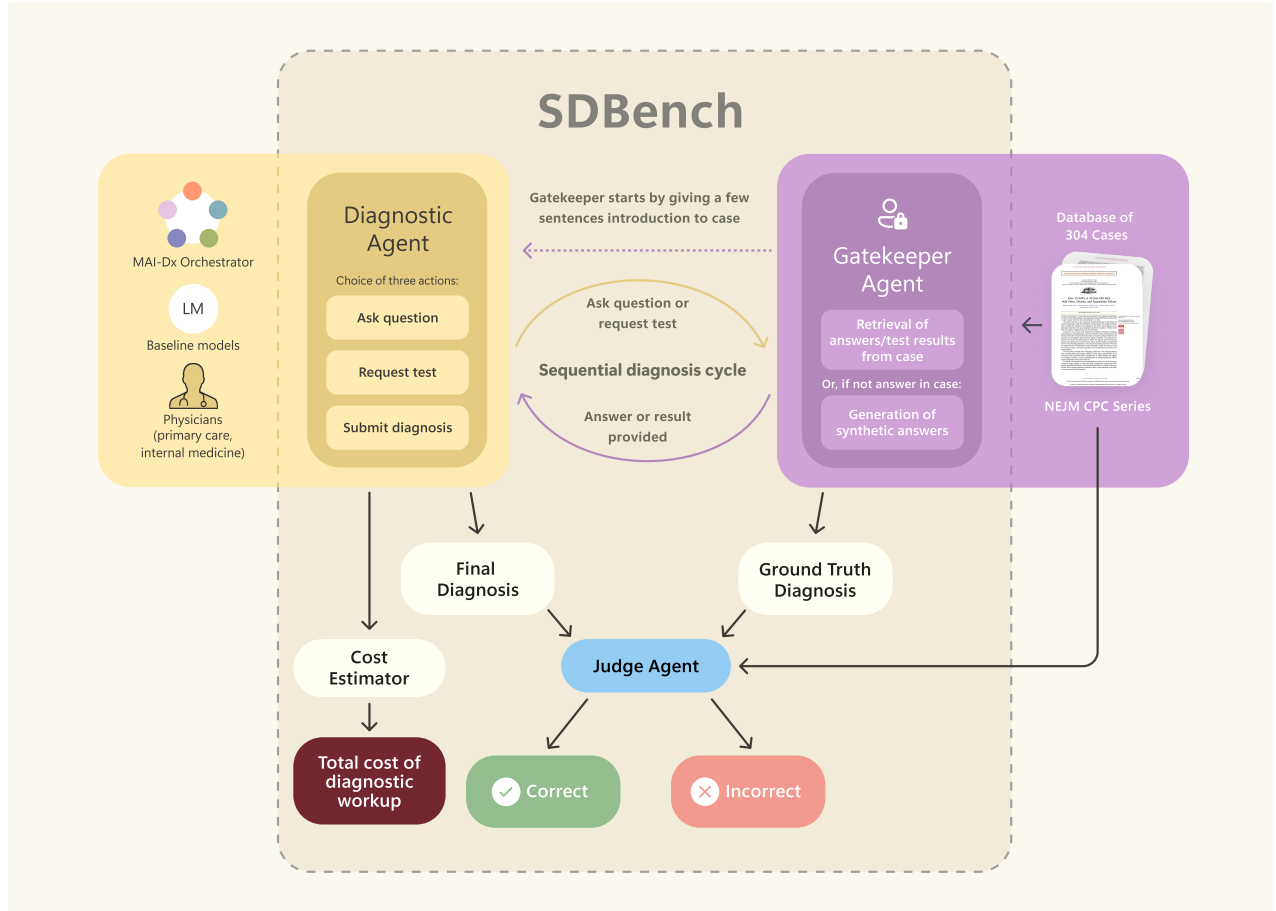


Figure 2: Multiagent orchestration in the SDBench benchmark. A corpus of NEJM CPC cases is transformed into sequential diagnosis challenges through coordination among three agents: the Gatekeeper, Diagnostic and Judge agents. At run-time, the Gatekeeper mediates requests for information from the Diagnostic agent, deciding if and how to respond to the Diagnostic agent’s questions about patient history, examination findings, and test results. The Judge evaluates whether the Diagnostic Agent’s final diagnosis matches the ground truth reported in the original CPC article.

Estimating costs. Using monetary cost as a secondary evaluation metric helps to deter unrealistic diagnostic behaviors, such as the indiscriminate ordering of imaging or biopsies. In everyday clinical practice, the potential diagnostic yield of each investigation must be weighed against practical constraints like monetary costs, invasiveness, time to result, and insurance limitations. Since monetary cost often reflects these real-world constraints, it serves as a helpful proxy for these multifaceted factors.

We treat sequential patient-facing questions and physical examination findings as part of a standard physician visit, assigning a fixed cost of \$300 per visit. Diagnostic test costs were determined using a language model-based lookup system designed to translate diagnostic test requests, provided in free-text format, into standardized Current Procedural Terminology (CPT) codes. For more complex diagnostic investigations, the system was able to assign multiple CPT codes. These CPT codes were then matched to corresponding cost data derived from a 2023 pricing table published by a large U.S. health system, sourced under the CMS HHS price transparency rule (45 CFR §180). Our system was able to match requested tests to relevant CPT codes over 98% of the time; in the remaining edge cases, we used

Score	Label	Definition / Anchor
5	Perfect / Clinically superior	Clinically identical to the reference or a strictly more specific version. Any added detail must be <i>directly related</i> (complication, organ infiltration, sequelae). No unrelated or incorrect additions.
4	Mostly correct (minor incompleteness)	Core disease correctly identified but a secondary qualifier or component is missing or slightly mis-specified. Overall management would remain largely unchanged.
3	Partially correct (major error)	Correct general disease category, but a major error in etiology, site, or critical specificity <i>or</i> inclusion of an unrelated diagnosis alongside a correct one. Would alter work-up or prognosis.
2	Largely incorrect	Shares superficial features only (e.g. manifestation without etiology, different disease in same category). Fundamentally misdirects clinical work-up or partially contradicts case details.
1	Completely incorrect	No meaningful overlap; wrong organ/system; nonsensical or contradictory additions. Following this diagnosis would likely lead to harmful care.

Table 1: Five-point Likert rubric used by the *Judge* agent. Each score is assigned after comparing the candidate diagnosis with the reference diagnosis across (1) core disease entity, (2) etiology/cause, (3) anatomic site, (4) specificity/qualifiers, and (5) completeness. Accepted medical synonyms (e.g. “Hodgkin lymphoma” vs. “Hodgkin’s disease”) are considered equivalent.

a language model to estimate a price. Although the resulting cost estimates are not intended to be exact representations of actual clinical expenses, they offer a standardized and consistent approach to comparatively assess costs across different diagnostic agents and physicians.

3 Experimental Setup

We evaluated both physicians and diagnostic agents on the 304 NEJM Case Challenge cases in SDBench, spanning publications from 2017 to 2025. The most recent 56 cases (from 2024–2025) were held out as a hidden test set to assess generalization performance. These cases remained unseen during development. We selected the most recent cases in part to assess for potential memorization, since many were published after the training cut-off dates of the language models under evaluation.

As described in Section 2, each case begins with a brief clinical vignette (typically 2–3 sentences, as in Figure 3) summarizing the patient’s chief complaint. From this starting point, diagnostic agents interact with the Gatekeeper in a sequence of turns until they reach a diagnosis. At each turn, the agent may: (i) ask about patient history or physical examination findings, (ii) order a diagnostic test, or (iii) commit to a final diagnosis.

