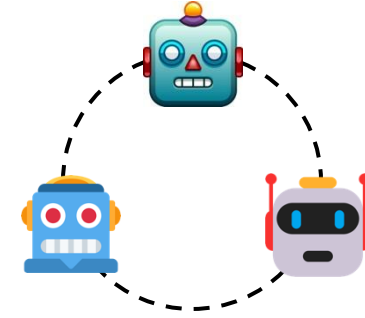


# MDAgents: An Adaptive Collaboration of LLMs in Medical Decision Making



Yubin Kim<sup>1</sup>, Chanwoo Park<sup>1</sup>, Hyewon Jeong<sup>1</sup>, Yik Siu Chan<sup>1</sup>, Xuhai Xu<sup>1</sup>, Daniel McDuff<sup>2</sup>, Hyeonhoon Lee<sup>3</sup>, Marzyeh Ghassemi<sup>1</sup>, Cynthia Breazeal<sup>1</sup>, Hae Won Park<sup>1</sup>

<sup>1</sup>Massachusetts Institute of Technology

<sup>2</sup>Google Research

<sup>3</sup>Seoul National University Hospital



# Robustness of LLM as a Medical Assistant



Hugh Harvey ✓  
@DrHughHarvey

...

Good work from @StanfordHAI showing that GPT 4 is not robust enough for use as a medical co-pilot for the following important reasons:

- 1) Non-deterministic: They found low similarity and high variability in responses to the same question. Jaccard and cosine similarity coefficients were merely 0.29 and 0.45 respectively.
- 2) Accuracy: Only 41% of GPT-4 responses agreed with the known answer to medical questions according to a consensus of 12 physicians.
- 3) Potential for harm: 7% of answers were deemed potentially harmful by the consensus physicians.



“Everyone in favor raise your hand!”

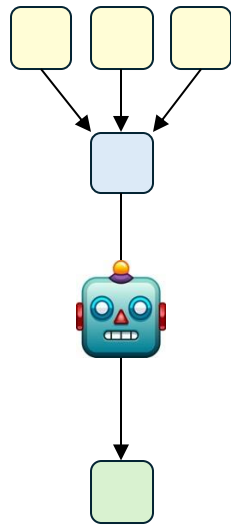
Fig. Typical problem in group decision making

*This work explores LLM collaborations in the context of medical decision-making, ranging from **individual perspectives** to **collective consensus***

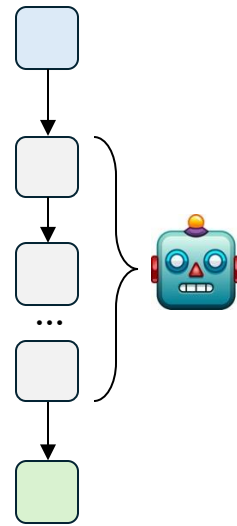
# Decision Making with LLM Agent(s)



Zero-shot

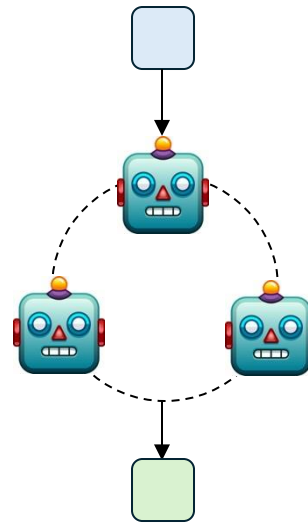


Few-shot

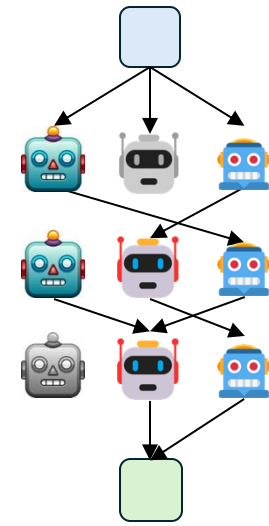


CoT

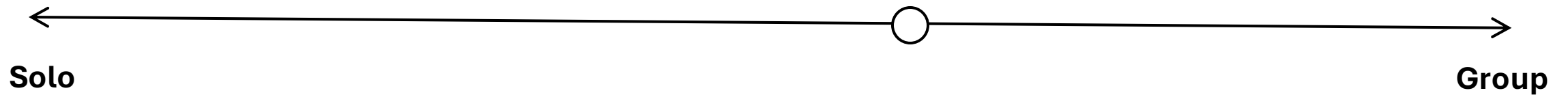
...



Voting



Debate

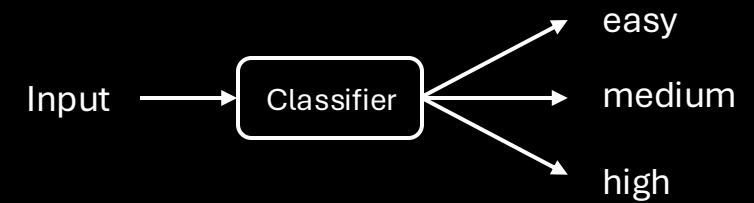


*Where is this optimal point within this spectrum?*

*Why do we need to select the **same** methods across the problems?*

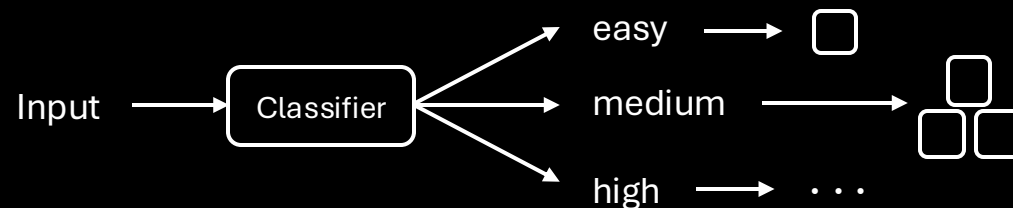
*Why do we need to select the **same** methods across the problems?*

❖ Let's consider the complexity of the input problem!

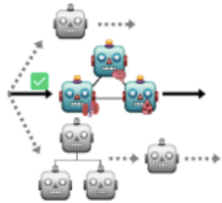



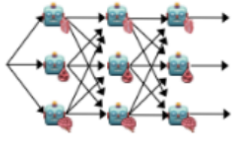
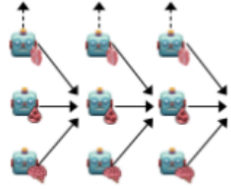


## Why do we need to select the *same* methods across the problems?

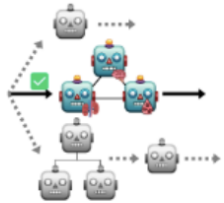

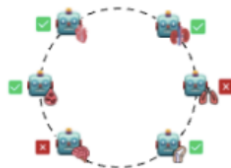
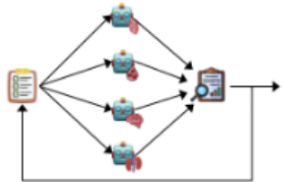
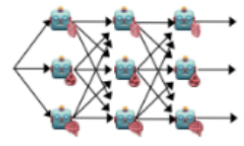
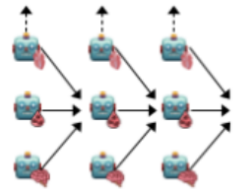
- ❖ Let's consider the complexity of the input problem!
- ❖ Based on the input complexity, let's use different methods to solve the problem!



