

SCUBA: Salesforce Computer Use Benchmark



Yutong Dai, Krithika Ramakrishnan, Jing Gu, Matthew Fernandez, Yanqi Luo, Viraj Prabhu, Zhenyu Hu, Silvio Savarese, Caiming Xiong, Zeyuan Chen, Ran Xu
Salesforce AI Research Project page: <https://sfrcua.github.io/SCUBA>

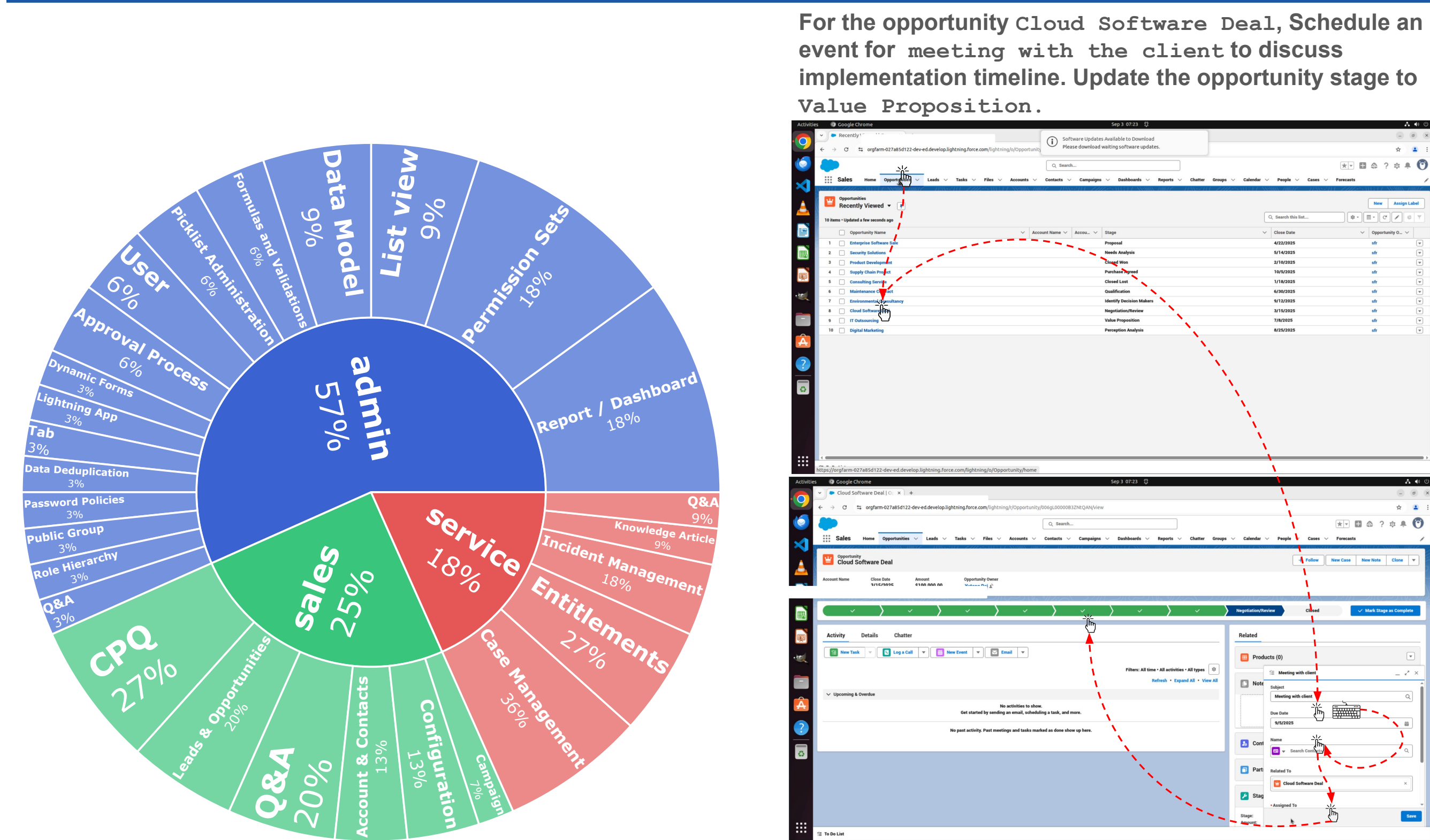
Motivation

Existing benchmarks (WebArena, OSWorld, WorkArena) do not capture the complexities of **enterprise-grade platforms**. The CRMarena series focus on tool-use/information retrieval and lack **state-modifying tasks**. We introduce SCUBA – a realistic benchmark for evaluating computer-use agents on CRM workflows within the Salesforce platform.