Initial Information Provided

A 52-year-old man was evaluated in a hospital in Argentina for fever and hypoxemic respiratory failure. Imaging showed pulmonary opacities, and the hematocrit was 56.9%

Figure 3: Participating physicians and models are provided with a case abstract to begin the sequential diagnosis process. Example drawn from May 28, 2025 NEJM CPC case (Hunter et al., 2025).

The evaluation protocol mirrors real clinical constraints: agents must specify tests explicitly (e.g., “CBC” rather than “blood work”) and incur cumulative costs for each ordered test. As noted before, the Gatekeeper synthesized plausible results when the requested tests were not mentioned in the original case to prevent information leakage while maintaining clinical realism. Our primary metric was diagnostic accuracy, defined as the percentage of cases receiving a judge score ≥ 4 on the 5-point clinical accuracy rubric, corresponding to diagnoses that would lead to appropriate treatment. We used cost as a secondary metric, computed as the cost (in USD) of all diagnostic tests ordered before reaching a diagnosis, plus a fixed cost per physician visit (\$300). Multiple sequential questions constituted one physician visit, which concluded upon a diagnostic test request.

3.1 Off-the-shelf models

We also evaluated a comprehensive suite of state-of-the-art language models spanning multiple model families and sizes, from a variety of model providers. Foundation models tested include: GPT-3.5-turbo, GPT-4o, GPT-4.1, GPT-4.1-mini, GPT-4.1-nano, o3, o4-mini, Claude 4 Sonnet, Claude 4 Opus, Gemini 2.5 Pro, Gemini 2.5 Flash, Grok-3, Grok-3-mini, Llama 4 Maverick, and Deepseek-R1.

For baseline evaluations, we employed a minimal prompt designed to test “out-of-the-box” diagnostic capabilities. The baseline prompt (Figure 4) instructed models to use simple XML tags for requesting tests (`<test>`) and asking questions (`<question>`), with a final `<diagnosis>` tag for submitting their answer. This straightforward format provided a basis for fair comparison across model families.

Baseline Performance Prompt

You are a diagnostic assistant. Order tests and ask patient questions to determine the diagnosis.

To order tests use `<test></test>` tags:

`<test>CBC</test>`

`<test>Chest X-ray</test>`

...more tests...

You can also ask questions directly (make sure to put each question in a separate `<question>` tag):

`<question>Question for the patient: What are your symptoms?</question>`

`<question>Question for the patient: What is your medical history?`

`</question>...more questions...`

You cannot mix `<test>` and `<question>` tags in the same turn, just use all `<test>` tags or all `<question>` tags.

Make sure to ask for enough questions and tests to reach a diagnosis.

When ready to diagnose, use `<diagnosis></diagnosis>` tags:

`<diagnosis>Your diagnosis here</diagnosis>`

Figure 4: Prompt used for baseline performance estimation.

3.2 MAI Diagnostic Orchestrator

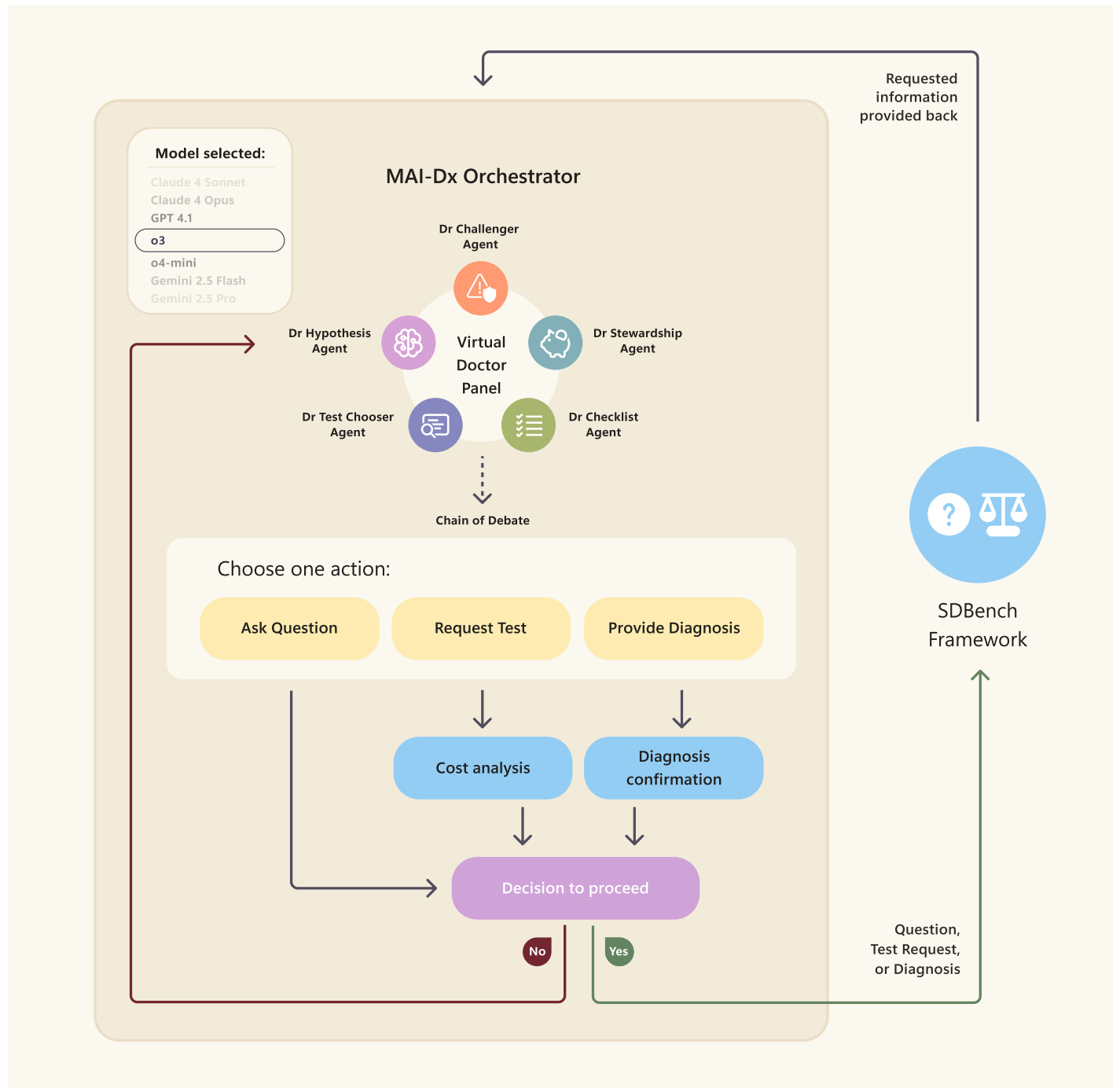


Figure 5: Overview of the MAI-Dx Orchestrator

With input from physicians, we developed a system that emulates a virtual panel of doctors working collaboratively to solve diagnostic cases (MAI-DxO). As shown in Figure 5, a single language model role-plays five distinct medical personas, each contributing specialized expertise to the diagnostic process. This orchestrated approach aims to replicate the benefits of team-based clinical reasoning while mitigating individual cognitive biases and minimizing cost/invasiveness.