# Ours and previous approaches

Method	MDAgents (Ours)	Single	Voting [75]	Delphi [64]	Debate [12]	ReConcile [6]
<b>Interaction Type</b>						
<b>Multiple Roles</b>	✓	✗	✓	✓	✓	✓
<b>Early Stopping</b>	✓	✗	✓	✓	✓	✗
<b>Adaptive Structure</b>	✓	✗	✗	✗	✗	✗
<b>Refinement</b>	✓	✗	✗	✓	✓	✗
<b>Conversation Pattern</b>	flexible	static	static	static	static	static

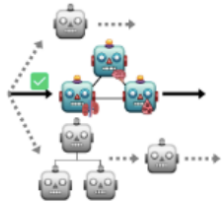



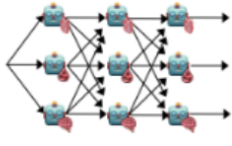
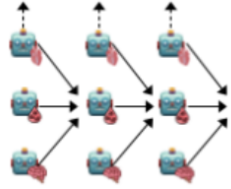
# Ours and previous approaches

Method	MDAgents (Ours)	Single	Voting [75]	Delphi [64]	Debate [12]	ReConcile [6]
<b>Interaction Type</b>						
<b>Multiple Roles</b>	✓	✗	✓	✓	✓	✓
<b>Early Stopping</b>	✓	✗	✓	✓	✓	✗
<b>Adaptive Structure</b>	✓	✗	✗	✗	✗	✗
<b>Refinement</b>	✓	✗	✗	✓	✓	✗
<b>Conversation Pattern</b>	flexible	static	static	static	static	static

- MDAgents is the first work to *adaptively* determine the inference framework based on the input complexity



# Ours and previous approaches

Method	MDAgents (Ours)	Single	Voting [75]	Delphi [64]	Debate [12]	ReConcile [6]
<b>Interaction Type</b>						
<b>Multiple Roles</b>	✓	✗	✓	✓	✓	✓
<b>Early Stopping</b>	✓	✗	✓	✓	✓	✗
<b>Adaptive Structure</b>	✓	✗	✗	✗	✗	✗
<b>Refinement</b>	✓	✗	✗	✓	✓	✗
<b>Conversation Pattern</b>	flexible	static	static	static	static	static

- MDAgents is the first work to *adaptively* determine the inference framework based on the input complexity
- This adaptive structure mirrors real-world clinical decision making process

# Framework

## Medical Knowledge

A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. Which of the following is the most likely diagnosis? A. Tinea versicolor B. Albinism **C. Vitiligo** D. Solar lentigo E. Melanoma

# Framework

## Medical Knowledge

A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. Which of the following is the most likely diagnosis? A. Tinea versicolor B. Albinism **C. Vitiligo** D. Solar lentigo E. Melanoma

## Differential Diagnosis (DDx)

Sex: M, Age: 47 Geographical region: North America

**Pathology: PSVT**

Symptoms:

- I feel pain.
- The pain is: tugging, Burning ...

Differential diagnosis:

PSVT: 0.22, Anemia: 0.16, Panic attack: 0.14, Atrial fibrillation: 0.11, Anaphylaxis: 0.11, Cluster headache: 0.09, Chagas: 0.07, Scombroid food poisoning: 0.07, HIV (initial infection): 0.01

# Framework

## Medical Knowledge

A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. Which of the following is the most likely diagnosis? A. Tinea versicolor B. Albinism **C. Vitiligo** D. Solar lentigo E. Melanoma

## Differential Diagnosis (DDx)

Sex: M, Age: 47 Geographical region: North America

**Pathology: PSVT**

Symptoms:

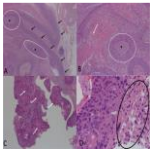
- I feel pain.
- The pain is: tugging, Burning ...

Differential diagnosis:

PSVT: 0.22, Anemia: 0.16, Panic attack: 0.14, Atrial fibrillation: 0.11, Anaphylaxis: 0.11, Cluster headache: 0.09, Chagas: 0.07, Scombroid food poisoning: 0.07, HIV (initial infection): 0.01

## Multi-modal Reasoning

What does the circle in image D surround?



