

AIM²: Adaptive Intelligent Medical Multi-Agents

EECS6895 Advanced Big Data and AI

Tianshuai Gao*

tg2935@columbia.edu

Columbia University, Department of Statistics
New York, NY, USA

Abstract

Large Language models (LLMs) have shown potential in medical applications. However, the best way to leverage Large Language models (LLMs) in complex clinical tasks remains an open question. I introduce a multi-agent framework named by Addaptive Intelligent Medical Multi-Agents (AIM²)^[1] that emulates dynamic clinical decision-making through structured, context-aware collaboration among large language models (LLMs). AIM² first interprets the complexity of the task and the clinical modality, then automatically assigns agents either alone or in a team with specialized roles and scopes of reasoning. These agents engage in deliberation when appropriate, simulating multidisciplinary team (MDT) workflows common in hospitals. We evaluate AIM² and baseline methods using state-of-the-art LLMs across a suite of benchmarks of medical knowledge and medical diagnosis in real-world settings. The results illustrate the capacity of AIM² to adaptively balance efficiency and depth of reasoning while maintaining transparent, role-based interactions. This framework bridges the gap between powerful foundation models and practical, adaptive medical reasoning systems.

My code can be found at

<https://github.com/tianshuai-gao/E6895-AIM>

My video can be found at

<https://www.youtube.com/watch?v=HJUf1HcermQ>.

Keywords

Adaptive multi-agent collaboration, Clinical decision support, Large language models, Multi-modal medical reasoning, Complexity-aware orchestration

1 Introduction

Medical decision-making (MDM) is a intricate collaborative process in which clinicians interpret complex multi-modal data—such

*This work is an intended work from EECS6895 Advanced Big Data and AI at Columbia University

^[1]The superscript “2” denotes two pillars: *multi-modal understanding* and *multi-agent coordination*.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
EECS6895 '25, New York, NY, USA

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnnn.nnnnnnn>

as image, electronic health records (EHR), physiological signals, and genetic information—while rapidly integrating new medical research into practice to reach specific and precise conclusions [65]. For example, when a case is complex, a primary care physician (PCP) may refer the patient to a specialist; similarly, patients arriving at the emergency department or urgent care are often triaged and then directed to a specialist for further evaluation [3, 31]. Recently, it is known that Large Language Models (LLMs) are able to synthesize large volumes of all kinds of literature and information, as well as enabling probabilistic, causal, and mathematical reasoning.

Recent research in multi-agent LLM systems has shown success in other domains, but their application to Medical decision-making (MDM) remains limited and underexplored. However, most existing approaches rely on static prompting and single-agent reasoning [55], failing to adapt to task-specific needs or simulate tiered collaboration structures used in real clinical workflows. In emergency care, triage stratifies patients by the severity and complexity of their conditions [7, 16, 59]. Low-complexity cases include pathognomonic, uncomplicated acute presentations or stable chronic issues that a PCP can manage [56]. In contrast, injuries involving multiple organ systems, multimorbidity with treatment side effects, or superimposed diseases typically require repeated multidisciplinary team (MDT) discussions or sequential inter-specialty consultations (ICT) [17, 37] and are classified as high-complexity cases [1].

To bridge this gap, we present Addaptive Intelligent Medical Multi-Agents (AIM²), a multi-agent collaboration framework designed to emulate real-world MDM processes using task-aware agent orchestration and role-based reasoning. Our system design is inspired by the way clinicians dynamically adjust their collaboration depth depending on the task at hand. Unlike static group prompting or fixed-agent pipelines, AIM² adapts reasoning granularity and communication pathways based on contextual cues and complexity signals. This allows it to support both rapid solo decisions in straightforward cases and collaborative exploration in ambiguous or high-stakes settings.

AIM² operate in four stages: (1) clinical-complexity assessment; (2) complexity-aware agent recruitment; (3) structured analysis and synthesis; and (4) decision generation.

My contributions are threefold:

- (1) I present AIM², an adaptive framework that mirrors real-world MDM via dynamic, complexity-driven collaboration among LLM agents.
- (2) On a suite of medical benchmarks, our reproduction attains accuracy comparable to strong single-agents and multi-agent baselines, with **modest improvements on several benchmarks**; we also characterize the performance–efficiency

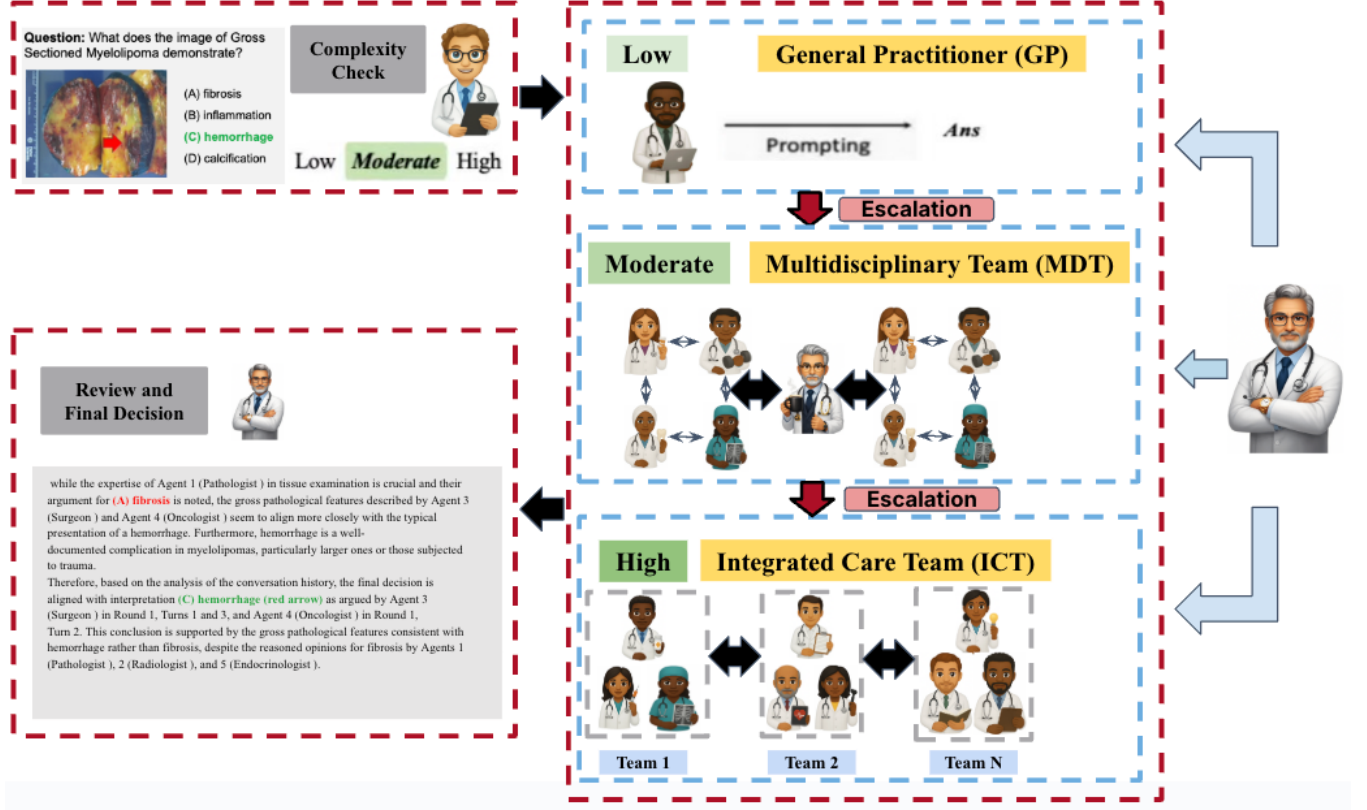


Figure 1: Decision pathway in AIM². Given a question and its context, the difficulty agent predicts whether the case is low, moderate or high complexity. Low cases are routed to a single general practitioner (GP) with simple prompting, moderate cases to a multidisciplinary team (MDT) and high cases to an integrated care team (ICT). If discussion at a lower level remains uncertain, the moderator escalates the case to the next tier. Finally, the moderator reviews all notes and produces the final decision and explanation.

trade-off by varying the **number of agents (measured via API calls)**.