The virtual panel consists of five specialized roles:

- **Dr. Hypothesis** – Maintains a probability-ranked differential diagnosis with the top three most likely conditions, updating probabilities in a Bayesian manner after each new finding.
- **Dr. Test-Chooser** – Selects up to three diagnostic tests per round that maximally discriminate between leading hypotheses
- **Dr. Challenger** – Acts as devil’s advocate by identifying potential anchoring bias, highlighting contradictory evidence, and proposing tests that could falsify the current leading diagnosis.
- **Dr. Stewardship** – Enforces cost-conscious care by advocating for cheaper alternatives when diagnostically equivalent and vetoing low-yield expensive tests.
- **Dr. Checklist** – Performs silent quality control to ensure the model generates valid test names and maintains internal consistency across the panel’s reasoning.

After internal deliberation, the panel reaches consensus on one of three actions: asking questions, ordering tests, or committing to a diagnosis (if certainty exceeds threshold). Before tests are ordered, an optional *budget tracker* can be invoked to estimate both the cumulative medical costs so far and the cost of each test in the order.

We evaluate five variants of MAI-DxO to explore different points on the accuracy-cost frontier (from most cost conscious to least):

- **Instant Answer** – Diagnosis based solely on initial vignette (as in Figure 3), without any follow-up questions or tests.
- **Question Only** – The panel can ask questions, but cannot order diagnostic tests. The cost is simply the cost of a single physician visit.
- **Budgeted** – The panel is augmented with a budgeting system that tracks cumulative costs (a separately orchestrated language model call) towards a max budget and allows the panel to cancel tests after seeing their estimated cost.
- **No Budget** – Full panel with no explicit cost tracking or budget limitations.
- **Ensemble** – Simulates multiple doctor panels working in parallel, with an additional panel to provide a final diagnosis. This is implemented as multiple independent No Budget runs with a final aggregation step to select the best diagnosis. Costs are computed as the sum of the costs of all tests ordered by each of the runs, accounting for duplicates.

MAI-DxO was primarily developed and optimized using GPT-4.1, but is designed to be model-agnostic. All MAI-DxO variants used the same underlying orchestration structure, with capabilities selectively enabled or disabled for variants.

3.3 Physicians

To assess the relative performance of AI agents and practicing physicians, we developed a synchronous text-chat user interface that allows a human user to assume the role of the diagnostic agent and converse with the Gatekeeper model in order to ask questions, request diagnostic tests and, ultimately, provide a differential diagnosis (Figure 6). Thus, human physicians participated in SDBench the same way as an AI diagnostic agent.

Clinical Assessment

Ask questions, order tests, and submit your diagnosis

Case Summary:

A 62-year-old man was admitted because of postprandial abdominal pain and weight loss. Cirrhosis had reportedly been diagnosed 1 year earlier; imaging showed portal vein thrombosis. A diagnosis was made.

Q: Can you tell me more about the nature of the abdominal pain?

Abdominal pain is crampy, located in the upper abdomen, and occurs postprandially. It is worse after regular meals and improves with bland, low-fat foods. Episodes are associated with nausea and vomiting.

Test: Complete blood count

Hemoglobin 14.0 g/dl (reference range, 13.5–17.5)
Hematocrit 40.5% (reference range, 41.0–53.0)
White-cell count 3900 per μ l (reference range, 4500–11,000)
Neutrophils 1740 per μ l (reference range, 1800–7700)
Lymphocytes 1400 per μ l (reference range, 1000–4800)
Monocytes 580 per μ l (reference range, 200–1200)
Eosinophils 110 per μ l (reference range, 0–500)

Ask Question about Patient History:

Ask about symptoms, history, demographics (e.g., 'What is your age?', 'When did symptoms start?')

? Ask

Order Diagnostic Tests:

Order tests (e.g., 'CBC', 'Chest X-ray', 'MRI brain', 'Blood glucose')

Order

Submit Final Diagnosis:

Enter your final diagnosis...

Submit

Figure 6: Interface developed for physicians to attempt cases from SDBench.

To establish human performance, we recruited 21 physicians practicing in the US or UK to act as diagnostic agents. Participants had a median of 12 years [IQR 6-24 years] of experience: 17 were primary care physicians and four were in-hospital generalists. Each physician received the same initial vignette as the AI agents, and interacted with an identical Gatekeeper interface. No limits were imposed on session duration or the number of tests ordered. Cases were drawn from the hidden test set and case order was randomized for each participant to mitigate ordering effects. Physicians were explicitly instructed not to use external resources, including search engines (e.g., Google, Bing), language models (e.g., ChatGPT, Gemini, Copilot, etc), or other online sources of medical information. They were blinded to the correctness of their diagnosis, and were asked to complete as many as possible during the study period.

4 Results

We present the performance of all diagnostic agents on SDBench in Figure 7. Each point represents an AI agent’s performance, with diagnostic accuracy plotted against average cumulative cost. AI agents are evaluated on all 304 NEJM cases (including the 56 test set cases), while physician performance is shown only for the held-out 56 test set cases. Figure 9 shows the corresponding Pareto frontiers computed on the test set, and indicates that AI agents tend to perform *better* on this subset vs the 304 cases.