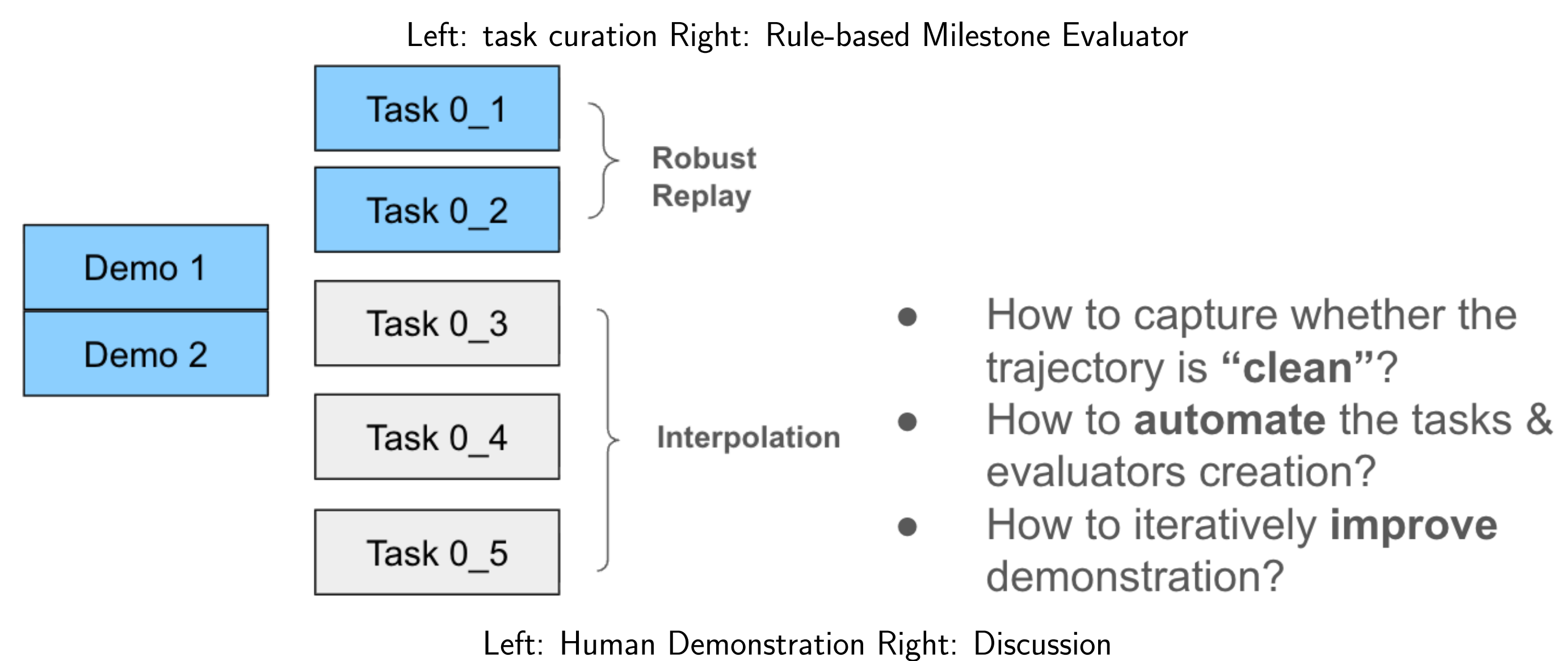
