TeamMedAgents Performance Comparison

Full Evaluation Benchmarks

Benchmark (Dataset)	TeamMedAgents (Adaptive)	TeamMedAgents (Ablation Best*)	MDAgents (Ablation*)
MedBullets	78.80%	74.00%	**80.8 ± 1.7**
PubMedQA	**78.30%**	76.60%	75.0 ± 1.0
MMLU-Pro Med	79.70%	**84.00%**	80.7 ± 2.0*
MedMCQA	**85.40%**	84.80%	80.8 ± 1.1*

1000q Subset Evaluations

Benchmark (Dataset)	TeamMedAgents (Adaptive)	TeamMedAgents (Ablation Best*)	MDAgents (Ablation*)
DDXPlus	74.90%	**81.50%**	77.9 ± 2.1
MedQA	90.70%	**92.60%**	88.7 ± 4.0
Path-VQA	**76.80%**	76.00%	65.3 ± 3.9
PMC-VQA	56.40%	**56.70%**	56.4 ± 4.5

Token Usage & Inference Time (Per Question)

Benchmark	Tokens/Question	Time/Question (s)	API Calls/Question
MedBullets	37,078	23.6	13.6
PubMedQA	34,305	19.4	13.4
MMLU-Pro	28,770	19.7	12.8
MedMCQA	26,812	23.3	12.6
DDXPlus	45,573	26.6	13.9
MedQA	32,444	23.0	13.2
Path-VQA	19,084	14.9	8.6
PMC-VQA	20,770	17.8	9.0

Disagreement Analysis

Benchmark	Total Disagreements	Disagreement Rate	Risk Level
PubMedQA	1	0.1%	Very Low

Benchmark	Total Disagreements	Disagreement Rate	Risk Level
PMC-VQA	151	15.1%	High
Path-VQA	51	5.1%	Medium
MMLU-Pro	2	0.24%	Very Low
MedQA	21	2.1%	Low
MedMCQA	21	2.1%	Low
MedBullets	3	1.0%	Very Low
DDXPlus	~45	4.5%	Medium