- (3) Through hyperparameter sweeps (e.g., temperature) and ablations, we observe greater robustness than single-agent or fixed-group setups and show that the orchestration module reliably selects an appropriate collaboration depth for each MDM instance.

2 Related Works

2.1 Language Models in Medical Decision-Making

Large language models (LLMs) have been applied across a broad spectrum of clinical tasks, spanning fact-seeking question answering and exam-style assessments, biomedical knowledge discovery, clinical risk stratification, and diagnostic support [24, 47, 64]. Beyond discriminative settings, recent work evaluates generative capabilities, including drafting and editing clinical reports, describing medical images, producing differential lists with rationales, conducting clinician-patient dialogues, and summarizing psychiatric [49, 50, 52].

Two complementary routes have emerged for improving medical LLMs. Training-time approaches adapt models with domain data via pretraining or finetuning, instruction tuning on clinical corpora, multimodal fusion for image-text inputs, and safety alignment [19, 32, 33, 41]. In parallel, inference-time techniques enhance reasoning without additional training through prompt design (e.g., chain-of-thought and self-consistency), retrieval-augmented generation (RAG) over guidelines or literature, tool use (e.g., drug databases and calculators), ensembles, and constrained decoding to improve factuality and completeness [21, 26, 53, 55, 61]. Notably, strong general-purpose LLMs, when paired with careful prompting or RAG, can match or surpass dedicated medical models on several standardized evaluations [34, 36].

Despite these advances, most pipelines still operate a single-agent paradigm with fixed prompting, which only partially reflects the dynamic, tiered, and collaborative nature of real-world medical decision-making (MDM). This observation motivates frameworks that organize multiple specialized agents to reason jointly under clinical constraints [25, 62].

Table 1: Comparison between our framework and previous methods (Solo and Group). Among these works, AIM² is the only one to perform all key dimensions of decision-making.

Method	AIM ²	Single	Voting [53]	Debate [10]	MedAgents [46]	ReConcile [6]
Multiple Roles	✓	✗	✓	✓	✓	✓
Early Stopping	✓	✗	✓	✓	✓	✗
Refinement	✓	✗	✗	✓	✓	✗
Complexity Check	✓	✗	✗	✗	✗	✗
Multi-party Chat	✓	✗	✗	✓	✗	✗
Conversation Pattern	Flexible	Static	Static	Static	Static	Static

2.2 Multi-Agent Collaboration

A growing line of research explores collaboration among multiple LLM agents to exceed the capability of a single model [51]. A common pattern is role-based cooperation, where agents assume complementary responsibilities (e.g., assistant, specialist, coordinator), decompose a task into subproblems, and iteratively refine intermediate outputs [28, 58]. Another family is debate/consensus methods: agents first reason independently, then critique or vote to converge on a final answer, which has been shown to improve factuality, mathematical reasoning, and reliability on complex problems [6, 10, 53]. Variants include group deliberation, multi-disciplinary collaboration, and negotiation-style protocols [6, 28, 46].

However, many existing systems predefine the number of agents, roles, and interaction rounds, yielding static orchestration. When task difficulty, uncertainty, or modality mix varies, static designs can be inefficient—incurring unnecessary computation—or suboptimal—failing to recruit sufficient expertise [30, 39]. Moreover, the additional coordination cost of multi-agent methods must be justified by measurable gains under realistic budgets (e.g., API calls or latency) [10, 58].

In response, recent frameworks introduce complexity-aware or adaptive orchestration that selects when and how to collaborate, and at what depth, given contextual signals [21, 29, 30]. Our work follows this direction: we incorporate an explicit complexity check and flexible conversation patterns to recruit appropriate agents (e.g., intra-specialty teams vs. MDT-style assemblies), aligning the interaction protocol with clinical MDM. A qualitative comparison of interaction dimensions (roles, early stopping, refinement, complexity assessment, multi-party chat, and conversation pattern) across prior methods is provided in Table 1.

3 AIM²: Adaptive Intelligent Medical Multi-Agents

The design of AIM² follows a staged pipeline. Figure 1 illustrates the process. The system first estimates the complexity of the incoming query. It then chooses a suitable collaboration level, recruits experts, lets them discuss in several rounds, and finally returns an answer that can be read by clinicians and patients.

3.1 Agent Roles

Moderator. The moderator is the first contact for each query. It receives the case description and attached tests or images. The moderator calls the difficulty agent, decides which collaboration

level should be used, tracks the progress of the case and makes the final decision to accept an answer or to escalate.

Difficulty Agent. The difficulty agent is a small classifier style agent. It reads the question and a short context summary and predicts whether the case is low, moderate or high complexity. It also outputs a short explanation and a confidence score. The prediction is used as the initial complexity label.

Recruiter. The recruiter forms the working team once the level is known. For low complexity it selects a single general practitioner (GP) agent. For moderate complexity it builds a multidisciplinary team (MDT) inside one department or closely related specialties. For high complexity it builds an integrated care team (ICT) that spans several departments. The recruiter draws agents from a library of common roles and reuses them across similar cases.

Generalist and Specialist Agents. Generalist agents manage routine problems that usually stay in primary care. Specialist agents cover domains such as cardiology, oncology, radiology and psychiatry. In MDTs specialists mostly share one organ system or modality. In ICTs they represent distinct disciplines. Each agent is instructed to explain its intermediate reasoning in natural language.

Synthesis Agent. The synthesis agent focuses on writing the final answer. It receives the notes, reports and votes from the discussion stage together with simple uncertainty scores. It produces a clinician facing explanation and, when needed, a patient facing summary.

3.2 Medical Complexity Check

The first step in AIM² is a medical complexity check. The moderator packages key information such as age, main symptoms, vital signs and imaging summaries and sends it to the difficulty agent. The difficulty agent returns one of three labels.

Low complexity. These are well defined questions that fit a standard clinical pathway. Typical examples are uncomplicated acute illnesses or stable chronic conditions where a single GP can manage the problem with little coordination [57].

Moderate complexity. These cases involve several interacting factors or mild uncertainty. They often require input from a small team inside one specialty or closely related specialties [44, 45]. Typical examples are patients with multiple comorbidities or partial conflicts between tests.

High complexity. These cases require coordinated input from many specialties or from several stages of care. Examples include complex cancer care, major trauma or rare diseases with limited guidelines [22, 38, 43].

Table 2: Accuracy (%) on selected Medical benchmarks with Solo/Group/Adaptive settings. Bold represents the best and underline the second best performance for each benchmark.