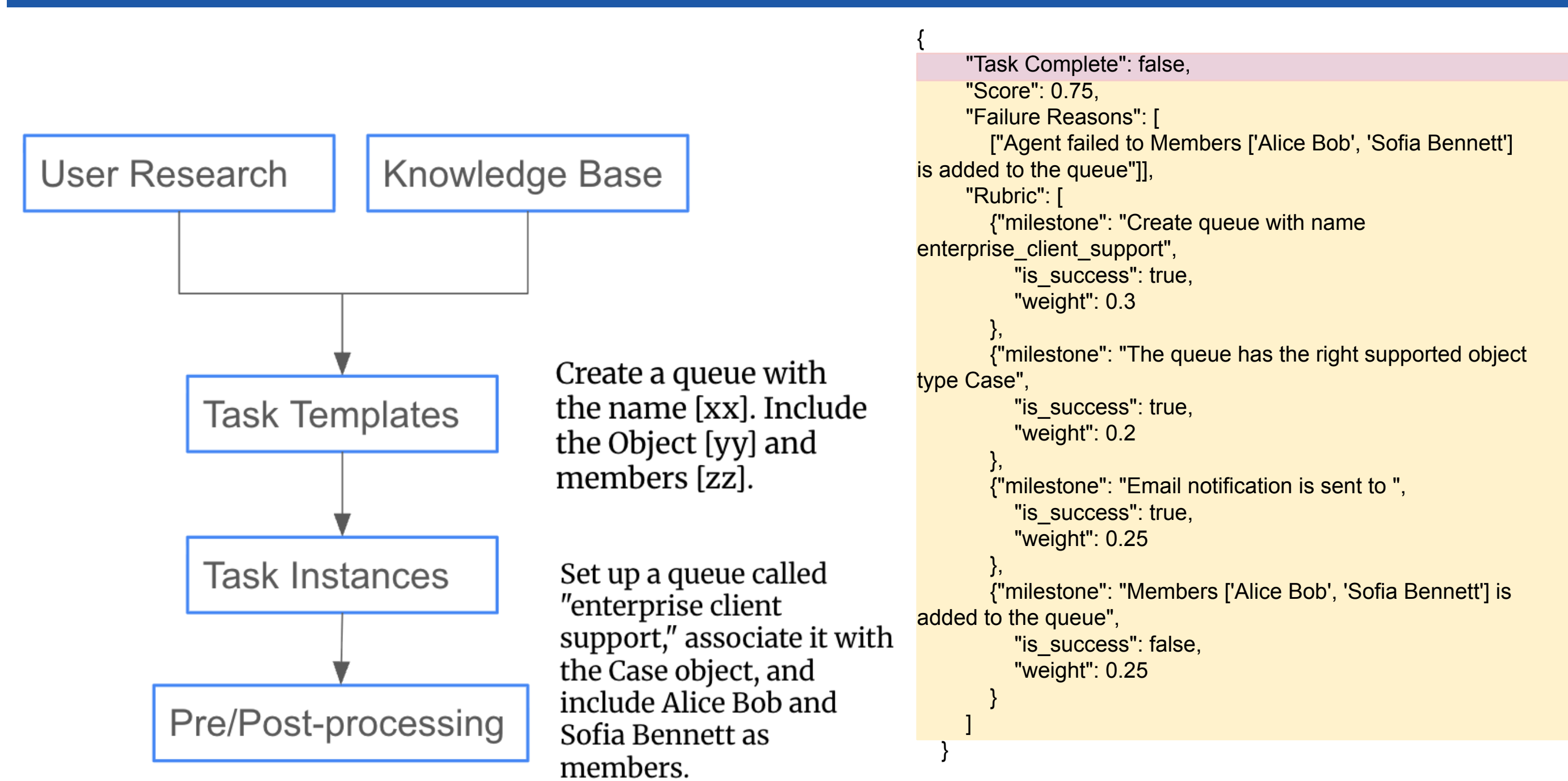
Key Contributions