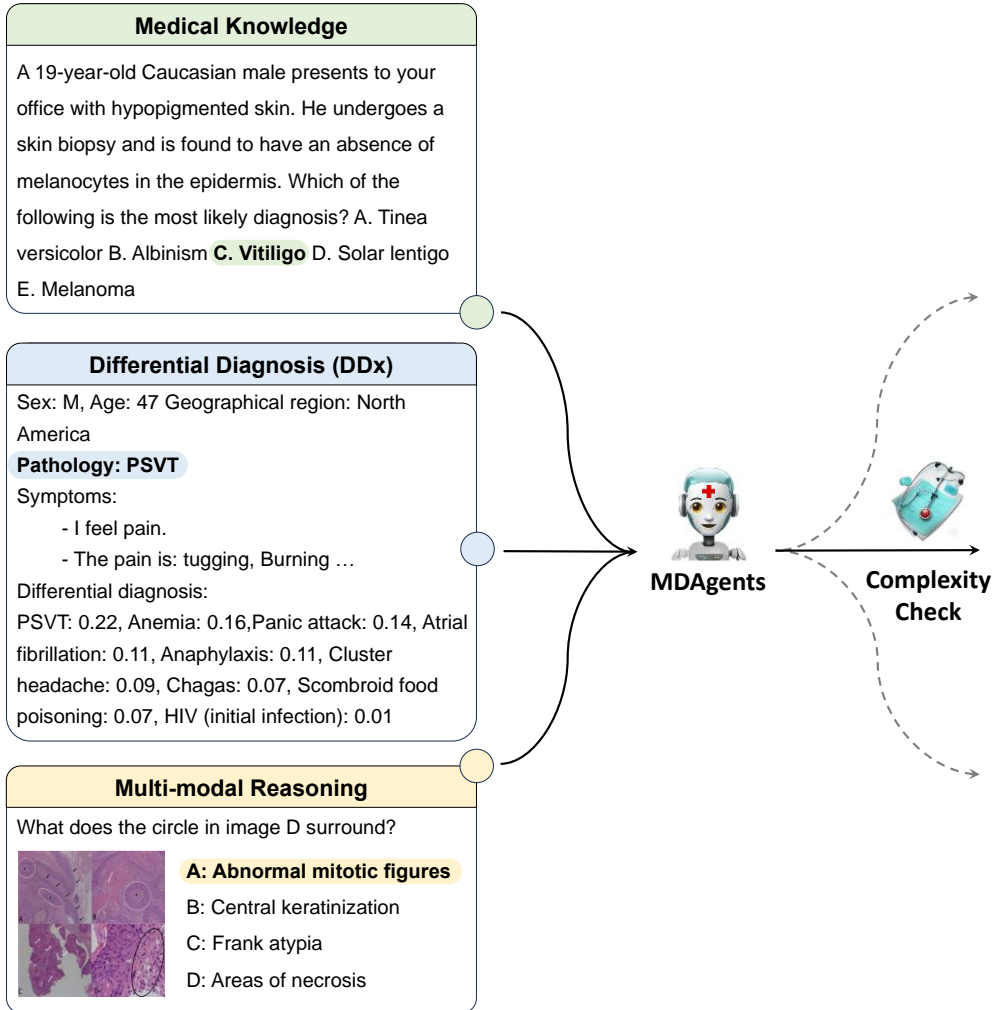
**A: Abnormal mitotic figures**

B: Central keratinization

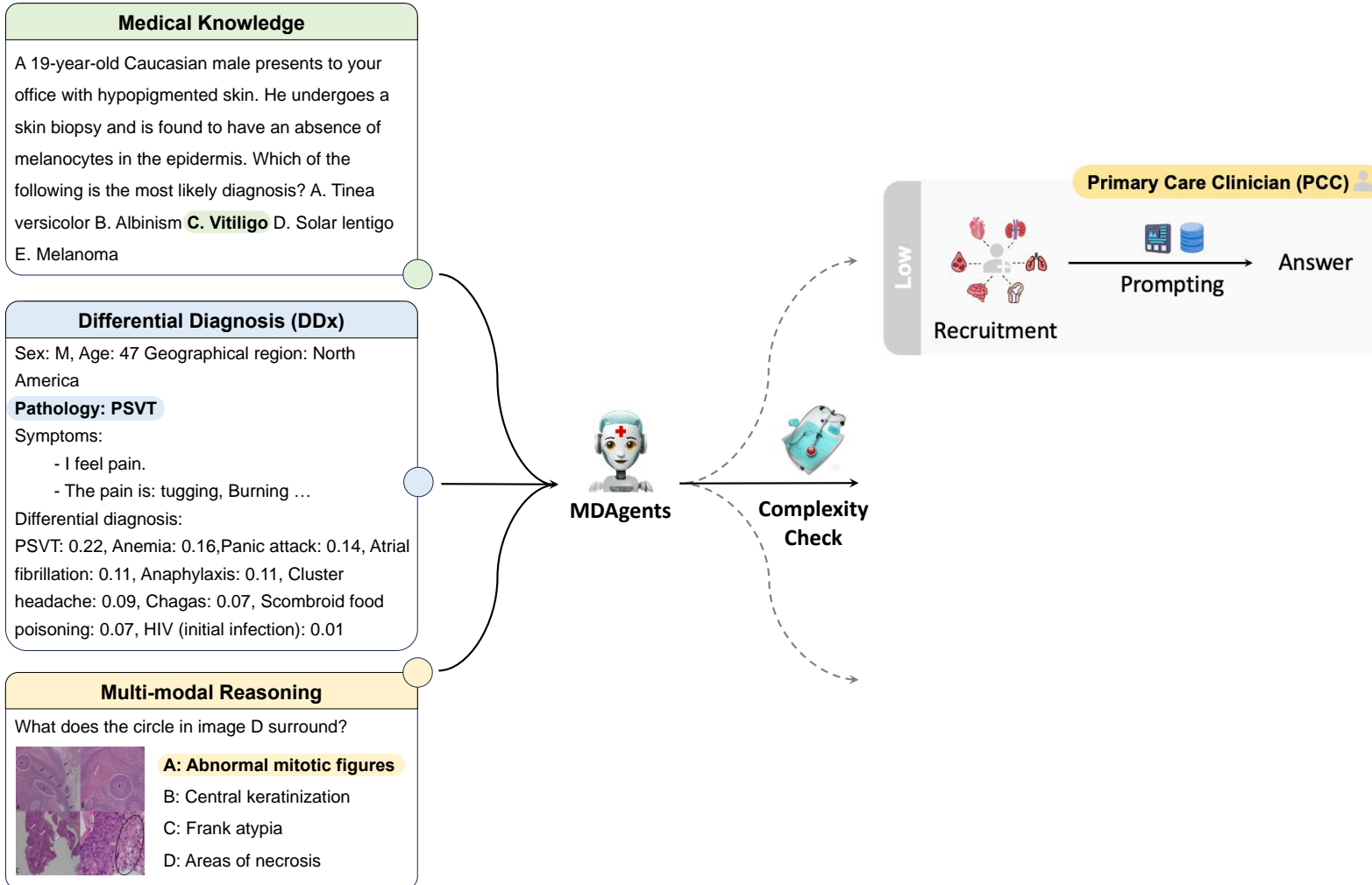
C: Frank atypia

D: Areas of necrosis

# Framework



# Framework



# Framework

**Medical Knowledge**

A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. Which of the following is the most likely diagnosis? A. Tinea versicolor B. Albinism **C. Vitiligo** D. Solar lentigo E. Melanoma

**Differential Diagnosis (DDx)**

Sex: M, Age: 47 Geographical region: North America

**Pathology: PSVT**

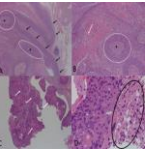
Symptoms:

- I feel pain.
- The pain is: tugging, Burning ...

Differential diagnosis:  
 PSVT: 0.22, Anemia: 0.16, Panic attack: 0.14, Atrial fibrillation: 0.11, Anaphylaxis: 0.11, Cluster headache: 0.09, Chagas: 0.07, Scombroid food poisoning: 0.07, HIV (initial infection): 0.01

**Multi-modal Reasoning**

What does the circle in image D surround?



**A: Abnormal mitotic figures**

B: Central keratinization

C: Frank atypia

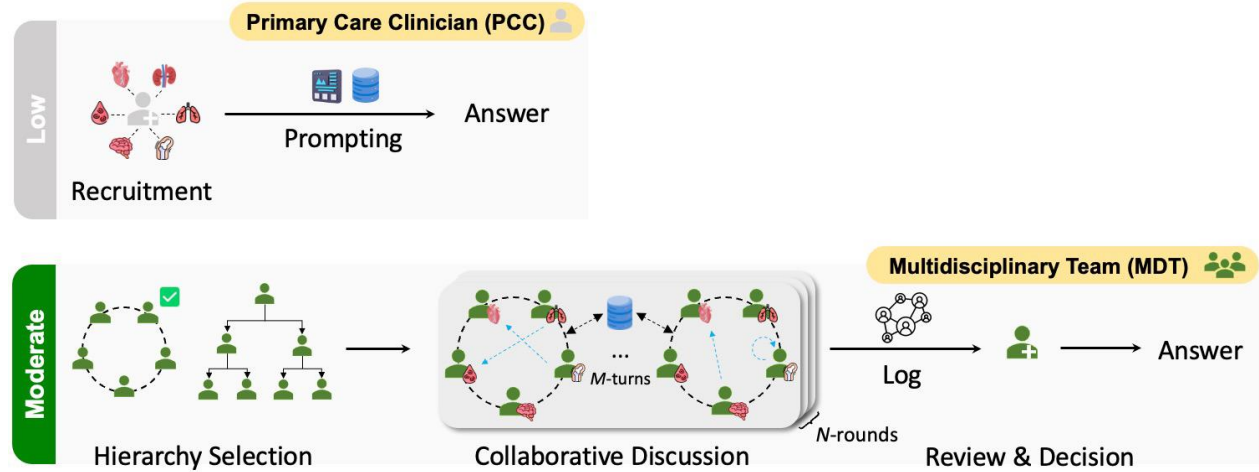
D: Areas of necrosis



MDAgents



Complexity Check



# Framework

**Medical Knowledge**

A 19-year-old Caucasian male presents to your office with hypopigmented skin. He undergoes a skin biopsy and is found to have an absence of melanocytes in the epidermis. Which of the following is the most likely diagnosis? A. Tinea versicolor B. Albinism **C. Vitiligo** D. Solar lentigo E. Melanoma

**Differential Diagnosis (DDx)**

Sex: M, Age: 47 Geographical region: North America

**Pathology: PSVT**

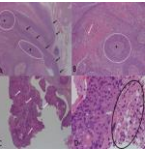
Symptoms:

- I feel pain.
- The pain is: tugging, Burning ...

Differential diagnosis:  
 PSVT: 0.22, Anemia: 0.16, Panic attack: 0.14, Atrial fibrillation: 0.11, Anaphylaxis: 0.11, Cluster headache: 0.09, Chagas: 0.07, Scombroid food poisoning: 0.07, HIV (initial infection): 0.01

**Multi-modal Reasoning**

What does the circle in image D surround?