Category	Method	MedQA [23]	PubMedQA [24]	Path-VQA [20]	PMC-VQA [63]	MIMIC-CXR [2]
6*Single-agent	Zero-shot	75.0 ± 1.3	61.5 ± 2.2	57.9 ± 1.6	49.0 ± 3.7	37.9 ± 8.4
	Few-shot	72.9 ± 1.4	63.1 ± 1.7	57.5 ± 3.5	52.2 ± 1.0	47.1 ± 8.6
	+CoT [55]	82.5 ± 4.2	57.6 ± 1.9	61.3 ± 5.1	51.5 ± 1.3	48.6 ± 5.4
	+CoT-SC [51]	83.9 ± 3.4	<u>58.7 ± 5.0</u>	61.0 ± 3.5	50.5 ± 3.0	49.2 ± 8.2
	ER [39]	<u>81.9 ± 2.1</u>	56.0 ± 7.0	61.4 ± 4.1	52.7 ± 2.9	48.5 ± 4.1
	Medprompt [34]	82.4 ± 5.4	51.8 ± 4.6	<u>59.2 ± 1.5</u>	<u>53.4 ± 3.7</u>	44.5 ± 7.2
5*Single-model Multi-agent	Majority Voting	80.6 ± 2.9	72.2 ± 6.9	56.9 ± 1.7	36.8 ± 6.7	50.8 ± 7.4
	Weighted Voting	78.8 ± 1.8	72.2 ± 8.9	62.1 ± 1.9	25.4 ± 5.7	57.8 ± 2.1
	Borda Count	70.3 ± 4.8	66.9 ± 4.3	55.6 ± 2.4	29.7 ± 3.3	54.5 ± 4.7
	MedAgents [28]	69.7 ± 4.1	73.7 ± 4.1	63.1 ± 3.2	29.5 ± 5.1	51.6 ± 4.8
	Meta-Prompting [29]	<u>80.6 ± 1.2</u>	73.3 ± 4.2	55.3 ± 4.9	42.6 ± 4.2	-
3*Multi-model Multi-agent	Reconcile [6]	81.3 ± 5.9	79.7 ± 1.8	55.9 ± 3.7	31.4 ± 5.1	-
	AutoGen [58]	60.6 ± 5.1	77.3 ± 4.2	43.0 ± 3.9	37.3 ± 4.1	43.3 ± 4.0
	DyLAN [30]	64.2 ± 2.3	73.6 ± 4.9	41.3 ± 4.1	34.0 ± 3.5	38.7 ± 1.2
Adaptive Intelligent Medical Multi-Agents	AIM ²	82.7 ± 4.0	74.0 ± 1.0	64.0 ± 3.9	55.1 ± 4.5	54.0 ± 9.1

The complexity label controls which team will be recruited and how strong the discussion protocol should be [15, 18].

3.3 Expert Recruitment

Given a query and its current complexity label the recruiter assembles an appropriate team.

For low complexity it assigns a single GP agent. The GP receives the question and a short context summary and reasons through the case with a simple prompting template.

For moderate complexity it constructs an MDT, such as several physicians within the same department and an imaging consultant. Each member has a brief role description, for example chair, diagnostic expert or treatment planner.

For high complexity it constructs an ICT that includes agents from several departments [11, 13, 45]. A typical ICT contains assessment, diagnostic and management roles. One member in each group acts as coordinator.

The recruiter also adds a challenger style agent that is instructed to criticize the current plan and to search for failure modes. This challenger is present for moderate and high cases.

3.4 Medical Collaboration and Refinement

The collaboration protocol in AIM² is driven by the complexity label. Each query is assigned to one of three levels and then follows a corresponding refinement pathway. These pathways are implemented in Algorithm 1.

Low – Straightforward cases (Lines 3–8 of Algorithm 1).

For queries labelled as low complexity the framework follows a simple GP centred route. The recruiter assigns a single general practitioner agent. The GP receives the question and a compact context summary and produces a provisional answer with a short hidden chain of thought. A light self check evaluates internal consistency and basic safety. If this check is passed the case is marked as resolved and sent directly to the decision stage. If the check fails or the GP expresses low confidence the moderator upgrades the case to moderate complexity and triggers MDT recruitment instead of starting a long discussion at the low level.

Moderate – Intermediate complexity cases (Lines 9–17 of Algorithm 1). Moderate complexity queries are handled by a multidisciplinary team [4, 14, 27]. The recruiter forms an MDT from several specialists within one department or closely related fields together with a challenger agent. The MDT enters the multi round discussion protocol with a modest cap on the number of rounds. After each round the moderator checks quality and consensus. If accuracy and safety look sufficient the conversation stops and the case moves to the decision stage with an MDT based answer. If disagreement remains high or the judge agent still gives a low quality score the moderator escalates the case to high complexity and an ICT is formed.

High – Complex care cases (Lines 18–21 of Algorithm 1). High complexity queries are assigned to an integrated care team [12, 22]. The recruiter divides ICT members into assessment, diagnostic and management roles together with a challenger. These agents follow the same multi round discussion protocol but with a higher cap on the number of rounds. In early rounds assessment agents summarize the case and identify red flags. Later rounds focus on differential diagnosis and on treatment planning. High level cases do not escalate further. If the quality and consensus checks indicate that the answer is still uncertain when the round limit is reached the moderator marks the case as high risk so that the decision stage can surface this uncertainty in the final report.

This design lets AIM² keep straightforward cases at the GP level, use MDT collaboration for intermediate problems and reserve full ICT style refinement for the most complex queries.

3.5 Multi-round Discussion and Escalation

After the team is formed AIM² enters a multi round discussion stage. The protocol is implemented in Algorithm 1. Each query can go through several cycles of discussion and possible escalation.

Round structure. At round r the moderator sends the current case state to all team members. Each expert writes a short note that contains a proposed answer, key evidence and explicit uncertainties. The challenger inspects these notes and lists potential errors, missing tests and unsafe suggestions. The moderator then builds a

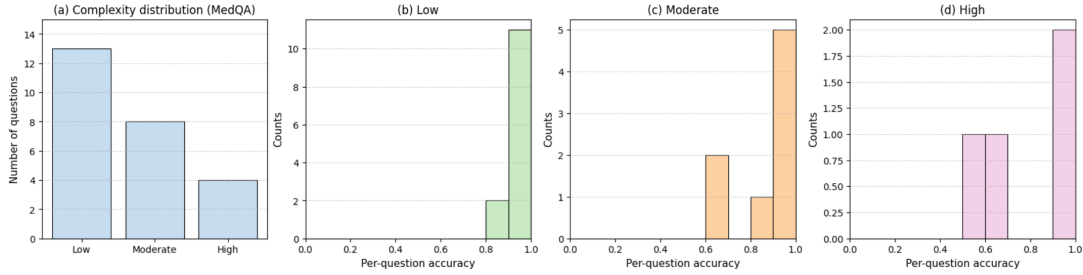


Figure 2: Analysis of complexity handling on MedQA [23] with AIM². (a) Distribution of questions assigned to low, moderate and high complexity by the difficulty agent. (b)–(d) Distributions of per-question accuracy under the low, moderate and high collaboration settings, respectively. Each question is solved multiple times and the histograms count how many questions achieve a given average accuracy at each complexity level.

brief summary of agreements and disagreements and shares it with the team.

Quality and consensus check. After each round the moderator runs a simple quality check. It asks a separate judge agent to score the current main answer along three axes: correctness, completeness and safety, each on a discrete scale. It also measures consensus by counting how many experts support the main answer. If the quality score is high and disagreements are small the discussion stops and the case moves to the decision stage.

