Table 1: Accuracy (%) on Medical benchmarks with **Solo/Group/Adaptive** settings with 100 samples. All benchmarks except for MedVidQA (`Gemini 1.5 Flash`) were evaluated with `GPT-4o mini`.

Category	Method	MedQA	PubMedQA	Path-VQA	PMC-VQA	MedVidQA
		T	T	T I	T I	V T
Single-agent	Zero-shot	75.0	54.0	58.0	48.0	50.0
	Few-shot	77.0	55.0	58.0	50.0	51.0
	+ CoT	78.0	50.0	59.0	52.0	53.0
	+ CoT-SC	79.0	51.0	60.0	53.0	53.0
	ER	76.0	51.0	61.0	51.0	52.0
	Medprompt	79.0	58.0	60.0	<u>54.0</u>	53.0
Multi-agent (Single-model)	Majority Voting	79.0	68.0	<u>63.0</u>	52.0	54.0
	Weighted Voting	80.0	68.0	**64.0**	51.0	<u>55.0</u>
	Borda Count	81.0	69.0	62.0	50.0	52.0
	MedAgents	80.0	69.0	55.0	52.0	50.0
	Meta-Prompting	82.0	69.0	56.0	49.0	-
Multi-agent (Multi-model)	Reconcile	<u>83.0</u>	<u>70.0</u>	58.0	45.0	-
	AutoGen	65.0	63.0	45.0	40.0	-
	DyLAN	68.0	67.0	42.0	48.0	-
Adaptive	**MDAgents (Ours)**	**87.0**	**71.0**	60.0	**55.0**	**56.0**
Category	**Method**	DDXPlus	SymCat	JAMA	MedBullets	MIMIC-CXR
		T	T	T	T	I T
Single-agent	Zero-shot	53.0	84.0	57.0	49.0	38.0
	Few-shot	60.0	<u>87.0</u>	58.0	52.0	33.0
	+ CoT	66.0	84.0	55.0	<u>64.0</u>	33.0
	+ CoT-SC	68.0	84.0	57.0	60.0	40.0
	ER	**76.0**	80.0	56.0	59.0	43.0
	Medprompt	70.0	84.0	**62.0**	60.0	43.0
Multi-agent (Single-model)	Majority Voting	53.0	82.0	56.0	59.0	<u>54.0</u>
	Weighted Voting	52.0	86.0	56.0	56.0	52.0
	Borda Count	53.0	86.0	56.0	59.0	51.0
	MedAgents	56.0	80.9	51.0	58.0	40.9
	Meta-Prompting	53.0	79.0	56.0	51.0	48.0
Multi-agent (Multi-model)	Reconcile	60.0	<u>87.0</u>	59.0	60.0	43.3
	AutoGen	47.0	<u>87.0</u>	53.0	55.0	47.0
	DyLAN	54.0	84.0	55.0	57.0	42.0
Adaptive	**MDAgents (Ours)**	<u>75.0</u>	**89.0**	<u>59.0</u>	**67.0**	**56.0**

Table 2: Accuracy (%) on entire MedQA 5-options dataset with `GPT-4o mini`

Category	Method	Accuracy (%)
Single-agent	Zero-shot	71.5
	3-shot	72.3
	+ CoT	76.6
	+ CoT-SC	77.2
Multi-agent (Multi-model)	Majority Voting	76.3
	Weighted Voting	79.1
	Borda Count	76.1
Multi-agent (Single-model)	Reconcile	80.2
Adaptive	**MDAgents (Ours)**	**83.6**

Table 3: Estimated costs for experimenting with entire test sets with `GPT-4 (Vision)` (in USD)

Method	MedQA	PubMedQA	Path-VQA	PMC-VQA	DDXPlus	SymCat	JAMA	MedBullets	MIMIC-CXR	Total Cost
Zero-Shot CoT	55.24	13.16	3,028.54	27,134.00	16,461.90	10,593.99	134.55	61.23	1,388.70	58,871.29
Ours	172.43	41.36	9,369.45	82,194.34	44,814.97	31,176.05	367.13	161.70	4,406.90	172,704.33

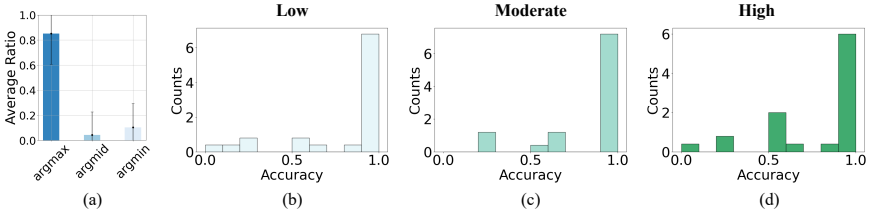

Figure 1: Experiment with the MedQA dataset (N=25 randomly sampled questions). (a) LLM's capability to classify complexity. (b-d) Accuracy distribution for low, moderate, and high complexity levels, with each question solved 10 times.