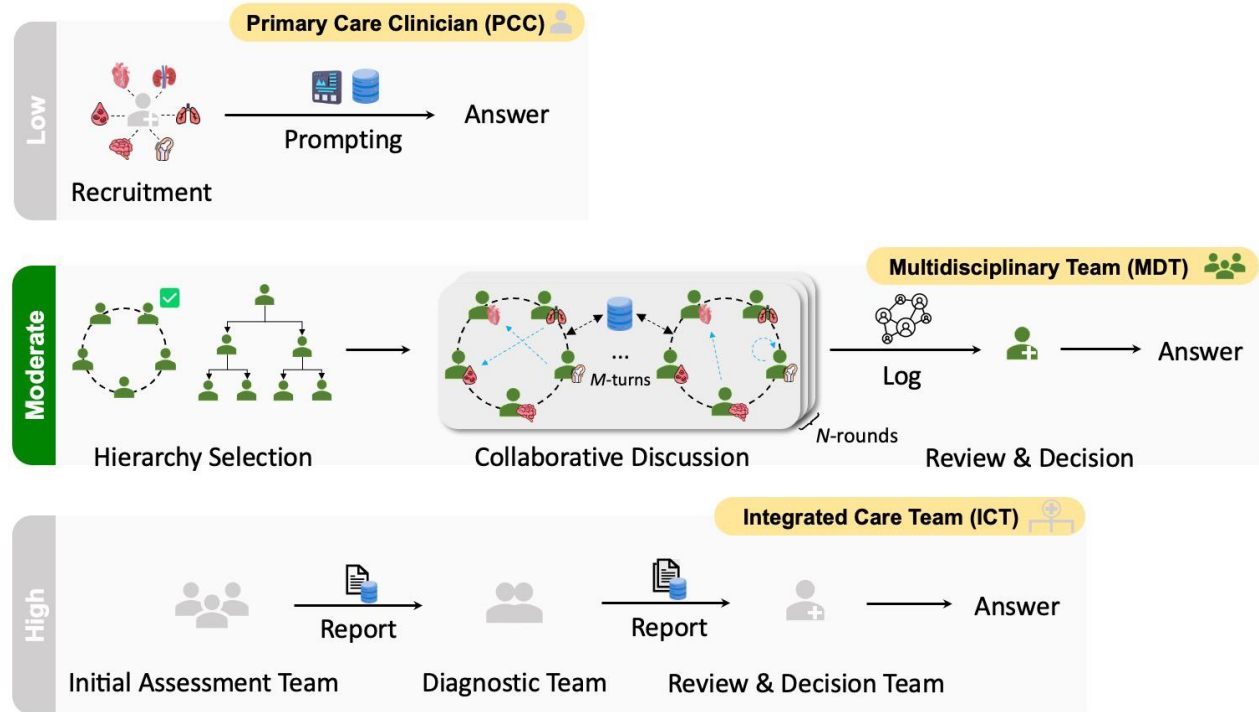
**A: Abnormal mitotic figures**  
 B: Central keratinization  
 C: Frank atypia  
 D: Areas of necrosis



MDAgents



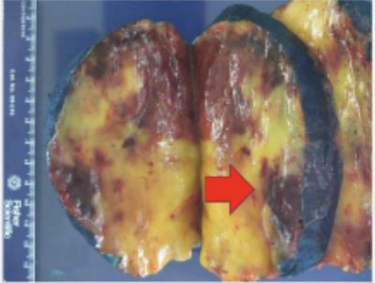
Complexity Check





# Case Study

**Question:** What does the image of Gross Sectioned Myelolipoma demonstrate?

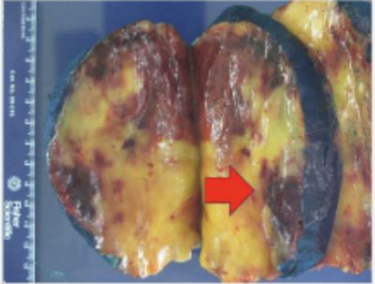


- (A) fibrosis
- (B) inflammation
- (C) hemorrhage
- (D) calcification

**GT:** (C) hemorrhage

## Case Study (Cont.)

**Question:** What does the image of Gross Sectioned Myelolipoma demonstrate?



- (A) fibrosis
- (B) inflammation
- (C) hemorrhage
- (D) calcification

**GT:** (C) hemorrhage

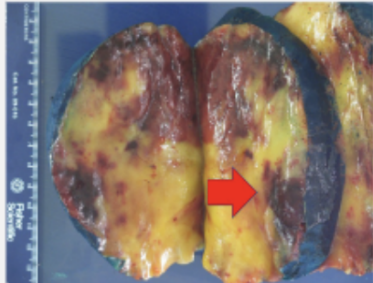
### Step 1. Complexity Check

#### Intermediate

While the identification of gross pathology might seem straightforward to a specialist in pathology, the distinction between fibrosis, hemorrhage, and other conditions can sometimes be subtle and subject to interpretation, as evidenced by the detailed discussion and disagreement among the agents with different medical backgrounds in the provided scenario.

# Case Study (Cont.)

**Question:** What does the image of Gross Sectioned Myelolipoma demonstrate?



- (A) fibrosis
- (B) inflammation
- (C) hemorrhage
- (D) calcification

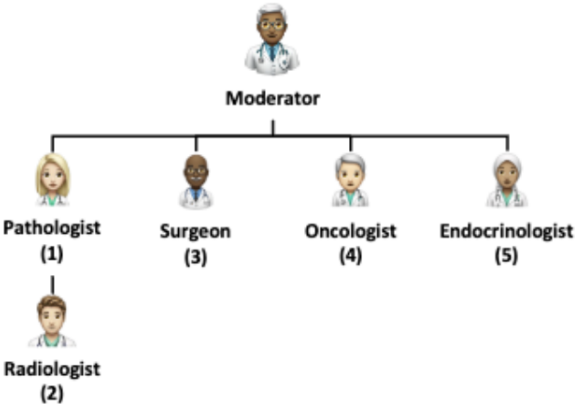
**GT:** (C) hemorrhage

## Step 1. Complexity Check

### Intermediate

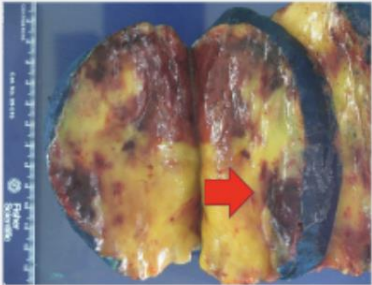
While the identification of gross pathology might seem straightforward to a specialist in pathology, the distinction between fibrosis, hemorrhage, and other conditions can sometimes be subtle and subject to interpretation, as evidenced by the detailed discussion and disagreement among the agents with different medical backgrounds in the provided scenario.

## Step 2. Expert Recruitment



# Case Study (Cont.)

**Question:** What does the image of Gross Sectioned Myelolipoma demonstrate?



- (A) fibrosis
- (B) inflammation
- (C) hemorrhage
- (D) calcification

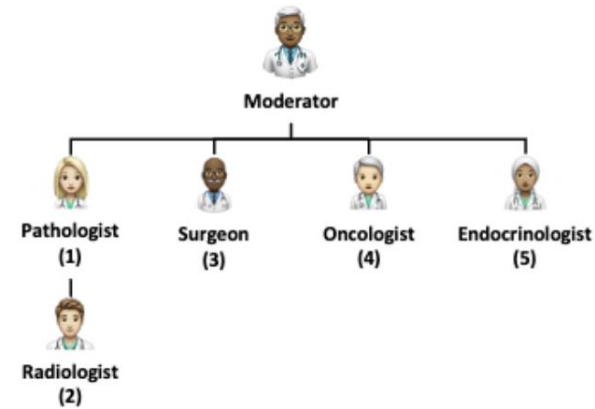
**GT:** (C) hemorrhage

## Step 1. Complexity Check

### Intermediate

While the identification of gross pathology might seem straightforward to a specialist in pathology, the distinction between fibrosis, hemorrhage, and other conditions can sometimes be subtle and subject to interpretation, as evidenced by the detailed discussion and disagreement among the agents with different medical backgrounds in the provided scenario.