Escalation rule. If the team cannot reach good quality within a small number of rounds the moderator considers escalation. For low complexity cases that fail the check the moderator upgrades the case to moderate complexity and calls the recruiter again to form an MDT. For moderate cases that still look uncertain the moderator upgrades them to high complexity and an ICT is formed [8, 60]. High complexity is the top level. High level cases do not escalate further but can run up to a larger maximum number of rounds with the ICT.

Stopping conditions. Each level has its own limit on the number of rounds. Low cases usually allow one or two rounds. Moderate cases allow a few more. High cases allow the largest number of rounds. If the limit is reached without strong agreement the moderator still moves to the decision stage but marks the case as uncertain so that the final answer can reflect this.

3.6 Decision Making

In the final stage AIM² turns intermediate outputs into a single answer. The synthesis agent uses the complexity label, the last round notes and the uncertainty flags to choose how to combine information.

Low complexity cases. Most low cases are solved by a single GP within one round. The synthesis agent lightly edits the GP answer for clarity. If the case was escalated from low to moderate the final answer is based on the MDT discussion instead. This keeps the cost close to Solo prompting while providing a safety net for harder questions.

Moderate complexity cases. For moderate cases the synthesis agent aggregates the MDT notes. It gives more weight to views that are supported by several team members and by clear evidence. Disagreements are briefly mentioned when clinically relevant. This acts like a structured form of voting with explanations. On MedQA

[23], PubMedQA [24] and Path-VQA [20] this procedure improves over majority voting, weighted voting and MedAgents, as reported in Table 2.

High complexity cases. For high cases the synthesis agent joins the reports from different ICT roles. It first checks that no recommendation violates basic safety rules. It then writes a narrative that explains the main plan and one or two reasonable alternatives and it highlights remaining uncertainty. In our experiments this report based synthesis is competitive with or better than fixed multi model systems such as ReConcile [6], AutoGen [58] and DyLAN [30] on difficult benchmarks like MIMIC-CXR [2], while keeping a clear medical decision structure [6, 40, 58].

Across all levels the output consists of a concise answer and an optional extended explanation. When needed AIM² generates both a clinician facing version with technical detail and a patient facing version that uses simpler language.

4 Experiments and Results

In this section we evaluate AIM² on five medical benchmarks in Solo, Group and Adaptive settings. We study overall accuracy, the effect of complexity selection, the impact of the number of agents and the robustness of the framework.

4.1 Setup

Datasets. We consider five datasets that cover both text based and image based clinical questions.

MedQA [23] and PubMedQA [24] evaluate medical question answering from text. MedQA [23] contains multiple choice questions from medical board style examinations. PubMedQA [24] is built from PubMed abstracts and asks whether the conclusion of a paper supports a given statement.

Path-VQA [20] and PMC-VQA [63] are visual question answering benchmarks. Path-VQA [20] uses pathology images with short clinical questions. PMC-VQA [63] uses figures from biomedical articles and requires models to combine visual content with captions and domain knowledge.

MIMIC-CXR [2] focuses on chest radiography. Each example pairs a chest X-ray and the corresponding report with a question about findings or impressions.

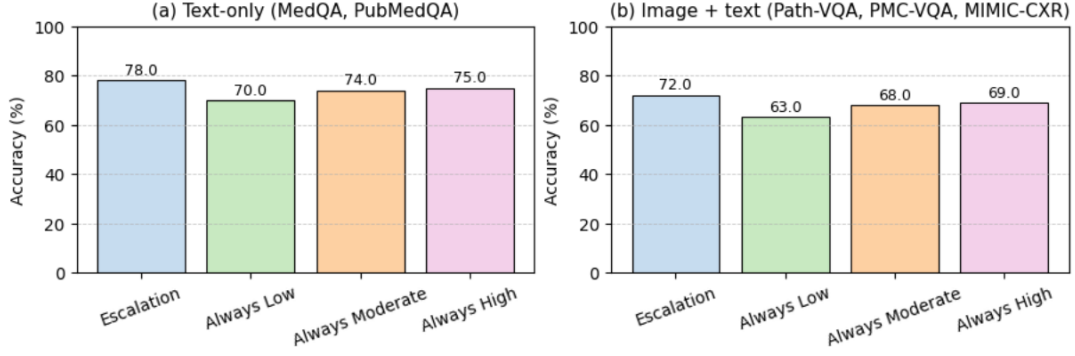


Figure 3: Impact of complexity selection strategies in AIM². Escalation denotes the proposed strategy that dynamically upgrades queries from low to moderate or high collaboration when the discussion remains uncertain. Always Low, Always Moderate and Always High treat all queries as low, moderate or high complexity respectively, without escalation. Results are reported on text-only benchmarks (MedQA [23] and PubMedQA [24]) and image+text benchmarks (Path-VQA [20], PMC-VQA [63] and MIMIC-CXR [2]).

Compared settings. All methods use the same backbone language model and the same pre processing for each dataset.

Solo A single agent answers the query. We include zero shot prompting, few shot prompting, chain of thought prompting, self consistent chain of thought, ensemble refinement and Medprompt as baselines [5, 26, 35, 42, 54, 55].

Group Several agents collaborate with fixed decision rules but without complexity awareness. We test majority voting, weighted voting, Borda count, MedAgents and meta prompting. These methods use the same prompts as Solo and only differ in how they aggregate answers.

Adaptive AIM² first predicts the complexity level of the query and then builds a pathway with one agent or with a team. Low complexity cases follow the single agent route. Moderate cases involve a small multidisciplinary team with a bounded number of discussion rounds. High complexity cases trigger the full multi stage collaboration described in the previous section.

4.2 Overall accuracy

Table 2 shows accuracy on MedQA [23], PubMedQA [24], Path-VQA [20], PMC-VQA [63] and MIMIC-CXR [2]. AIM² reaches the best or second best performance on all benchmarks. On MedQA [23] and PubMedQA [24] it slightly improves over the strongest Solo baselines such as self consistent chain of thought and Medprompt. On the visual datasets AIM² clearly improves over Solo methods and over most Group baselines, although specialised multi model systems such as ReConcile [6] sometimes remain stronger. On MIMIC-CXR [2] the adaptive framework approaches or surpasses these group methods while keeping a single backbone model.

Figure 2 illustrates results on MedQA [23] in more detail. Panel (a) shows how AIM² distributes queries across low, moderate and high complexity. Panels (b) to (d) plot the accuracy of the framework at each level. Low complexity questions are often solved with near perfect accuracy. Moderate and high levels are more challenging, but the adaptive design keeps performance stable across all three regions.

5 Ablation Studies

5.1 Effect of complexity selection

To study the value of the complexity check we compare four variants. The full AIM² uses an escalation strategy that routes queries to low, moderate or high collaboration and upgrades them when the discussion remains uncertain. Always Low sends every query to the GP pathway, Always Moderate always uses an MDT, and Always High always activates the full ICT protocol.

Figure 3 reports accuracy for these variants on text-only benchmarks (MedQA [23] and PubMedQA [24]) and on image+text benchmarks (Path-VQA [20], PMC-VQA [63] and MIMIC-CXR [2]). On text-only datasets, escalation improves several points over all three static baselines. On image+text datasets the gap is larger: Always Low underfits because many cases require multiple specialists, while Always Moderate and Always High waste computation on easy queries and can even hurt accuracy when large teams introduce noisy opinions. The escalation policy balances these extremes, achieving the best trade off between performance and cost.