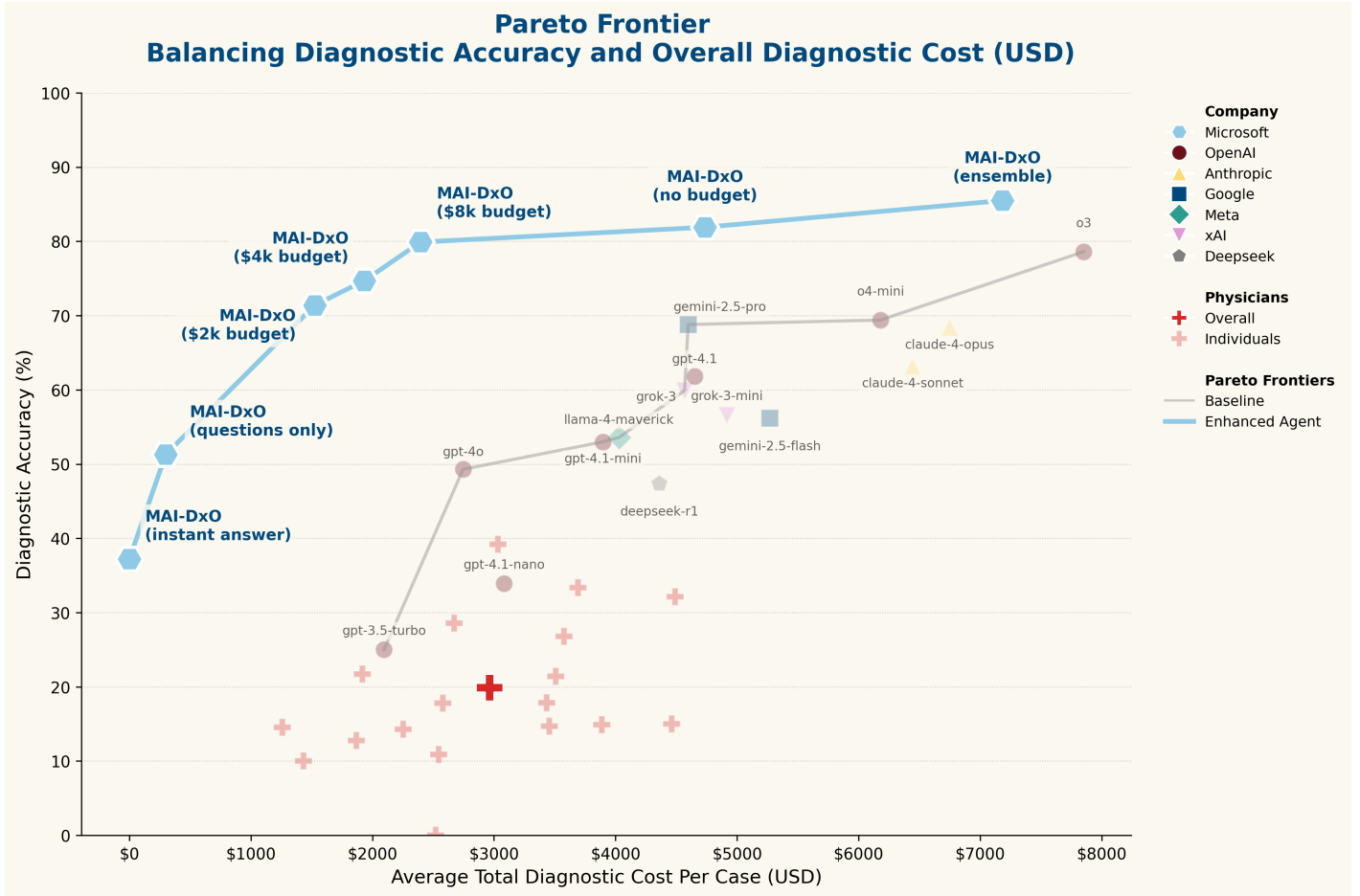


Figure 7: Pareto-frontier showing diagnostic accuracy versus average cumulative monetary cost for each agent. Off-the-shelf models were evaluated using a uniform baseline prompt (see Figure 4). MAI-DxO, built on top of the o3 model, achieves Pareto dominance over both off-the-shelf models and practicing physicians.

Off-the-shelf model performance. The Pareto frontier for off-the-shelf models ranged from modest accuracy (30-50%) with minimal testing to 70-79% accuracy with extensive testing (incurring \$4,000-7,900 in cost). While some models dominated others (e.g. Gemini-2.5-Pro had higher accuracy than Claude-4 Sonnet and Opus, at *less* cost), there was a correlation between diagnostic accuracy and cost, especially for reasoning models. Off-the-shelf o3 achieved the highest accuracy at 78.6%, but also incurred the highest cost of \$7,850 per case. This correlation indicates that information gathering remains crucial for diagnosis even for the most advanced AI systems, and supports the importance of information gathering as a key capability when evaluating models as diagnostic agents. Further, it suggests that more capable reasoning models are more skilled at assessing the complexity of cases, generating a broader set of differential diagnoses, and ordering more tests to confirm or exclude options. Meanwhile, weaker models achieved a false economy by considering fewer possible differential diagnoses, thus never ordering the tests that would confirm or exclude them.

Physician performance. Each physician completed an average of 36 cases, providing a total of 764 human diagnoses over the 56 cases in the test set. On average, physicians spent 11.8 minutes per case, requesting 6.6 questions and 7.2 tests, resulting in an average cost of \$2,963. Their average diagnostic

accuracy was 19.9%. As with language models, we observed a correlation between diagnostic accuracy and cost incurred, although the variance for physicians is higher. Even the physician performing at the highest diagnostic accuracy (41%) was outperformed by GPT-4o in both accuracy and cost. We note again that NEJM cases are selected for difficulty and instructional potential, so it is not surprising that physicians found them difficult.

MAI-Diagnostic Orchestrator performance. The blue curve in Figure 7 represents different operational modes of MAI-DxO, allowing for precise navigation of the accuracy-cost tradeoff. MAI-DxO, establishes a new Pareto frontier that outperforms off-the-shelf models *and* physicians at every cost or accuracy level. We use o3 as the model for every MAI-DxO configuration throughout in Figure 7, as model inference costs ($\approx \$0.15$ per case) are negligible compared to diagnostic test costs, making compute optimization largely irrelevant in the context of clinical diagnostic costs.

The standard MAI-DxO configuration (no budget) demonstrated particular effectiveness. When applied to o3, it achieved 81.9% accuracy (vs off-the-shelf o3 at 78.6%) while reducing average test costs to \$4,735 (from \$7,850). This simultaneous improvement in both metrics illustrates that structured reasoning mitigates the accuracy-cost trade-off present in off-the-shelf models and physicians. As expected, adding budget constraints significantly lowered costs, with progressively more loss in accuracy as the budget was reduced. Notably, one configuration achieves 79.9% accuracy at the low cost of \$2,396. On the other end of the spectrum, we were able to improve accuracy to 85.5% with ensembling techniques, while still keeping the costs at \$7,184 (from o3’s \$7,850).

We manually inspected a variety of cases where MAI-DxO (no budget) disagreed with off-the-shelf o3 with the baseline prompt. In these cases, MAI-DxO often sought disconfirming evidence and switched its diagnostic path (as a result of Dr. Hypothesis’ explicit hypothesis tracking and Dr. Challenger’s adversarial role), while off-the-shelf o3 seemed to anchor on initial impressions. Furthermore, the baseline seemed to lack a theory of information value, ordering tests that are “reasonable” given the current differential, rather than what maximally reduced diagnostic uncertainty per dollar spent. MAI-DxO’s Dr. Stewardship did not reject expensive tests outright, but forced the panel to ask whether the same information could be acquired at lower cost (in particular by asking the patient questions). As an example, one particular case featured a patient hospitalized for alcohol withdrawal who ingested hand sanitizer, leading to intoxication. Off-the-shelf o3 fixated on antibiotic toxicity, ordering expensive imaging (including a brain MRI and EEG), and finally produced an incorrect diagnosis at a high cost of \$3,431. In contrast, Dr. Hypothesis flagged the need to consider in-hospital toxin exposures given the timing in the very first round, and the panel asked about hand sanitizer ingestion before ordering tests. This direct question elicited the patient’s confession, leading to targeted confirmatory testing (toxic alcohol panel showing elevated acetone) and a correct diagnosis at a total cost of only \$795.