## Step 2. Expert Recruitment

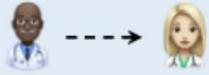


## Step 3. Initial Assessment

- (1) Answer: (A) fibrosis
- (2) Answer: (A) fibrosis
- (3) Answer: (C) hemorrhage
- (4) Answer: (C) hemorrhage
- (5) Answer: (A) fibrosis

# Case Study (Cont.)

## Step 4. Collaborative Discussion



### Turn 1

(3) → (1)

: I am a surgeon who specializes in surgical procedures for various medical conditions. I have extensive experience in diagnosing and treating a wide range of surgical conditions, including cancer, trauma, cardiovascular disease, and gastro intestinal disorders.

I believe that the image shows (C) hemorrhage (red arrow). Hemorrhage is the escape of blood from a blood vessel. It can occur due to a variety of factors, including trauma, bleeding disorders, and certain medical conditions. In the image, the red arrow points to an area of the tissue that is dark red and appears to be surrounded by a thin rim of white tissue. This is consistent with the appearance of hemorrhage.



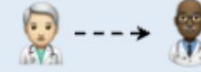
### Turn 1

(3) → (2)

: I am a surgeon who specializes in surgical procedures for various medical conditions.

I believe that the image shows (C) hemorrhage (red arrow). Hemorrhage is the escape of blood from a blood vessel. It can occur due to a variety of factors, including trauma, bleeding disorders, and certain medical conditions.

(...)



## Round 1

### Turn 2

(4) → (3)

: I am an oncologist who specializes in the diagnosis and treatment of cancer.

I believe that the image of Gross Sectioned Myelolipoma demonstrates (C) hemorrhage (red arrow). Hemorrhage is the escape of blood from a blood vessel. It can occur due to a variety of factors, including trauma, bleeding disorders, and certain medical conditions.

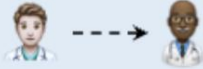
In the image, the red arrow points to an area of the tissue that is dark red and appears to be surrounded by a thin rim of white tissue. This is consistent with the appearance of hemorrhage.

I am confident in my opinion because I have extensive experience in diagnosing and treating cancer. I have seen many cases of hemorrhage, and I am familiar with the appearance of this condition.

I would like to discuss my interpretation of the image with you. I believe that the image shows hemorrhage. I am confident that we can come to a consensus on the diagnosis.

# Case Study (Cont.)

## Step 4. Collaborative Discussion



### Turn 2

(2) → (3)

: I am a radiologist who focuses on the interpretation of medical imaging, such as X-rays, CT scans, MRIs, and ultrasounds. I have extensive experience in diagnosing and treating a wide range of medical conditions using medical imaging.

I believe that the image demonstrates **(A) fibrosis (red arrow)**. Fibrosis is a common finding in myelolipomas, which are benign tumors composed of mature fat cells and hematopoietic tissue. **(C) Hemorrhage**, on the other hand, is less common and typically appears as a dark red or black area on imaging.

I am confident in my interpretation of the image and believe that the surgeon may have misinterpreted the findings.

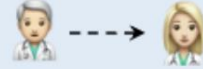


### Turn 2

I do not want to talk to any other experts.

I am confident in my opinion that the image shows **(C) hemorrhage**. I have carefully considered the opinions of the other medical experts, but I believe that the area pointed to by the red arrow is more consistent with the appearance of hemorrhage than fibrosis.

I am open to discussing my opinion further with the other medical experts, but I do not believe that it is necessary to talk to any other experts at this time.



### Turn 3

(4) → (1)

: I am an oncologist who specializes in the diagnosis and treatment of cancer.

I believe that the image of Gross Sectioned Myelolipoma demonstrates **(C) hemorrhage (red arrow)**. Hemorrhage is the escape of blood from a blood vessel.

I understand that you believe the image shows **(A) fibrosis**, but I disagree. I believe that the image shows hemorrhage because the area pointed to by the red arrow is dark red and appears to be surrounded by a thin rim of white tissue. This is consistent with the appearance of hemorrhage, not fibrosis.

I am also aware that you are a specialist in the study and diagnosis of disease through examination of tissue samples, but I believe that my expertise in oncology gives me a unique perspective on this image.

I believe that my argument is strong and that I can convince the pathologist of my opinion. I am confident in my diagnosis of hemorrhage, and I am willing to discuss the image further to support my opinion.

# Case Study (Cont.)

## Step 5. Review and Final Decision

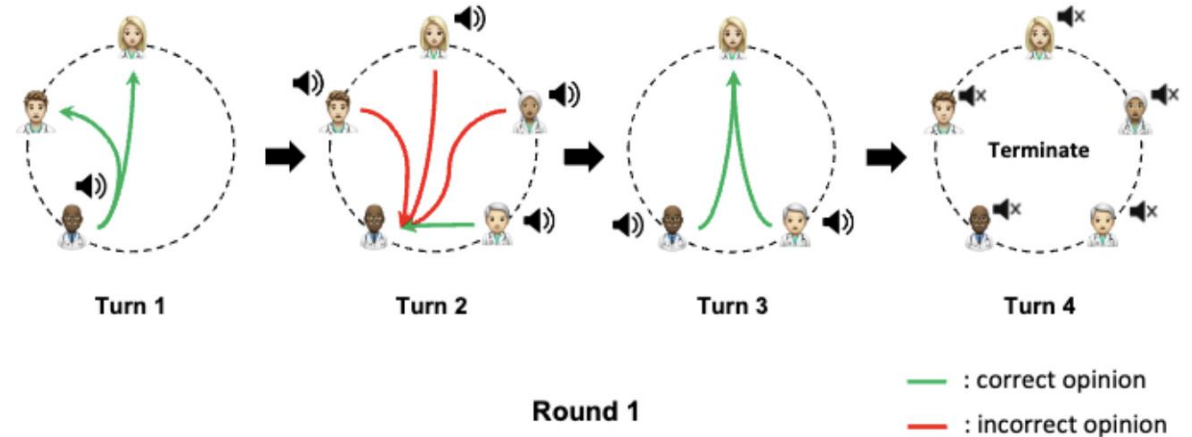


The decision is challenging due to the conflicting interpretations of the image. However, the arguments for hemorrhage by Agent 3 (Surgeon) and Agent 4 (Oncologist) are based on the visual characteristics that are grossly consistent with hemorrhage – specifically, the dark red area suggested by the red arrow. In clinical practice, the appearance of hemorrhage in a gross section is often a dark red coloration, which is indeed what the image seems to display.

While the expertise of Agent 1 (Pathologist) in tissue examination is crucial and their argument for **fibrosis** is noted, the gross pathological features described by Agent 3 (Surgeon) and Agent 4 (Oncologist) seem to align more closely with the typical presentation of a hemorrhage. Furthermore, hemorrhage is a well-documented complication in myelolipomas, particularly larger ones or those subjected to trauma.

Therefore, based on the analysis of the conversation history, the final decision is aligned with interpretation (C) **hemorrhage (red arrow)** as argued by Agent 3 (Surgeon) in Round 1, Turns 1 and 3, and Agent 4 (Oncologist) in Round 1, Turn 2. This conclusion is supported by the gross pathological features consistent with hemorrhage rather than fibrosis, despite the reasoned opinions for fibrosis by Agents 1 (Pathologist), 2 (Radiologist), and 5 (Endocrinologist).