5.2 Impact of the number of agents

We next vary the maximum team size for moderate and high complexity cases. Figure 4(a) shows the average accuracy as a function of the maximum number of agents that AIM² is allowed to recruit. Small teams with two or three specialists already capture most of the gains over the solo GP baseline. Beyond this point the escalation strategy quickly saturates and remains close to the performance of a system that always uses a full ICT, while fixed MDT and ICT settings continue to require larger teams to reach similar accuracy.

Figure 4(b) plots the corresponding number of model calls per query. Fixed ICT and MDT baselines incur a high and almost constant call count once the team size is large, whereas escalation grows much more slowly because many low complexity queries are resolved by a single GP and moderate cases often reach consensus early. As a result AIM² attains near-ICT accuracy with a budget that is much closer to the solo setting, indicating that selectively

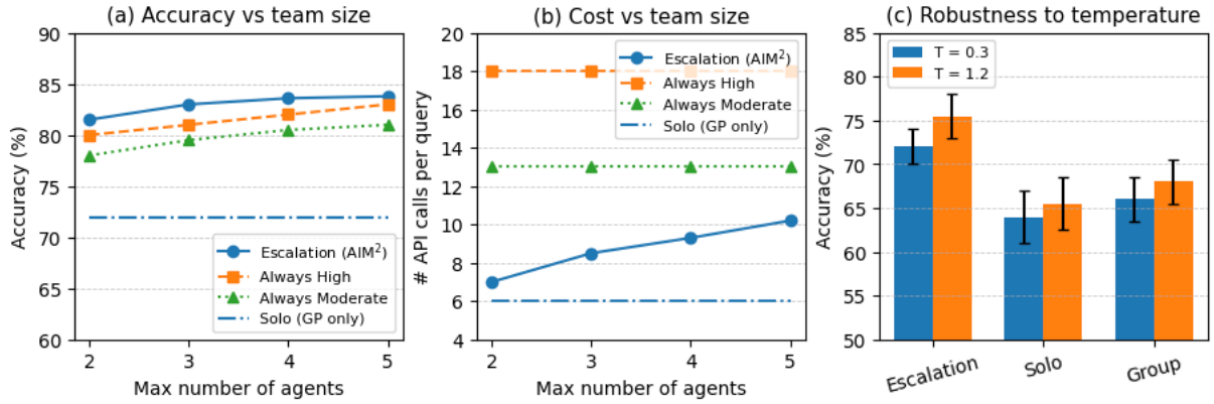


Figure 4: Impact of the number of agents and temperature in AIM². (a) Accuracy as a function of the maximum number of agents. The escalation strategy reaches high accuracy with small teams, approaching the performance of always using an ICT while outperforming single-agent and always-MDT settings. (b) Average number of API calls per query under the same configurations. Escalation substantially reduces cost compared to always-high and always-moderate collaboration. (c) Robustness to decoding temperature: escalation maintains strong performance and can exploit higher temperatures better than solo and fixed group baselines.

recruiting a small team with escalation is more efficient than simply adding more agents for every query.

5.3 Robustness and consensus behaviour

Finally we examine robustness to sampling temperature and how the multi-round protocol behaves in practice. Figure 5 reports accuracy at a low temperature ($T=0.3$) and a higher temperature ($T=1.2$) for three settings: a solo baseline that always uses a single GP agent, a fixed group baseline that always uses the same MDT or ICT team, and our escalation strategy in AIM². Solo and fixed group performance varies with temperature and shows only modest gains at higher temperature. In contrast, the escalation setting remains stable and even improves when T increases, suggesting that the multi-round discussion with upgrading can better exploit the additional diversity in sampled reasoning paths.

To understand consensus behaviour, we also inspected intermediate transcripts from moderate and high complexity cases. At the beginning of a discussion agents often propose diverse diagnoses and plans, especially on image-heavy questions. As the moderator highlights disagreements and requests short follow-up replies, the proposals gradually align and repeated iterations become rare. This qualitative pattern indicates that the collaboration protocol in AIM² tends to turn diverse opinions into a coherent final answer rather than amplifying randomness.

6 Conclusion and Limitations

This work introduced AIM², an adaptive multi-agent framework for medical decision-making that combines complexity-aware orchestration with structured collaboration among LLM agents. The system first estimates query difficulty, then routes each case to an appropriate collaboration level ranging from a single GP agent to an MDT or a full ICT. Within each level, multi-round discussion and a simple escalation rule allow the framework to deepen reasoning only when needed while keeping a clear record of intermediate

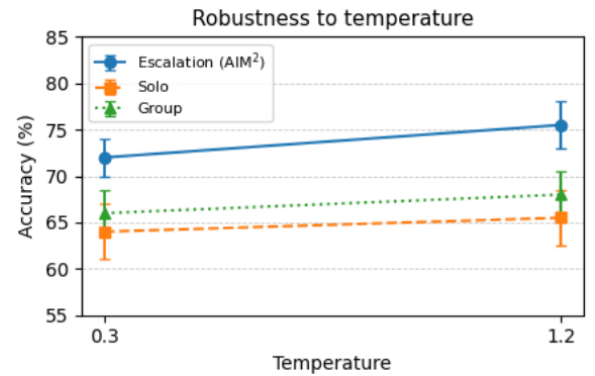


Figure 5: Robustness of AIM² to decoding temperature. We compare the proposed escalation strategy with a solo baseline and a fixed group baseline under low temperature ($T=0.3$) and high temperature ($T=1.2$). Escalation maintains strong performance and even benefits from a higher temperature, while solo and fixed group settings show smaller gains, indicating that multi-round discussion with escalation can better exploit diverse reasoning paths.

notes and disagreements. Across five medical benchmarks AIM² attains competitive or superior accuracy compared with strong single-agent and multi-agent baselines, especially on visual and image+text datasets. Ablation studies show that the escalation policy improves over static complexity settings, that small teams can capture most of the benefits of collaboration and that the framework remains robust across different sampling temperatures.

Despite these encouraging results, AIM² has several limitations that motivate future work.

- **Backbone model and data.** All experiments rely on a single proprietary LLM and a fixed set of public benchmarks.

Performance may change with other models, languages or local hospital data. Evaluating AIM² with open-weight models, additional modalities and non-English datasets is an important next step.

- **Clinical realism and safety.** Although the framework is inspired by MDT and ICT workflows, it has not been tested in real clinical environments and does not have formal safety guarantees. The current work measures accuracy and simple error modes but does not include prospective studies or detailed human-in-the-loop audits. Future work should involve clinicians, examine failure cases in depth and integrate stricter guardrails and calibration.
- **Complexity prediction and orchestration.** The difficulty agent and escalation rules are based on prompt-engineered heuristics and light supervision. Misclassified cases can lead to under- or over-collaboration. Learning these policies from data, for example via reinforcement learning or bandit methods, may yield better trade-offs between cost and accuracy.
- **Cost and latency.** While AIM² reduces API calls relative to always using large teams, multi-round discussions still incur higher cost and latency than single-agent baselines. Exploring more aggressive early-stopping criteria, partial context sharing and model distillation for frequent patterns could make the system more practical at scale.
- **Evaluation of explanations.** AIM² produces rationales and brief reports, but our evaluation focuses mainly on answer correctness. We do not systematically assess the faithfulness, usefulness or cognitive load of these explanations for clinicians or patients. Future studies should include human evaluation of explanation quality and investigate how evidence presentation affects trust and decision-making.

