Results

ReDel sets the new SotA in 2 of 3 benchmarks!

- We compared three ReDel systems with two baseline single-agent systems
- All experiments were run with zero-shot prompts

System	FanOutQA		TravelPlanner			WebArena		
	Loose	Model Judge	CS-Micro	H-Micro	Final	SR	SR (AC)	SR (UA)
ReDel (GPT-4o)	**0.687**	**0.494**	**67.49**	9.52	**2.78**	0.203	**0.179**	0.643
ReDel (hybrid)	0.551	0.255	65.90	2.14	0.56	0.188	0.171	0.500
ReDel (GPT-3.5-turbo)	0.300	0.087	54.58	0	0	0.092	0.066	0.571
Baseline (GPT-4o)	0.650	0.394	50.83	**18.81**	0	0.162	0.128	**0.786**
Baseline (GPT-3.5-turbo)	0.275	0.077	48.75	0.24	0	0.085	0.058	0.571
Published SotA	0.580	0.365	61.1	15.2	1.11	**0.358**	—	—
Human	0.685	0.452	100	100	100	0.782	0.773	1.000

ReDel outperforms the corresponding single-agent baselines across all benchmarks and improves over published SotA in two of three.

29

Results

ReDel sets the new SotA in 2 of 3 benchmarks!

- We compared three ReDel systems with two baseline single-agent systems
- All experiments were run with zero-shot prompts

System	FanOutQA		TravelPlanner			WebArena		
	Loose	Model Judge	CS-Micro	H-Micro	Final	SR	SR (AC)	SR (UA)
ReDel (GPT-4o)	**0.687**	**0.494**	**67.49**	9.52	**2.78**	0.203	**0.179**	0.643
ReDel (hybrid)	0.551	0.255	65.90	2.14	0.56	0.188	0.171	0.500
ReDel (GPT-3.5-turbo)	0.300	0.087	54.58	0	0	0.092	0.066	0.571
Baseline (GPT-4o)	0.650	0.394	50.83	**18.81**	0	0.162	0.128	**0.786**
Baseline (GPT-3.5-turbo)	0.275	0.077	48.75	0.24	0	0.085	0.058	0.571
Published SotA	0.580	0.365	61.1	15.2	1.11	**0.358**	—	—
Human	0.685	0.452	100	100	100	0.782	0.773	1.000

ReDel outperforms the corresponding single-agent baselines across all benchmarks and improves over published SotA in two of three.

Cost Analysis

More Agents = More Cost?

System	MJ	Δ (↑)	Cost	Δ (↓)
ReDel (4o)	0.494	+25.4%		
RD Short (4o)	0.426	+8.2%		
Base Short (4o)	0.361	-8.2%		
Baseline (4o)	0.394	—	$102.65	—
ReDel (hybrid)	0.255	+229%		
ReDel (3.5-t)	0.087	+12.5%		
Baseline (3.5-t)	0.077	—	$1.24	—