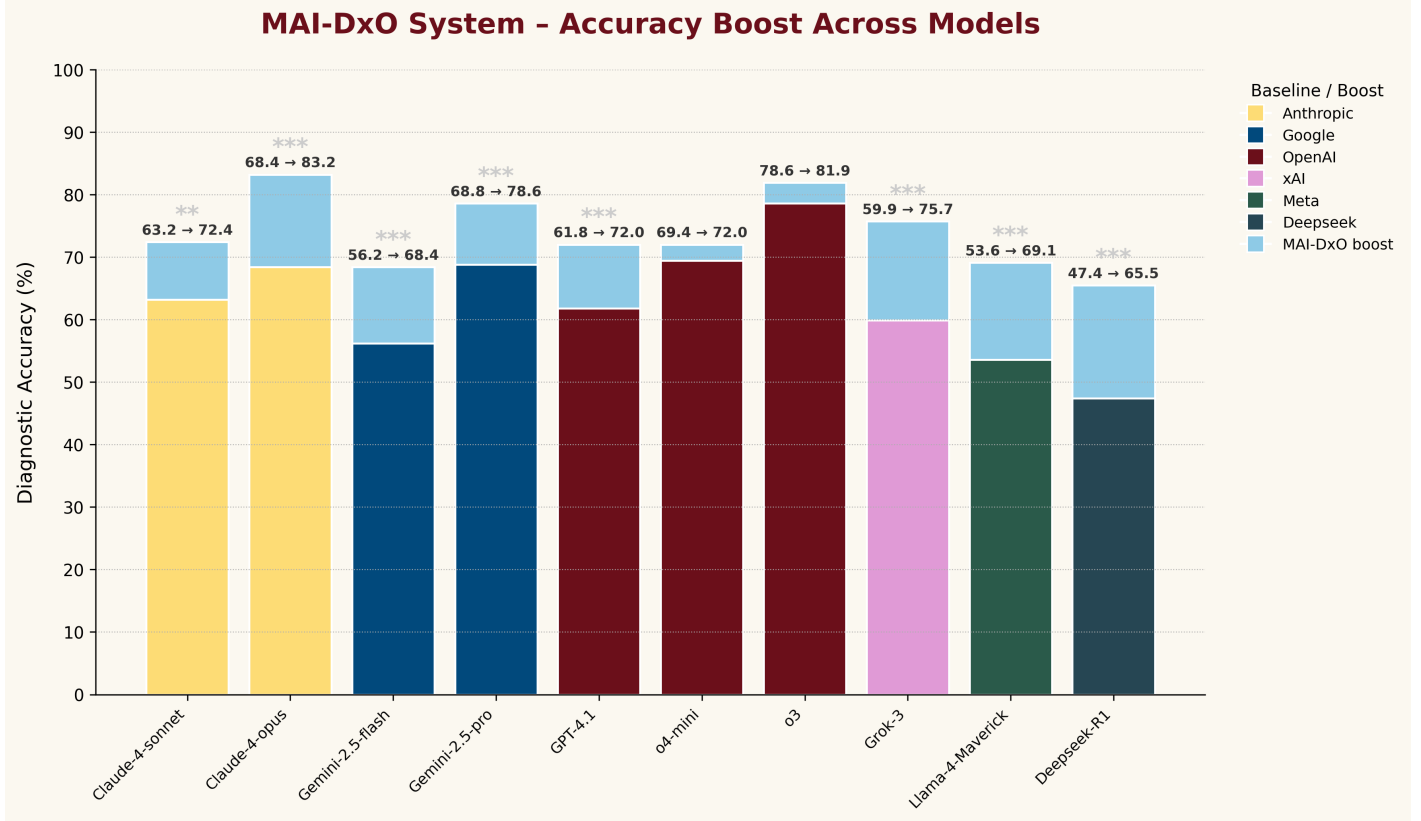


Figure 8: Accuracy improvements delivered by MAI-DxO (no budget constraints) across different large language models. Asterisks indicate statistical significance.

MAI-DxO improved all off-the-shelf models. Even though MAI-DxO was primarily developed using GPT-4.1, its structured reasoning approach proved remarkably model-agnostic. Figure 8 demonstrates that MAI-DxO consistently improves diagnostic accuracy across all sufficiently capable foundation models, with particularly pronounced gains for weaker baselines, suggesting the framework helps weaker models overcome their limitations through structured reasoning. We computed the statistical significance of all accuracy gains in Figure 8 using a one-sided paired permutation test with 10000 resamples. The gains for MAI-DxO (no budget) were statistically significant for all models ($p < 0.005$), except o3 and o4-mini which had very significant cost reductions over baseline ($p < 0.005$). Significant accuracy gains were also observed for o3 with MAI-DxO (ensemble) ($p < 0.005$).

This convergence likely reflects how MAI-DxO compensates for different types of model weaknesses. For less capable models, the explicit maintenance of a differential diagnosis and systematic test selection provide scaffolding for medical reasoning they struggle with on their own. The virtual physician panel prevents common errors like premature closure or overlooking rare diseases. For more capable models, MAI-DxO appears to impose useful discipline—ensuring comprehensive differentials, reducing anchoring bias, and encouraging cost-conscious testing. Under simple, baseline prompting we hypothesize that models may rely on sets of inductive biases introduced during post-training for preparing them for general uses. Applying MAI-DxO may help override or reorient these inductive biases.

Interestingly, we found that raw diagnostic accuracy gains were more modest for OpenAI’s reasoning models—likely because their baseline performance is already high. However, MAI-DxO was able to significantly improve their cost efficiency, as illustrated by the performance of o3 in Figure 7.

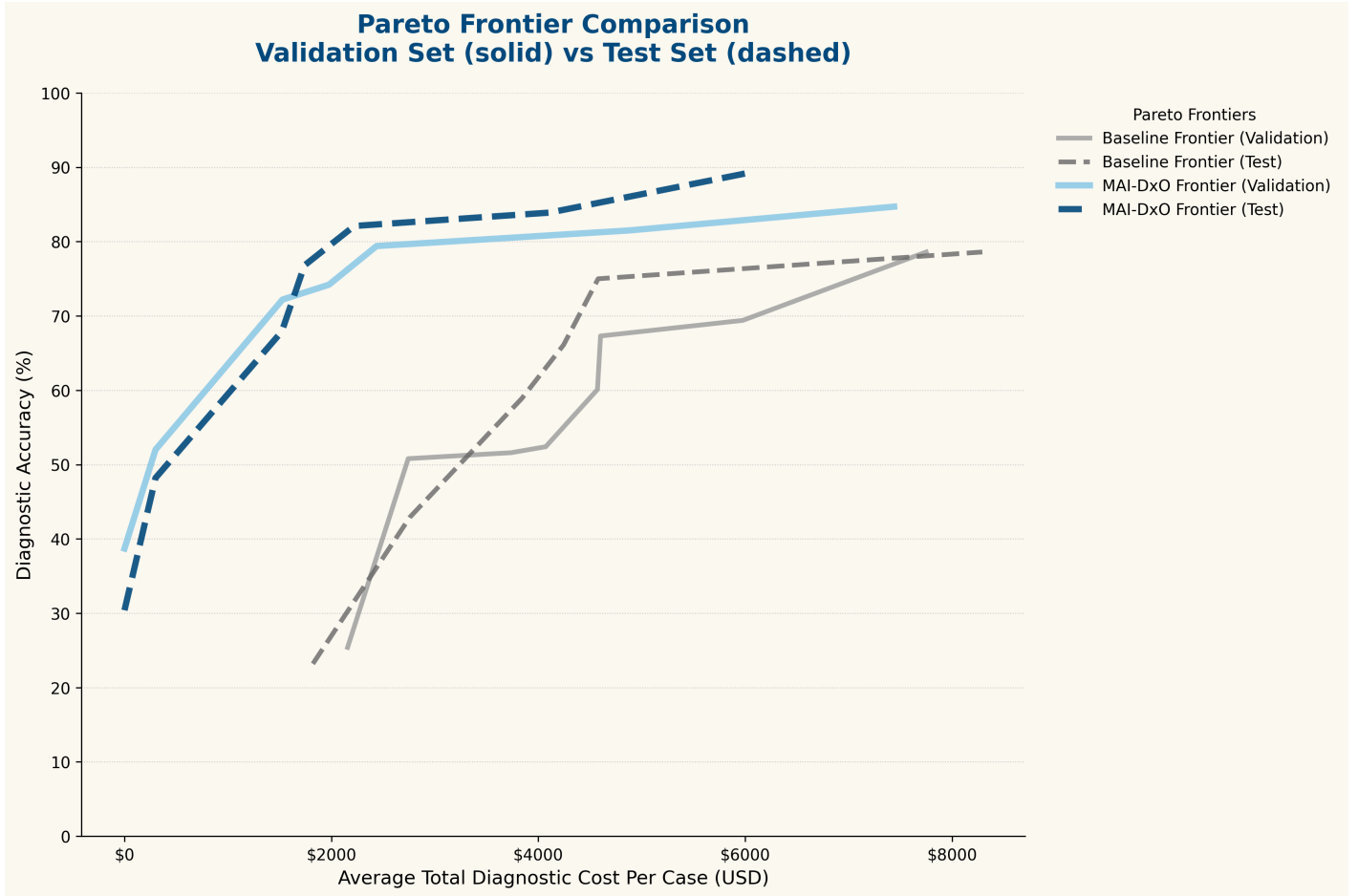


Figure 9: Pareto frontier curves of MAI-DxO and baseline prompting across validation and held-out test data. MAI-DxO continues to show significant improvement on performance on NEJM CPC cases published after model training cutoff, and thus provably outside the training corpora.