Overall, AIM² shows that complexity-aware escalation and structured collaboration can make LLM-based systems more aligned with real medical decision processes, while keeping computational costs manageable [9, 48]. We hope that this framework and its analysis will encourage further work on reliable, interpretable and clinically grounded multi-agent systems.

References

- [1] Seongsu Bae, Daeun Kyung, Jaehye Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kwon, Jungwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. 2023. EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images. In *Advances in Neural Information Processing Systems (NeurIPS)*. Datasets and Benchmarks Track. doi:10.48550/arXiv.2310.18652 arXiv:2310.18652.
- [2] Seongsu Bae, Daeun Kyung, Jaehye Ryu, Eunbyeol Cho, Gyubok Lee, Sunjun Kwon, Jeongwoo Oh, Lei Ji, Eric I-Chao Chang, Tackeun Kim, and Edward Choi. 2023. EHRXQA: A Multi-Modal Question Answering Dataset for Electronic Health Records with Chest X-ray Images. In *Proceedings of the 37th International Conference on Neural Information Processing Systems (NeurIPS 2023)*. Association for Computing Machinery, New Orleans, LA, USA, 3867–3880. doi:10.5555/3666122.3666292
- [3] Michael L. Barnett, Nancy L. Keating, Nicholas A. Christakis, A. James O'Malley, and Bruce E. Landon. 2012. Reasons for choice of referral physician among primary care and specialist physicians. *Journal of General Internal Medicine* 27, 5 (2012), 506–512. doi:10.1007/s11606-011-1861-z
- [4] Justin Bitter, Elizabeth van Veen-Berkx, Hein Gooszen, and Ludovic van Amelsvoort. 2013. Multidisciplinary teamwork is an important issue to healthcare professionals. *Team Performance Management* 19, 5/6 (2013), 263–278. doi:10.1108/TPM-10-2012-0034
- [5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165* (2020). <https://arxiv.org/abs/2005.14165>
- [6] Justin Chih-Yao Chen, Swarnadeep Saha, and Mohit Bansal. 2023. ReConcile: Round-Table Conference Improves Reasoning via Consensus Among Diverse LLMs. *arXiv preprint arXiv:2309.13007* (2023).
- [7] Michael Christ, Florian Grossmann, Daniela Winter, Roland Bingisser, and Elke Platz. 2010. Modern triage in the emergency department. *Deutsches Ärzteblatt International* 107, 50 (2010), 892–898. doi:10.3238/arztebl.2010.0892
- [8] Michael Christ, Florian Grossmann, Daniela Winter, Roland Bingisser, and Elke Platz. 2010. Modern triage in the emergency department. *Deutsches Ärzteblatt International* 107, 50 (2010), 892–898. doi:10.3238
- [9] Jan Clusmann, Fiona R. Kolbinger, Hannah Sophie Muti, Zunamys I. Carrero, Jan-Niklas Eckardt, Narmin Ghaffari Laleh, Chiara Maria Lavinia Löffler, Sophie-Caroline Schwarzkopf, Michaela Unger, Gregory P. Veldhuizen, Sophia J. Wagner, and Jakob Nikolas Kather. 2023. The future landscape of large language models in medicine. *Communications Medicine* 3, 1 (2023), 141. doi:10.1038/s43856-023-00370-1
- [10] Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. *arXiv preprint arXiv:2305.14325* (2023).
- [11] C. Ee, J. Lake, J. Firth, F. Hargraves, M. de Manincor, T. Meade, W. Marx, and J. Sarris. 2020. An integrative collaborative care model for people with mental illness and physical comorbidities. *International Journal of Mental Health Systems* 14, 1 (2020), 83. doi:10.1186/s13033-020-00410-6
- [12] C. Ee, J. Lake, J. Firth, F. Hargraves, M. de Manincor, T. Meade, W. Marx, and J. Sarris. 2020. An integrative collaborative care model for people with mental illness and physical comorbidities. *International Journal of Mental Health Systems* 14, 1 (2020), 83. doi:10.1186/s13033-020-00410-6
- [13] Nancy E. Epstein. 2014. Multidisciplinary in-hospital teams improve patient outcomes: A review. *Surgical Neurology International* 5, Suppl 7 (2014), S295–S303. doi:10.4103/2152-7806.139612
- [14] Nancy E. Epstein. 2014. Multidisciplinary in-hospital teams improve patient outcomes: A review. *Surgical Neurology International* 5, Suppl 7 (2014), S295–S303. doi:10.4103/2152-7806.139612
- [15] Amy L. Garcia. 2017. Variability in Acuity in Acute Care. *Journal of Nursing Administration* 47, 10 (2017), 476–483. doi:10.1097/NNA.0000000000000518
- [16] Nicki Gilboy, Paula Tanabe, Debbie Travers, and Alexander M. Rosenau. 2011. *Emergency Severity Index (ESI): A triage tool for emergency department care, version 4. Implementation handbook, 2012 edition*. Technical Report. Agency for Healthcare Research and Quality. AHRQ publication.
- [17] David Grembowski, Judith Schaefer, Karin E. Johnson, Henry Fischer, Susan L. Moore, Ming Tai-Seale, Richard Ricciardi, James R. Fraser, Donald Miller, and Lisa LeRoy. 2014. A conceptual model of the role of complexity in the care of patients with multiple chronic conditions. *Medical Care* 52, Suppl 3 (2014), S7–S14. doi:10.1097/MLR.0000000000000045
- [18] David Grembowski, Judith Schaefer, Karin E. Johnson, Henry Fischer, Susan L. Moore, Ming Tai-Seale, Richard Ricciardi, James R. Fraser, Donald Miller, and Lisa LeRoy. 2014. A conceptual model of the role of complexity in the care of patients with multiple chronic conditions. *Medical Care* 52, Suppl 3 (2014), S7–S14. doi:10.1097/MLR.0000000000000045
- [19] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Transactions on Computing for Healthcare* 3, 1 (2021), 1–23. doi:10.1145/3458754
- [20] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. 2020. PathVQA: 30000+ Questions for Medical Visual Question Answering. *arXiv preprint arXiv:2003.10286* (2020). doi:10.48550/arXiv.2003.10286
- [21] Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. 2024. Uncertainty of Thoughts: Uncertainty-Aware Planning Enhances Information Seeking in Large Language Models. *arXiv preprint arXiv:2402.03271* (2024). arXiv:2402.03271 [cs.CL]
- [22] Maria Jimenez-Lara. 2016. Reaping the Benefits of Integrated Health Care. *Stanford Social Innovation Review* (Sept. 2016). https://ssir.org/articles/entry/reaping_the_benefits_of_integrated_health_care Accessed: 2025-11-26.
- [23] Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What Disease Does This Patient Have? A Large-Scale Open Domain Question Answering Dataset from Medical Exams. *Applied Sciences* 11, 14 (2021), 6421. doi:10.3390/app11146421
- [24] Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. PubMedQA: A Dataset for Biomedical Research Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language*