300 realistic tasks from user interviews · **Milestone scores** + latency + cost metrics · **Human demos** that boost performance · **Async pipeline** (full eval <90 min)

Task Distribution & Trajectory



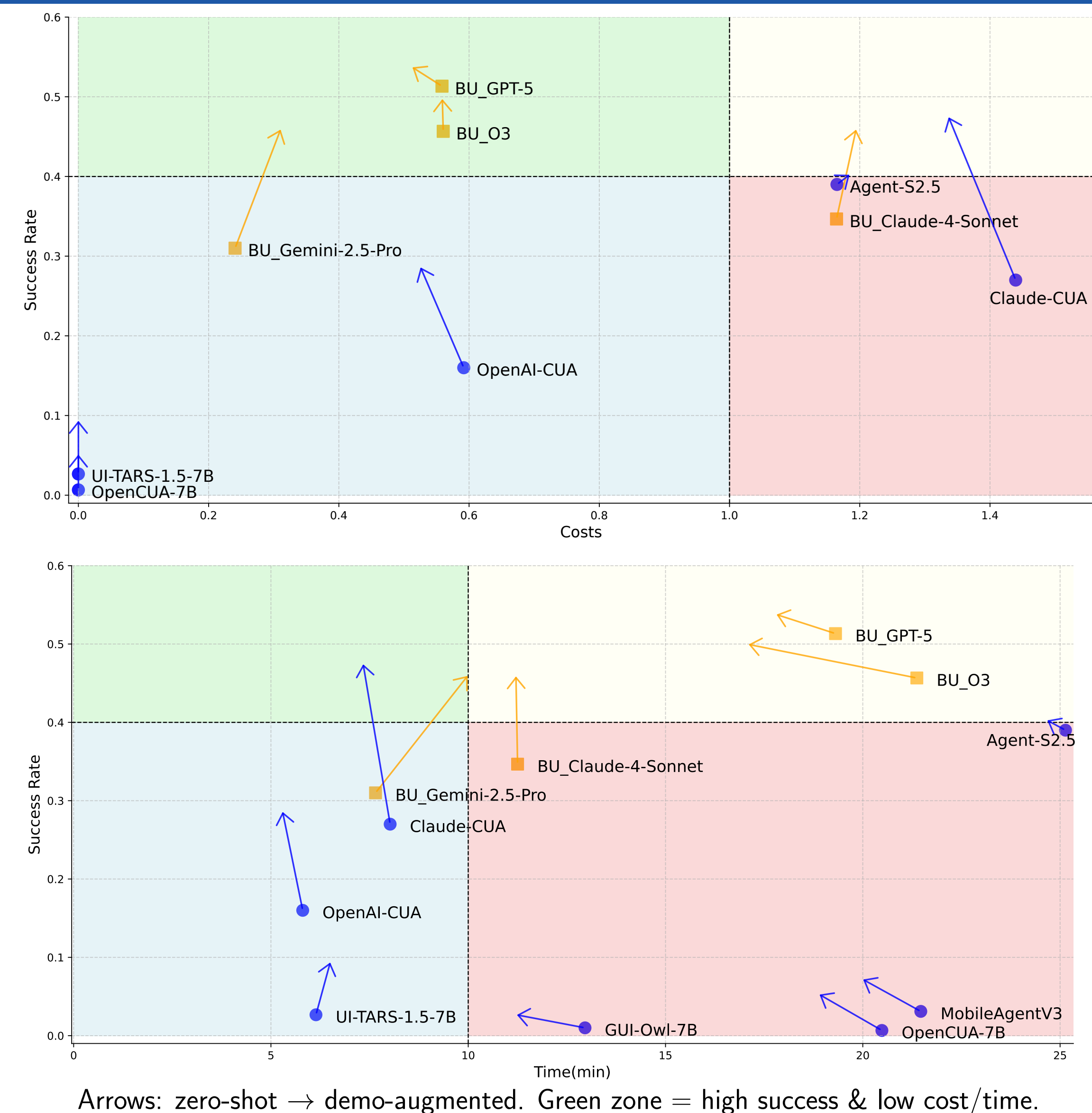
Benchmark Design



Main Results

Method	Milestone ↑	Success Rate ↑	Time (min) ↓	Steps ↓	Tokens (k) ↓	Cost (\$) ↓
Zero-shot Setting						
Browser-Use Agents						
GPT-5	0.73	51.33%	19.31	22.9	245.24	0.55
o3	0.65	45.67%	21.37	26.74	237.16	0.56
Gemini-2.5-Pro	0.47	31.00%	7.27	15.61	134.94	0.24
Claude-4-sonnet	0.56	34.67%	11.25	24.12	353.47	1.16
Computer-Use Agents						
UI-TARS-1.5-7B	0.10	2.67%	6.14	35.43	183.69	–
OpenCUA-7B	0.05	0.67%	20.48	31.21	133.88	–
GUI-Owl-7B	0.05	1.00%	12.96	34.05	167.64	–
OpenAI-CUA	0.29	16.00%	5.80	31.32	193.83	0.59
Claude-4-sonnet (comp.)	0.48	27.00%	8.02	35.24	457.77	1.44
MobileAgentV3	0.11	3.10%	21.47	41.96	755.63	1.17
Agent-S2.5(w/GPT-5)	0.58	39.00%	25.13	34.23	597.92	1.17
Demonstration-Augmented Setting						
Browser-Use Agents						
GPT-5	0.75	53.85%	17.76	21.33	232.76	0.51
o3	0.69	50.00%	17.05	25.33	236.63	0.55
Gemini-2.5-Pro	0.71	46.15%	10.02	19.48	177.13	0.31
Claude-4-sonnet	0.68	46.15%	11.21	23.92	364.10	1.19
Computer-Use Agents						
UI-TARS-1.5-7B	0.24	9.16%	6.52	37.65	213.07	–
OpenCUA-7B	0.12	5.38%	18.86	33.57	159.11	–
GUI-Owl-7B	0.12	2.70%	11.18	36.81	181.50	–
OpenAI-CUA	0.50	28.85%	5.29	27.29	171.93	0.52
Claude-4-sonnet (comp.)	0.67	47.69%	7.33	32.05	424.90	1.34
MobileAgentV3	0.18	7.30%	19.96	42.98	830.85	1.17
Agent-S2.5(w/GPT-5)	0.57	40.38%	24.63	34.43	626.26	1.19