Results were robust across dataset splits. As noted earlier, the 56 most recent CPC cases, published between 2024 and 2025, were kept completely hidden from the development team as a “test set”, and no variants were run on them until methodologies were finalized. Partitioning data in this way is a common practice to measure and prevent overfitting, wherein a system fails to generalize beyond the data used during its training or validation. Strong performance on a truly held-out test set increases confidence on the system’s ability to generalize. While MAI-DxO does not update model weights (relying instead on prompting and orchestration), it is still possible for the system design choices to inadvertently overfit to cases employed during its validation. This particular train-test split also checks against potential memorization. While the NEJM cases are hidden behind a paywall, it is still possible that some off-the-shelf models were trained on them in some way. However, the majority of cases in the test set occurred *after the training data cutoff* of the models we report on.

In Figure 9, we report stratified Pareto frontier curves of model performance across the validation (248 cases) and test (56 cases) sets. The MAI-DxO system exhibited comparable absolute performance on the test set, with the relative improvements over off-the-shelf models preserved in both diagnostic accuracy and cost efficiency. These results suggest that the performance gains are robust and not driven by memorization effects.

5 Discussion

We introduce SDBench, a benchmark that transforms 304 New England Journal of Medicine CPC cases into interactive, multi-turn diagnostic challenges. Unlike static medical benchmarks that present all information upfront, SDBench more closely mirrors real-world clinical practice: diagnosticians start with minimal information and must actively decide which questions to ask, which tests to order, and when to issue a final diagnosis, with each decision incurring realistic costs. Through careful engineering, including a Gatekeeper that can synthesize plausible results for tests not described in the original cases and a clinically validated Judge to assess diagnostic accuracy, we introduce a robust evaluation environment for sequential clinical reasoning.

Within this framework, we present MAI-DxO, a system that simulates panels of different clinical personas in order to decide which questions or tests to request. MAI-DxO significantly improved diagnostic accuracy beyond strong off-the-shelf models, while simultaneously reducing cumulative test costs in SDBench, thereby establishing a new Pareto frontier between accuracy and medical cost.

5.1 Explaining superhuman performance

When doctors begin their careers, they face a key decision: should they become generalists, with broad knowledge across many medical areas, or specialists, with deep expertise in a narrow field? This division is necessary because medicine is too vast for any one person to master in full. To manage this complexity, healthcare systems rely on collaboration: generalists and specialists work together in clinics and hospitals, combining their diverse and complementary knowledge and decision-making skills to provide patients with the comprehensive and effective care that they need.

Today, frontier AI language models are challenging this traditional structure. These advanced systems show remarkable versatility, demonstrating both broad and deep medical understanding, and the polymathic ability to reason across specialties. In effect, they combine the generalist’s range with specialists’ depth. As a result, they significantly outperform individual physicians on complex diagnostic problems, such as those featured in the NEJM CPC cases. Our findings highlight this impressive capability. Expecting any single doctor to master the full range of such cases is unrealistic.

Consider, for example, a complex undiagnosed cancer case. A primary care physician’s role is to generate initial hypotheses and to refer the patient to the appropriate oncology specialist for further diagnostic workup. The specialist then oversees advanced diagnostic tests to reach a conclusive diagnosis—steps that the generalist would not typically manage.

This raises an intriguing question: When evaluating frontier AI systems, should we evaluate frontier AI systems by comparing them to individual physicians, or to entire hospital-like teams of generalists and specialists? The answer to this question will help both define and shape the future role of AI in healthcare.

5.2 Related Work

Medical problem solving has been a longstanding field of study within the medical community. In the medical AI literature, sequential diagnosis was formalized several decades ago through *normative* models grounded in Bayesian probability and decision theory (Horvitz et al., 1988). This framework enabled expert-level sequential diagnostic systems in domains such as nephrology (Gorry and Barnett, 1968), pathology (Heckerman et al., 1992; Horvitz et al., 1984), and trauma care (Horvitz and Seiver,

1997). However, widespread adoption was hindered by the practical challenges of engineering these systems, particularly bottlenecks around the need to acquire detailed, expert-curated data on probabilistic relationships between findings and diseases.

More recent work has shifted toward the application of LMs to medical challenge problems, which typically include clinical reasoning as part of a broader evaluation suite (Bedi et al., 2025a; Brin et al., 2023; Chakraborty et al., 2020; Gilson et al., 2023; Gu et al., 2021; Singhal et al., 2023). While these studies demonstrated foundational performance leaps at their time of publication, existing multiple-choice benchmarks have now become saturated, highlighting the need for more complex and realistic assessments, as well as careful end-to-end agent optimization in healthcare tasks (Bedi et al., 2025b).

To this end, there have been multiple studies, notably the *Articulate Medical Intelligence Explorer* (AMIE) line of work, which leveraged NEJM content as source material for challenging benchmarks. For diagnostic capability assessments, AMIE also leveraged NEJM-CPC cases; however, this line of work assessed models in a fixed "vignette" style setting in which the case information was summarized into a compact prompt and the models were asked to make a top-10 differential diagnosis (McDuff et al., 2025). In contrast, our key differentiation was to transform the static clinical case information into the real-world evidential reasoning challenge characterized by sequential diagnosis, which assesses models on their ability to iteratively ask for information, *starting from minimal information*, in a cost-sensitive manner and decide when a diagnosis should be made. Of note, in a parallel paper (Tu et al., 2025) AMIE was also assessed on conversational quality dimensions, such as empathy. While these represent critical dimensions of interaction with physicians and patients, we chose to frame physicians' and agents' interaction with SDBench as an interaction with an "oracle" about the patient, and so primarily focused on measures of cost and diagnostic accuracy. We note that (Li et al., 2024) also tests language models on information gathering capabilities; however, this work builds on much simpler, multiple-choice USMLE-style questions (which are a few sentences long; by contrast, NEJM CPC cases are several pages long). The authors also focus purely on information gathering via patient questions; in this work, we enable the extra dimension of ordering diagnostic tests and measuring cost efficiency.

More recently, (Brodeur et al., 2024) utilized cases from NEJM's Healer Platform, a digital platform designed to teach and assess clinical reasoning through interactive, case-based learning, in addition to NEJM-CPC cases. Notably, the cases within the NEJM Healer platform are designed to serve as educational aids for healthcare professionals in training and do not pose the same diagnostic challenge as the NEJM-CPC cases. As with McDuff et al. (2025), the presenting information from the NEJM-CPC cases were presented as fixed vignettes to a model, which generated a differential diagnosis and the next most appropriate diagnostic test. Similarly, (Schmidgall et al., 2024) leverages the NEJM Image Challenges, which are multiple choice image question and answer tasks.