- Processing (EMNLP-IJCNLP). Association for Computational Linguistics, Hong Kong, China, 2567–2577.
- [25] Qiao Jin, Zhizheng Wang, Yifan Yang, Qingqing Zhu, Donald Wright, Thomas Huang, Nikhil Khandekar, Nicholas Wan, Xuguang Ai, W. John Wilbur, Zhe He, R. Andrew Taylor, Qingyu Chen, and Zhiyong Lu. 2025. AgentMD: Empowering language agents for risk prediction with large-scale clinical tool learning. *Nature Communications* 16 (2025), 9377. doi:10.1038/s41467-025-64430-x
 - [26] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.
 - [27] Danielle L. LaFrance, Mary Jane Weiss, Ellie Kazemi, Joanne Gerenscser, and Jacqueline Dobres. 2019. Multidisciplinary Teaming: Enhancing Collaboration through Increased Understanding. *Behavior Analysis in Practice* 12, 3 (2019), 709–726. doi:10.1007/s40617-019-00331-y
 - [28] Junkai Li, Yunhui Lai, Weitao Li, Jingyi Ren, Meng Zhang, Nianqi Zhang, Xinhui Kang, Siyu Wang, Peng Li, Ya-Qin Zhang, Weizhi Ma, and Yang Liu. 2024. Agent Hospital: A Simulacrum of Hospital with Evolvable Medical Agents. *arXiv preprint arXiv:2405.02957* (2024). arXiv:2405.02957 [cs.AI]
 - [29] Shuyue Stella Li, Vidhisha Balachandran, Shangqin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024. MediQ: Question-Asking LLMs and a Benchmark for Reliable Interactive Clinical Reasoning. *arXiv preprint arXiv:2406.00922* (2024). arXiv:2406.00922 [cs.CL]
 - [30] Zijun Liu, Yanzhe Zhang, Peng Li, Yang Liu, and Diyi Yang. 2023. A Dynamic LLM-Powered Agent Network for Task-Oriented Agent Collaboration. *arXiv preprint arXiv:2310.02170* (2023).
 - [31] Ateev Mehrotra, Christopher B. Forrest, and Caroline Y. Lin. 2011. Dropping the baton: specialty referrals in the United States. *The Milbank Quarterly* 89, 1 (2011), 39–68. doi:10.1111/j.1468-0009.2011.00619.x
 - [32] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M. Krumholz, Jure Leskovec, Eric J. Topol, and Pranav Rajpurkar. 2023. Foundation models for generalist medical artificial intelligence. *Nature* 616, 7956 (2023), 259–265. doi:10.1038/s41586-023-05881-4
 - [33] Tong Mu, Alec Helyar, Johannes Heidecke, Joshua Achiam, Andrea Vallone, Ian Kivlichan, Molly Lin, Alex Beutel, John Schulman, and Lilian Weng. 2024. Rule Based Rewards for Language Model Safety. *arXiv preprint arXiv:2411.01111* (2024).
 - [34] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375* (2023). arXiv:2303.13375 [cs.CL]
 - [35] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. *arXiv preprint arXiv:2303.13375* (2023). <https://arxiv.org/abs/2303.13375>
 - [36] Harsha Nori, Yin Tat Lee, Sheng Zhang, Dean Carignan, Richard Edgar, Nicolo Fusi, Nicholas King, Jonathan Larson, Yuanzhi Li, Weishung Liu, Rengqian Luo, Scott Mayer McKinney, Robert Osazuwa Ness, Hoifung Poon, Tao Qin, Naoto Usuyama, Chris White, and Eric Horvitz. 2023. Can Generalist Foundation Models Outperform Special-Purpose Tuning? Case Study in Medicine. *arXiv preprint arXiv:2311.16452* (2023). arXiv:2311.16452 [cs.CL]
 - [37] Anand K. Parekh, Richard A. Goodman, Catherine Gordon, and Howard K. Koh. 2011. Managing multiple chronic conditions: a strategic framework for improving health outcomes and quality of life. *Public Health Reports* 126, 4 (2011), 460–471. doi:10.1177/00333549112600403
 - [38] Anand K. Parekh, Richard A. Goodman, Catherine Gordon, and Howard K. Koh. 2011. Managing multiple chronic conditions: a strategic framework for improving health outcomes and quality of life. *Public Health Reports* 126, 4 (2011), 460–471. doi:10.1177/00333549112600403
 - [39] Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. 2024. Do LLM Agents Have Regret? A Case Study in Online Learning and Games. *arXiv preprint arXiv:2403.16843* (2024).
 - [40] Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. 2024. Do LLM Agents Have Regret? A Case Study in Online Learning and Games. *arXiv preprint arXiv:2403.16843* (2024). <https://arxiv.org/abs/2403.16843>
 - [41] Khaled Saab, Tao Tu, Wei-Hung Weng, Ryutaro Tanno, David Stutz, Ellery Wulczyn, Fan Zhang, Tim Strother, Chunjong Park, Elahe Vedadi, Juanma Zambrano Chaves, Szu-Yeu Hu, Mike Schaekermann, Aishwarya Kamath, Yong Cheng, David G. T. Barrett, Cathy Cheung, Basil Mustafa, Anil Palepu, Daniel McDuff, Le Hou, Tomer Golany, Luyang Liu, Jean-baptiste Alayrac, Neil Houlsby, Nenad Tomasev, Jan Freyberg, Charles Lau, Jonas Kemp, Jeremy Lai, Shekoofeh Azizi, Kimberly Kanada, SiWai Man, Kavita Kulkarni, Ruoxi Sun, Siamak Shakeri, Luheng He, Ben Caine, Albert Webson, Natasha Latysheva, Melvin Johnson, Philip Mansfield, Jian Lu, Ehud Rivlin, Jesper Anderson, Bradley Green, Renee Wong, Jonathan Krause, Jonathan Shlens, Ewa Dominowska, S. M. Ali Eslami, Katherine Chou, Claire Cui, Oriol Vinyals, Koray Kavukcuoglu, James Manyika, Jeff Dean, Demis Hassabis, Yossi Matias, Dale Webster, Joelle Barral, Greg Corrado, Christopher Semturs, S. Sara Mahdavi, Juraj Gottweis, Alan Karthikesalingam, and Vivek Natarajan. 2024. Capabilities of Gemini Models in Medicine. *arXiv preprint arXiv:2404.18416* (2024).
 - [42] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, Perry Payne, Martin Seneviratne, Paul Gamble, Chris Kelly, Nathanael Schärli, Aakanksha Chowdhery, Philip Mansfield, Dina Demner-Fushman, Blaise Agüera y Arcas, Dale Webster, Greg S Corrado, Yossi Matias, Katherine Chou, Juraj Gottweis, Nenad Tomasev, Yun Liu, Alvin Rajkomar, Joelle Barral, Christopher Sumsur, Alan Karthikesalingam, and Vivek Natarajan. 2023. Large Language Models Encode Clinical Knowledge. *Nature* 620, 7972 (2023), 172–180. doi:10.1038/s41586-023-06291-2
 - [43] Harold C. Sox, Michael C. Higgins, Douglas K. Owens, and Gillian Sanders Schilder. 2024. *Medical Decision Making*. John Wiley & Sons, Hoboken, NJ. doi:10.1002/9781119627876
 - [44] J. Stairmand, L. Signal, D. Sarfati, C. Jackson, L. Batten, M. Holdaway, and C. Cunningham. 2015. Consideration of comorbidity in treatment decision making in multidisciplinary cancer team meetings: a systematic review. *Annals of Oncology* 26, 7 (2015), 1325–1332. doi:10.1093/annonc/mdv025
 - [45] Miren Taberna, Francisco Gil Moncayo, Enric Jané-Salas, Maite Antonio, Lorena Arribas, Esther Vilajosana, Elisabet Peralvez Torres, and Ricard Mesia. 2020. The Multidisciplinary Team (MDT) Approach and Quality of Care. *Frontiers in Oncology* 10 (2020), 85. doi:10.3389/fonc.2020.00085
 - [46] Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. MedAgents: Large Language Models as Collaborators for Zero-shot Medical Reasoning. *arXiv preprint arXiv:2311.10537* (2023).
 - [47] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine* 29, 8 (2023), 1930–1940. doi:10.1038/s41591-023-02448-8
 - [48] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature Medicine* 29, 8 (2023), 1930–1940. doi:10.1038/s41591-023-02448-8
 - [49] Tao Tu, Mike Schaekermann, Anil Palepu, Khaled Saab, Jan Freyberg, Ryutaro Tanno, Amy Wang, Brenna Li, Mohamed Amin, Yong Cheng, Elahe Vedadi, Nenad Tomasev, Shekoofeh Azizi, Karan Singhal, Le Hou, Albert Webson, Kavita Kulkarni, S. Sara Mahdavi, Christopher Semturs, Juraj Gottweis, Joelle Barral, Katherine Chou, Greg S. Corrado, Yossi Matias, and Vivek Natarajan. 2025. Towards conversational diagnostic artificial intelligence. *Nature* 642, 8067 (2025), 442–450. doi:10.1038/s41586-025-08866-7
 - [50] Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine* 30, 4 (2024), 1134–1142. doi:10.1038/s41591-024-02855-5
 - [51] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji-Rong Wen. 2023. A Survey on Large Language Model based Autonomous Agents. *arXiv preprint arXiv:2308.11432* (2023). arXiv:2308.11432 [cs.AI]
 - [52] Sheng Wang, Zihao Zhao, Xi Ouyang, Tianming Liu, Qian Wang, and Dinggang Shen. 2024. Interactive computer-aided diagnosis on medical image using large language models. *Communications Engineering* 3 (2024), 133. doi:10.1038/s44172-024-00271-8
 - [53] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171* (2022).
 - [54] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint arXiv:2203.11171* (2023). <https://arxiv.org/abs/2203.11171>
 - [55] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.
 - [56] Robin M. Weinick, Rachel M. Burns, and Ateev Mehrotra. 2010. Many emergency department visits could be managed at urgent care centers and retail clinics. *Health Affairs* 29, 9 (2010), 1630–1636. doi:10.1377/hlthaff.2009.0748
 - [57] Robin M. Weinick, Rachel M. Burns, and Ateev Mehrotra. 2010. Many emergency department visits could be managed at urgent care centers and retail clinics. *Health Affairs (Millwood)* 29, 9 (2010), 1630–1636. doi:10.1377/hlthaff.2009.0748
 - [58] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryan W. White, Doug Burger, and Chi Wang. 2023. AutoGen: Enabling Next-Gen LLM Applications via Multi-Agent Conversation. *arXiv preprint arXiv:2308.08155* (2023). arXiv:2308.08155 [cs.CL]