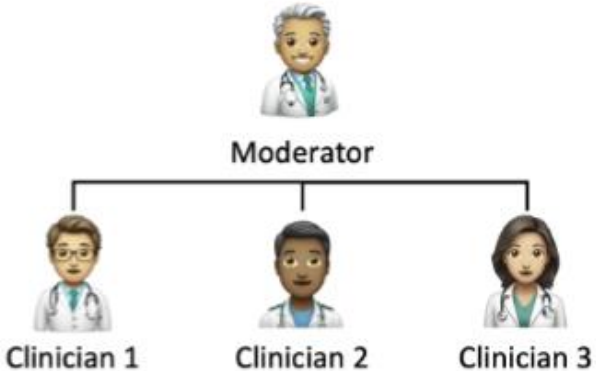
### Interaction Log



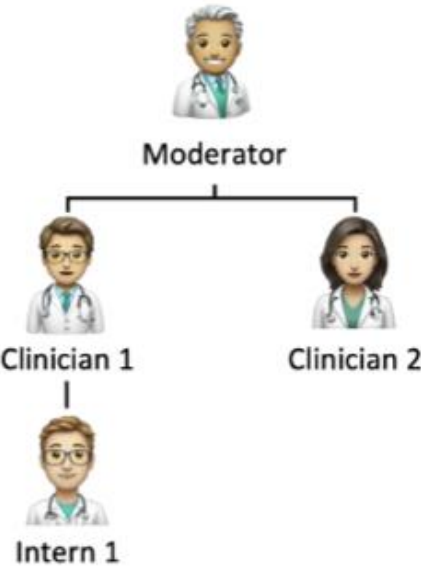
# LLM Recruitment Example



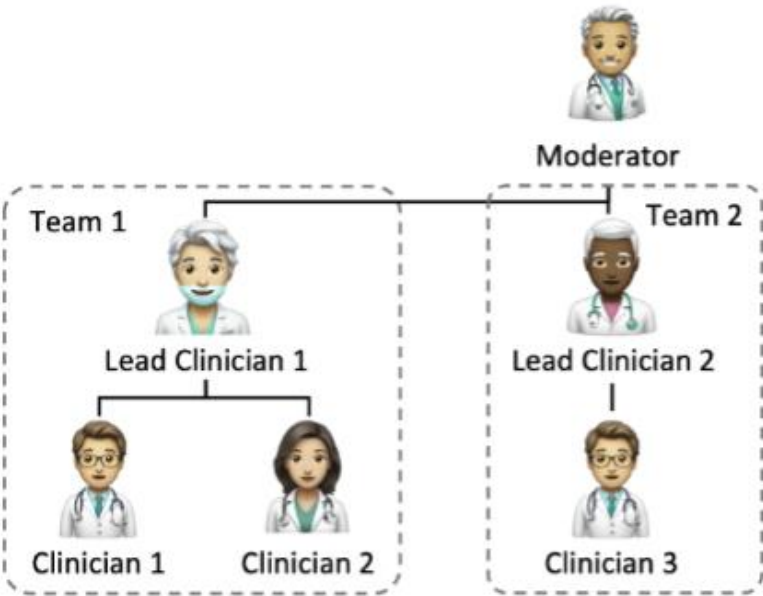
PCP



MDT



MDT w/ hierarchy













ICT

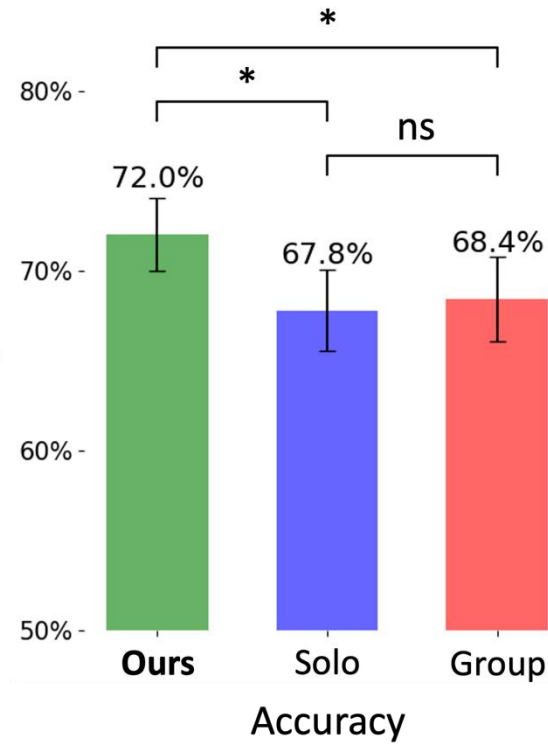
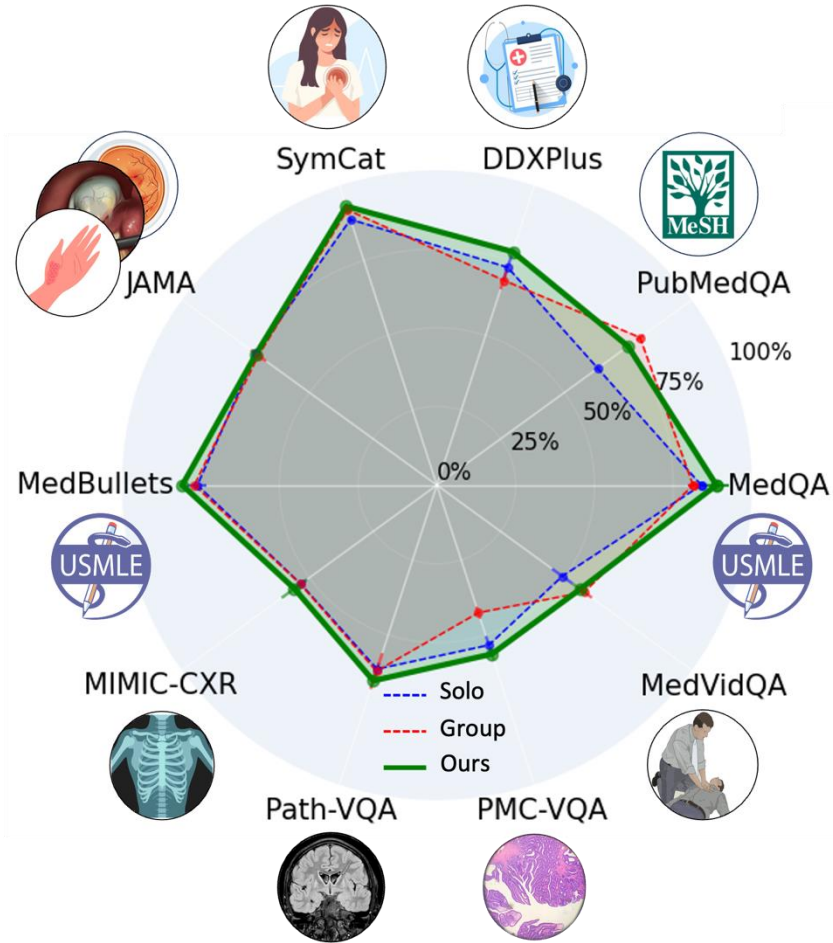


## Datasets

Table 2: Summary of the Datasets. : Text, : Image, : Video. In Appendix [A](#), we provide detailed sample information for each benchmark.