5.3 Limitations

Since SDBench is built from complex, pedagogically curated NEJM CPC cases, the case distribution does not match that of a real-world deployment scenario, and indeed there are no cases where the patients are in fact healthy or have benign syndromes. Thus, we do not know whether MAI-DxO's performance gains on hard cases generalize to common, everyday clinical conditions, and could not measure false positive rates. Additionally, a practical diagnostic agent must incorporate patient-specific risk factors, and consider additional factors beyond cost, e.g. invasiveness and risk to patients, patient discomfort and wait times, expected delays before receiving results in the face of acute illness, availability of tests at current location, and constraints of authorization and reimbursement.

While our estimates of medical costs reflected test costs in the United States, in reality costs vary across geography, health systems, payers, and providers. Further, there are costs beyond the tests themselves, such as physician reporting time, device maintenance, patient travel costs for tests, etc. While recognizing that our medical cost estimates are best viewed as first-order approximations, they are consistent across all agents, and thus still help quantify relative trade-offs between accuracy and resource use.

While our report of physician performance is useful in comparing humans to AI diagnostic systems, it is also meant as a first-order approximation. Given the breadth of diagnoses represented within NEJM CPCs, we opted to recruit medical generalists only (primary care physicians and internal physicians), while in reality these might refer more complex cases to specialists. Further, we asked the participants in our study to refrain from using search engines (to prevent them finding the exact NEJM cases online), while in reality physicians are free to use such tools, including electronic medical records that often contain care guidelines, consult colleagues or textbooks, or even off-the-shelf LMs. While acknowledging these limitations, our results indicate possible accuracy gains, especially when considering clinicians working in remote and under-resourced settings, and also give us a picture of how LMs could augment medical expertise to improve health outcomes even in well-resourced settings.

5.4 Implications and Future Work

Our findings demonstrate the promise of AI methods for sequential diagnosis, including the ability to explicitly model working differential diagnoses and reason about informational value and cost of diagnostic tests. While these results do not yet establish the clinical efficacy of MAI-DxO in real-world decision support, they underscore AI’s increasing potential to address urgent challenges in healthcare delivery. Our model-agnostic system design may alleviate risks and implementation challenges for health systems aiming to adopt best-in-class language model-based diagnostic support in a rapidly evolving field. By reducing reliance on any single model, it avoids the need to “version chase” each new model release. In terms of practical application, future work should validate MAI-DxO in everyday clinical environments, where disease prevalence and presentations reflect routine practice rather than the rare, complex cases featured in the NEJM CPC corpus. An immediate goal is to identify the settings in which MAI-DxO could address unmet needs and deliver the greatest value to health outcomes and societal benefit.

We hypothesize that access to superhuman diagnostic capabilities requiring minimal health IT infrastructure could improve quality of care globally, helping to mitigate the costly impact of clinical workforce shortages and variability in care delivery Mandl (2025); Wennberg et al.. In resource-limited settings especially, cost-effective strategies may enable health systems to impact more lives per dollar spent, allowing scarce medical resources to be reserved for those with the most urgent clinical needs. More broadly, such systems might even make direct-to-consumer tools possible, such as smartphone-based triage, provided that safety, regulatory clearance, and data-privacy safeguards are demonstrably in place.

Progress toward effective clinical decision support will require the development of diagnostic corpora that mirror real-world prevalence patterns. Such benchmarks will help to surface limitations and opportunities for refinement that may be obscured by our current emphasis on especially difficult diagnostic scenarios. Second, our synthetic findings framework could support the development of large-scale interactive medical benchmarks beyond the 304 cases available here. Beyond evaluation of AI systems, the methodology we have developed could be used to enhance medical education and training, enabling students and practitioners to practice diagnostic reasoning in simulated interactive environments, potentially guided by AI-based pedagogical support. Finally, incorporating visual and other sensory modalities, such as imaging, could push diagnostic accuracy even higher while maintaining cost efficiency.

Code and data availability

We are in the process of submitting this work for external peer review and are actively working with partners to explore the potential to release SDBench as a public benchmark.

Acknowledgments

We are grateful to NEJM Group for permission to use the NEJM cases in the research reported in this paper. The research described here has benefited from work across many teams at Microsoft. We thank esteemed colleagues both inside and outside of Microsoft for sharing their insights including Bryan Bunning, Nando de Freitas, Kenneth D Mandl, Andrija Milicevic, Joseph Petro, Hoifung Poon, David Rhew, Karén Simonyan, Eric Topol, and Jim Weinstein. Gianluca Fontana and Kevin Hawkins (Prova Health) provided support on the health economics and outcomes section.

References

- Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, et al. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*, 2025.
- Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, et al. Medhelm: Holistic evaluation of large language models for medical tasks. *arXiv preprint arXiv:2505.23802*, 2025a.
- Suhana Bedi, Iddah Mlauzi, Daniel Shin, Sanmi Koyejo, and Nigam H Shah. The optimization paradox in clinical ai multi-agent systems. *arXiv preprint arXiv:2506.06574*, 2025b.
- Donald M Berwick, Thomas W Nolan, and John Whittington. The triple aim: care, health, and cost. *Health affairs*, 27(3):759–769, 2008.
- 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):16492, 2023.
- Peter G Brodeur, Thomas A Buckley, Zahir Kanjee, Ethan Goh, Evelyn Bin Ling, Priyank Jain, Stephanie Cabral, Raja-Elie Abdunour, Adrian Haimovich, Jason A Freed, et al. Superhuman performance of a large language model on the reasoning tasks of a physician. *arXiv preprint arXiv:2412.10849*, 2024.
- Stephanie Cabral, Daniel Restrepo, Zahir Kanjee, Philip Wilson, Byron Crowe, Raja-Elie Abdunour, and Adam Rodman. Clinical reasoning of a generative artificial intelligence model compared with physicians. *JAMA Internal Medicine*, 184(5):581–583, 2024.
- Souradip Chakraborty, Ekaba Bisong, Shweta Bhatt, Thomas Wagner, Riley Elliott, and Francesco Mosconi. BioMedBERT: A pre-trained biomedical language model for QA and IR. In Donia Scott, Nuria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics*, pages 669–679, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.59. URL <https://aclanthology.org/2020.coling-main.59/>.