- [59] R. C. Wuerz, L. W. Milne, D. R. Eitel, D. Travers, and N. Gilboy. 2000. Reliability and validity of a new five-level triage instrument. *Academic Emergency Medicine* 7, 3 (2000), 236–242. doi:10.1111/j.1553-2712.2000.tb01066.x
- [60] R. C. Wuerz, L. W. Milne, D. R. Eitel, D. Travers, and N. Gilboy. 2000. Reliability and validity of a new five-level triage instrument. *Academic Emergency Medicine* 7, 3 (2000), 236–242. doi:10.1111/j.1553-2712.2000.tb01066.x
- [61] Guangzhi Xiong, Qiao Jin, Zhiyong Lu, and Aidong Zhang. 2024. Benchmarking Retrieval-Augmented Generation for Medicine. In *Findings of the Association for Computational Linguistics: ACL 2024*. Association for Computational Linguistics, Bangkok, Thailand, 6233–6251.
- [62] Weixiang Yan, Haitian Liu, Tengxiao Wu, Qian Chen, Wen Wang, Haoyuan Chai, Jiayi Wang, Weishan Zhao, Yixin Zhang, Renjun Zhang, Li Zhu, and Xuandong Zhao. 2024. ClinicaLab: Aligning Agents for Multi-Departmental Clinical Diagnostics in the Real World. *arXiv preprint arXiv:2406.13890* (2024). arXiv:2406.13890 [cs.CL]
- [63] Xiaoman Zhang, Chaoyi Wu, Ziheng Zhao, Weixiong Lin, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2023. PMC-VQA: Visual Instruction Tuning for Medical Visual Question Answering. *arXiv preprint arXiv:2305.10415* (2023). doi:10.48550/arXiv.2305.10415 last revised 2024.
- [64] Hongjian Zhou, Fenglin Liu, Boyang Gu, Xinyu Zou, Jinfa Huang, Jing Wu, Yiru Li, Sam S. Chen, Peilin Zhou, Junling Liu, Yining Hua, Chengfeng Mao, Chenyu You, Xian Wu, Yefeng Zheng, Lei Clifton, Zheng Li, Jiebo Luo, and David A. Clifton. 2023. A Survey of Large Language Models in Medicine: Progress, Application, and Challenge. *arXiv preprint arXiv:2311.05112* (2023). arXiv:2311.05112 [cs.CL]
- [65] Junbin Zhou and Xiao Xu. 2023. The difficulty of medical decision-making: should patients be involved? *Hepatobiliary Surgery and Nutrition* 12, 3 (2023), 407–409. doi:10.21037/hbsn-23-245

A AIM² Decision Procedure

Algorithm 1 AIM² Medical Collaboration and Decision Process

Require: query q , context c

Ensure: final answer a , explanation e , complexity level L

```

1:  $L \leftarrow \text{DIFFICULTYAGENT}(q, c)$ 
   while true do
   — line:loop-begin
2:  $\mathcal{T} \leftarrow \text{RECRUITTEAM}(L)$ 
   if  $L = \text{Low}$  then
   — line:low-begin
3:  $(a_{\text{gp}}, n_{\text{gp}}) \leftarrow \text{GPANSWER}(q, c, \mathcal{T})$  if  $\text{SELFHECK}(a_{\text{gp}}, n_{\text{gp}})$  then
4:   return  $(a_{\text{gp}}, n_{\text{gp}}, L)$  else
5:    $L \leftarrow \text{Moderate}$  ▷ escalate
6: continue
7:
8:
9:  $\text{history} \leftarrow \emptyset$  for  $r = 1$  to  $\text{MAXROUNDS}(L)$  do
10:    $\text{opinions} \leftarrow \text{COLLECTNOTES}(\mathcal{T}, q, c, \text{history})$ 
11:    $\text{critique} \leftarrow \text{CHALLENGERCRTIQUE}(\mathcal{T}, \text{opinions})$ 
12:    $\text{summary} \leftarrow \text{MODERATORSUMMARY}(\text{opinions}, \text{critique})$ 
13:    $(\text{quality}, \text{consensus}) \leftarrow \text{JUDGESCORE}(\text{summary})$ 
14:    $\text{history} \leftarrow \text{history} \cup \{(\text{opinions}, \text{summary})\}$  if
      $\text{ACCEPT}(\text{quality}, \text{consensus})$  then
15:     break
16:
17:
   if  $L = \text{Moderate}$  and  $\text{NEEDESCALATION}(\text{quality}, \text{consensus})$ 
   then
   — line:mod-escalate-begin
18:  $L \leftarrow \text{High}$  ▷ MDT  $\rightarrow$  ICT
19: continue
20:
21:  $(a, e) \leftarrow \text{SYNTHESISAGENT}(\text{history}, L)$ 
22: return  $(a, e, L)$ 
23:

```