Dataset	Modality	Format	Choice	Testing Size	Domain
MedQA		Question + Answer	A/B/C/D	1273	US Medical Licensing Examination
MedMCQA		Question + Answer	A/B/C/D and Explanations	6.1K	AIIMS and NEET PG entrance exams
PubMedQA		Question + Context + Answer	Yes/No/Maybe	500	PubMed paper abstracts
DDxPlus		Question + Answer	A/B/C/D/ . . .	134 K	Pathologies, Symptoms and Antecedents from Patients
PMC-VQA	 	Question + Answer	A/B/C/D	50 K	VQA pairs across Images, spanning diverse Modalities and Diseases
Path-VQA	 	Question + Answer	Yes/No	3391	Open-ended Questions from Pathology Images
MedVidQA	 	Question + Answer	A/B/C/D	155	First Aids, Medical Emergency, and Medical Education Questions

# Main Experiment Results



- Need more experiments with different seeds to better understand the **trends**
- Currently running experiments with more **benchmarks**

## Ablation 1:

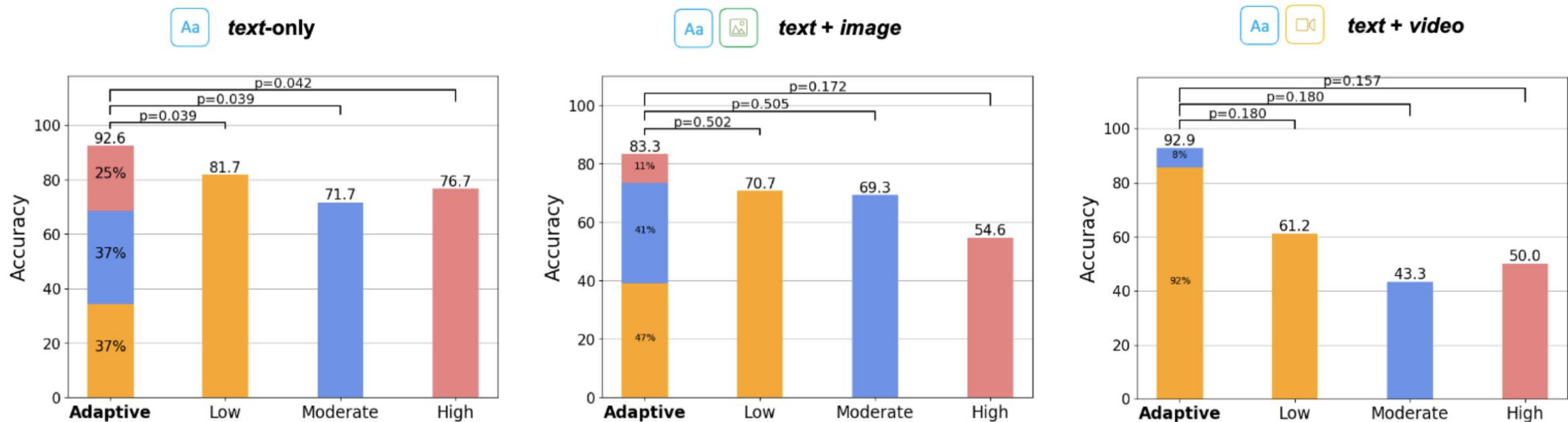


Figure 7: Impact of complexity selection of the query. Accuracy of each ablation on *text-only* (left), *text+image* (center) and *text+video* (right) benchmarks are reported.

## Ablation 2:

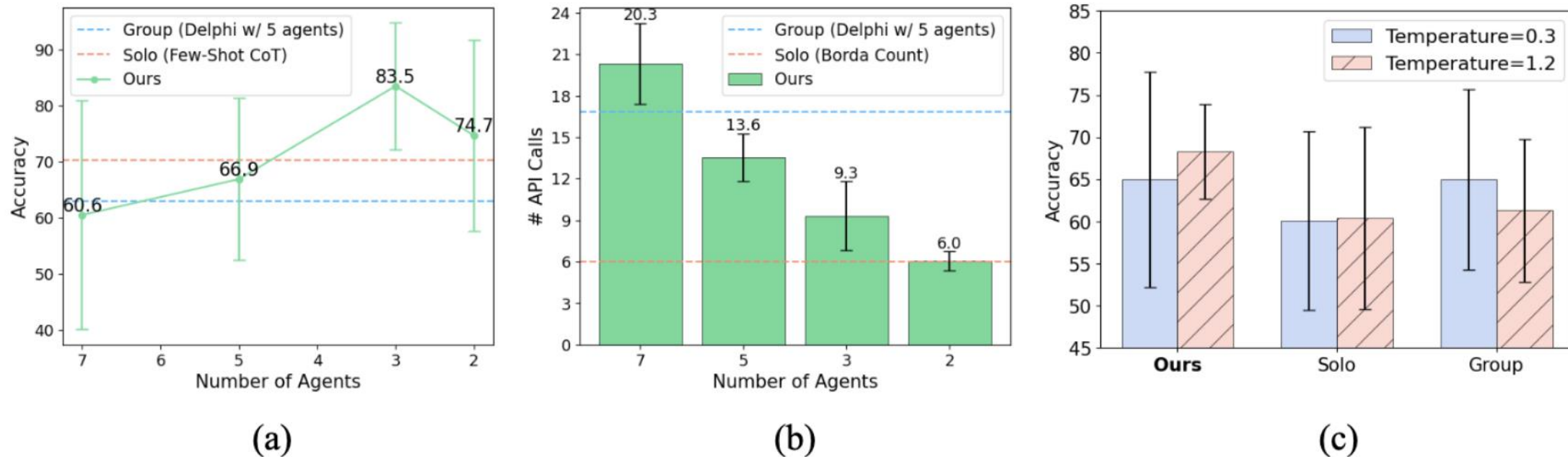


Figure 6: Impact of the number of agents on (a) Accuracy, (b) Number of API Calls on medical benchmarks with GPT-4 (V) and (c) Performance of three different settings under low ( $T=0.3$ ) and high ( $T=1.2$ ) temperatures on medical benchmarks. Our Adaptive setting shows better robustness to different temperatures and even takes advantage of higher temperatures.

### Ablation 3:

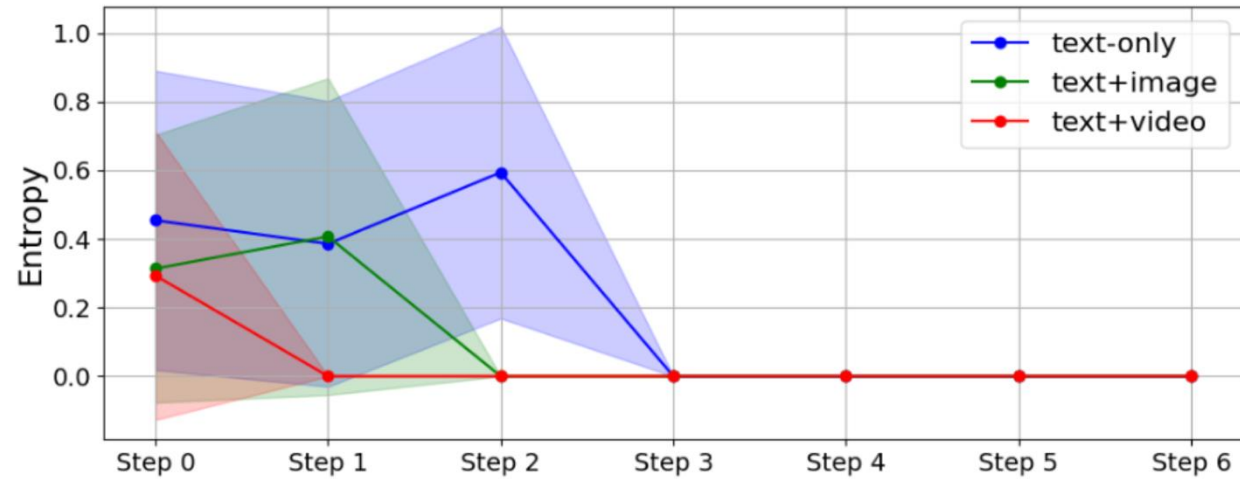


Figure 5: An illustration of consensus entropy in group collaboration process of MDAgents (w/ Gemini-Pro (Vision),  $N=30$  for each dataset) on medical benchmarks with different modality inputs.