- Aidan Gilson, Conrad W Safranek, Thomas Huang, Vimig Socrates, Ling Chi, Richard Andrew Taylor, David Chartash, et al. How does chatgpt perform on the united states medical licensing examination (usmle)? the implications of large language models for medical education and knowledge assessment. *JMIR medical education*, 9(1):e45312, 2023.
- Ethan Goh, Robert Gallo, Jason Hom, Eric Strong, Yingjie Weng, Hannah Kerman, Joséphine A Cool, Zahir Kanjee, Andrew S Parsons, Neera Ahuja, et al. Large language model influence on diagnostic reasoning: a randomized clinical trial. *JAMA Network Open*, 7(10):e2440969–e2440969, 2024.
- G Anthony Gorry and G Octo Barnett. Experience with a model of sequential diagnosis. *Computers and Biomedical Research*, 1(5):490–507, 1968.
- Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.
- David E Heckerman, Eric J Horvitz, and Bharat N Nathwani. Toward normative expert systems: Part I The Pathfinder project. *Methods of information in medicine*, 31(02):90–105, 1992.
- Eric Horvitz and Adam Seiver. Time-critical action: Representations and application. In *Proceedings of the Thirteenth Conference on Uncertainty in Artificial Intelligence*, pages 250–257, 1997.
- Eric J Horvitz, David E Heckerman, Bharat N Nathwani, and Lawrence M Fagan. Diagnostic strategies in the hypothesis-directed pathfinder system. In *Proceedings of the First IEEE Conference on Applications of Artificial Intelligence*, 1984.
- Eric J Horvitz, John S Breese, and Max Henrion. Decision theory in expert systems and artificial intelligence. *International journal of approximate reasoning*, 2(3):247–302, 1988.
- Martín Hunter, Ignacio Lopez Saubidet, Tomás Amerio, and Maria V. Leone. Case 15-2025: A 52-year-old man with fever, nausea, and respiratory failure. *New England Journal of Medicine*, 392(20):2049–2057, 2025. doi: 10.1056/NEJMcp2412526. URL <https://www.nejm.org/doi/full/10.1056/NEJMcp2412526>.
- Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei W Koh, and Yulia Tsvetkov. Mediq: Question-asking llms and a benchmark for reliable interactive clinical reasoning. *Advances in Neural Information Processing Systems*, 37:28858–28888, 2024.
- Kenneth D Mandl. How ai could reshape health care—rise in direct-to-consumer models. *JAMA*, 333(19):1667–1669, 2025.
- Daniel McDuff, Mike Schaekermann, Tao Tu, Anil Palepu, Amy Wang, Jake Garrison, Karan Singhal, Yash Sharma, Shekoofeh Azizi, Kavita Kulkarni, et al. Towards accurate differential diagnosis with large language models. *Nature*, pages 1–7, 2025.
- Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. Capabilities of GPT-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*, 2023a.
- Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. *arXiv preprint arXiv:2311.16452*, 2023b.

- Harsha Nori, Naoto Usuyama, Nicholas King, Scott Mayer McKinney, Xavier Fernandes, Sheng Zhang, and Eric Horvitz. From Medprompt to o1: Exploration of run-time strategies for medical challenge problems and beyond. *arXiv preprint arXiv:2411.03590*, 2024.
- 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. *arXiv preprint arXiv:2405.07960*, 2024.
- Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.
- Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, et al. Towards conversational diagnostic artificial intelligence. *Nature*, pages 1–9, 2025.
- John E Wennberg, S Brownle, Elliot S Fisher, Jonathan S Skinner, and James N Weinstein. Improving quality and curbing health care spending: opportunities for the congress and the obama administration. dartmouth atlas white paper. 2008.

A Batched vs Single Testing

The main results for MAI-DxO and baseline agents were all performed using batch testing mode, where the agents were allowed to ask many questions or request many tests within one turn. MAI-DxO agents were instructed to limit this to 5 questions and 3 tests. Clinician review of our benchmark setup suggested that batching questions and tests reflected more realistic testing strategies, to alleviate patient discomfort, reduce test result wait times and for practical reasons concerning laboratory test processing. However, for usability reasons the human clinicians used the UI-version of the benchmark in a single-turn fashion. We ran an ablation on the test set only for MAI-DxO (o3) to compare single vs batched mode. We found that both single and batched mode agents had identical accuracy (83.9%), but the single-mode agent actually had a lower cost (\$3,991 vs \$5,084). However, to preserve the aforementioned realism, we report MAI-DxO results in the batch setting.

B Visualizing Model and Clinician Performance

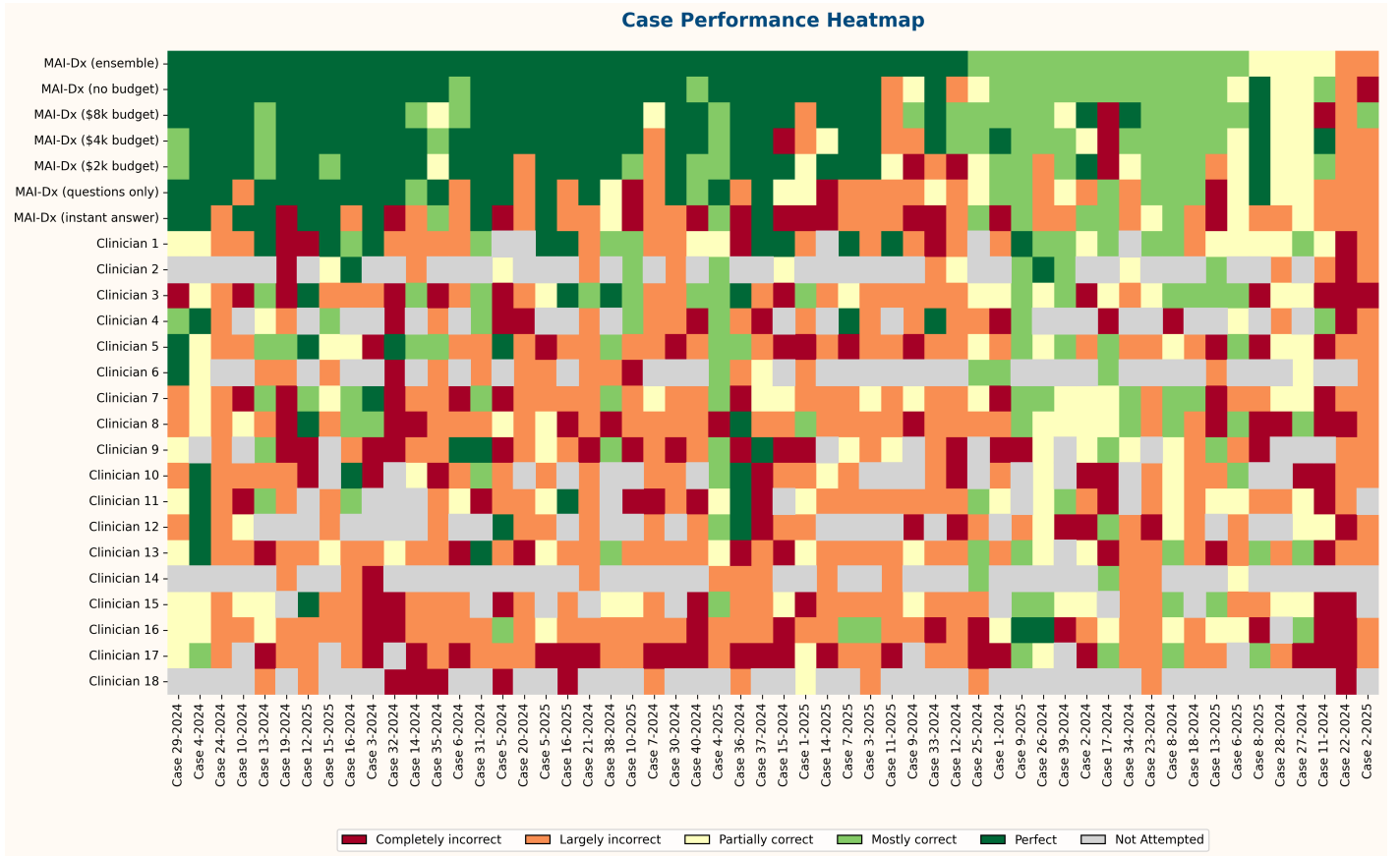


Figure 10: Case level scores of MAI-DxO variants and clinicians across the 56 case test set. All clinicians who assessed at least 10 cases are shown. Model variants and clinicians are ordered by their diagnostic accuracy and cases are ordered by their perceived difficulty according to MAI-DxO (ensemble).