*They eventually **converges**...!*

## Complexity Assessment Comparison w/ Human Physicians

**Dataset:** MedQA

**Number of Samples:** N=50

Two among the three physicians had two years of medical training Internal Medicine (Post graduate year 2 (PGY-2) and one among them is a general physician.

# Complexity Assessment Comparison w/ Human Physicians

**Dataset:** MedQA

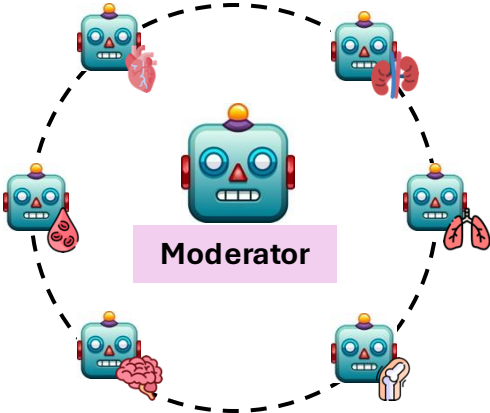
**Number of Samples:** N=50

Two among the three physicians had two years of medical training Internal Medicine (Post graduate year 2 (PGY-2) and one among them is a general physician.

- ICC2k: 0.269
- ICC3k: 0.280

- ➔ Moderate agreement among the raters, highlighting the inherent complexity and subjectivity in evaluating medical questions.
- ➔ The variability in ratings could be attributed to **differences in individual experience, interpretation of the question's context, and the nuances of medical knowledge.**

# Research Direction

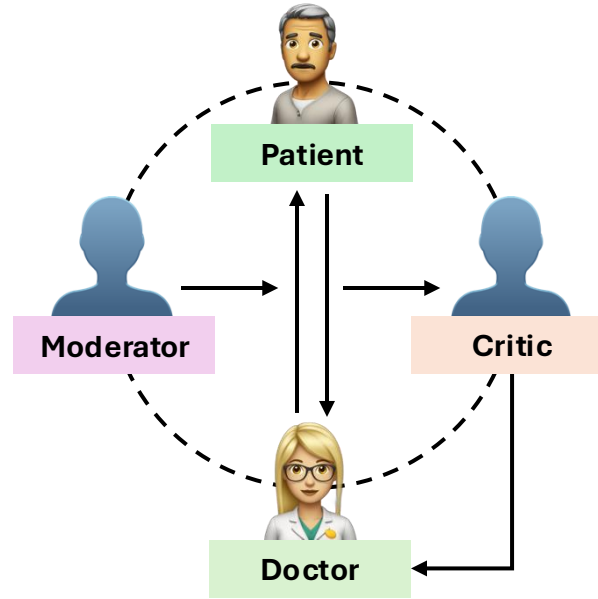
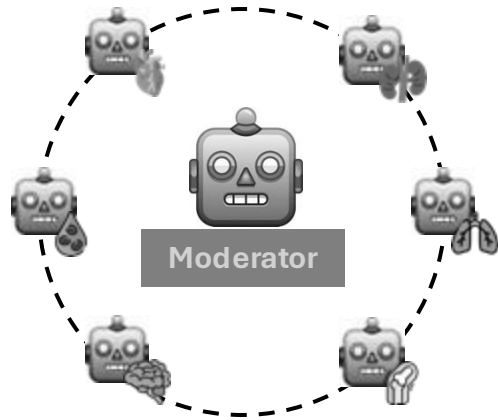


## Multi-LLMs for Decision Support

- Multi-modal Medical Reasoning
- Multi-modal Medical QA
- Diagnostic Support
- ...



# Research Direction



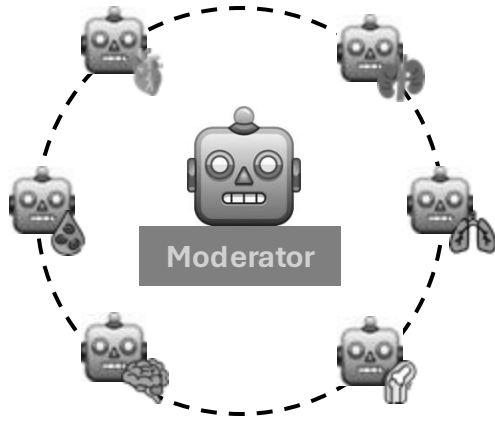
## Multi-LLMs for Decision Support

- Multi-modal Medical Reasoning
- Multi-modal Medical QA
- Diagnostic Support
- ...

## Patient-Doctor Interaction

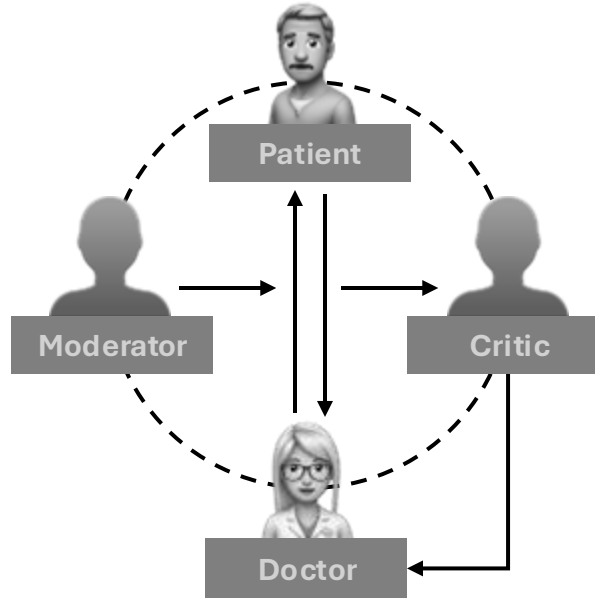
- Medical Reasoning
- Medical QA
- Medical Summarization
- Real-world Dialogue
- Simulated Dialogue

# Research Direction



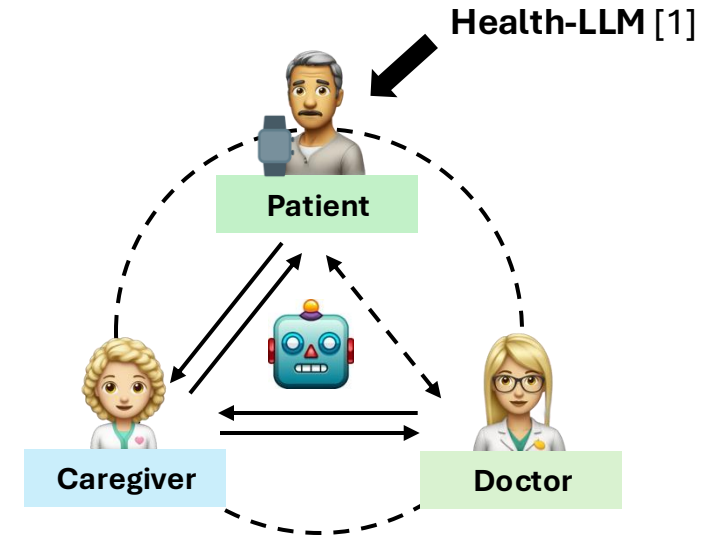
## Multi-LLMs for Decision Support

- Multi-modal Medical Reasoning
- Multi-modal Medical QA
- Diagnostic Support
- ...



## Patient-Doctor Interaction

- Medical Reasoning
- Medical QA
- Medical Summarization
- Real-world Dialogue
- Simulated Dialogue



## Patient-Caregiver-Doctor Interaction

- Medical Summarization (hospitalization/discharge)
- Real-world Dialogue
- Simulated Dialogue
- Personalized Monitoring/Intervention

# Thank you for listening!



Project Page:



Contact: [ybkim95@mit.edu](mailto:ybkim95@mit.edu)