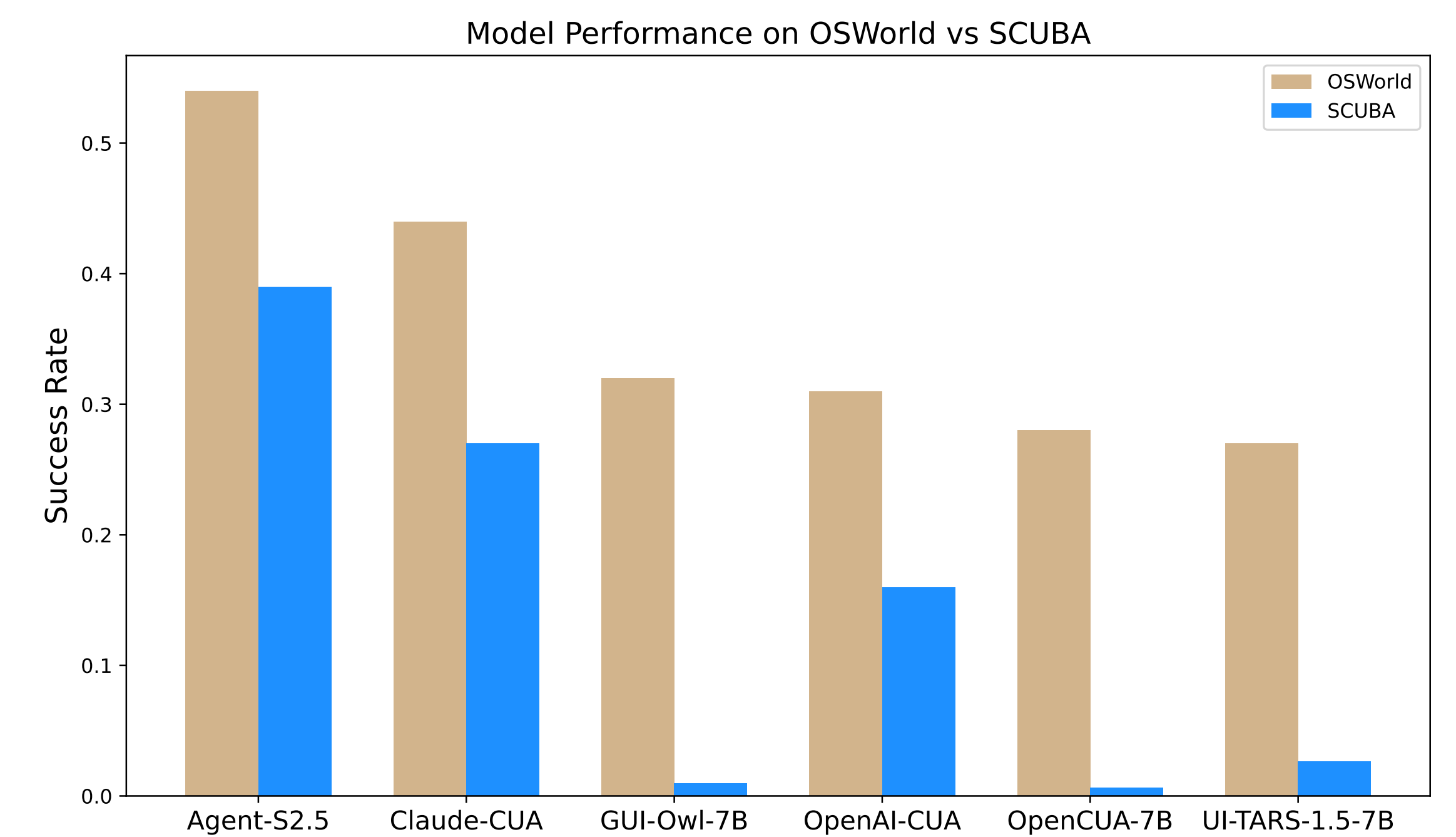
Cost & Time vs. Success Rate



Key Findings

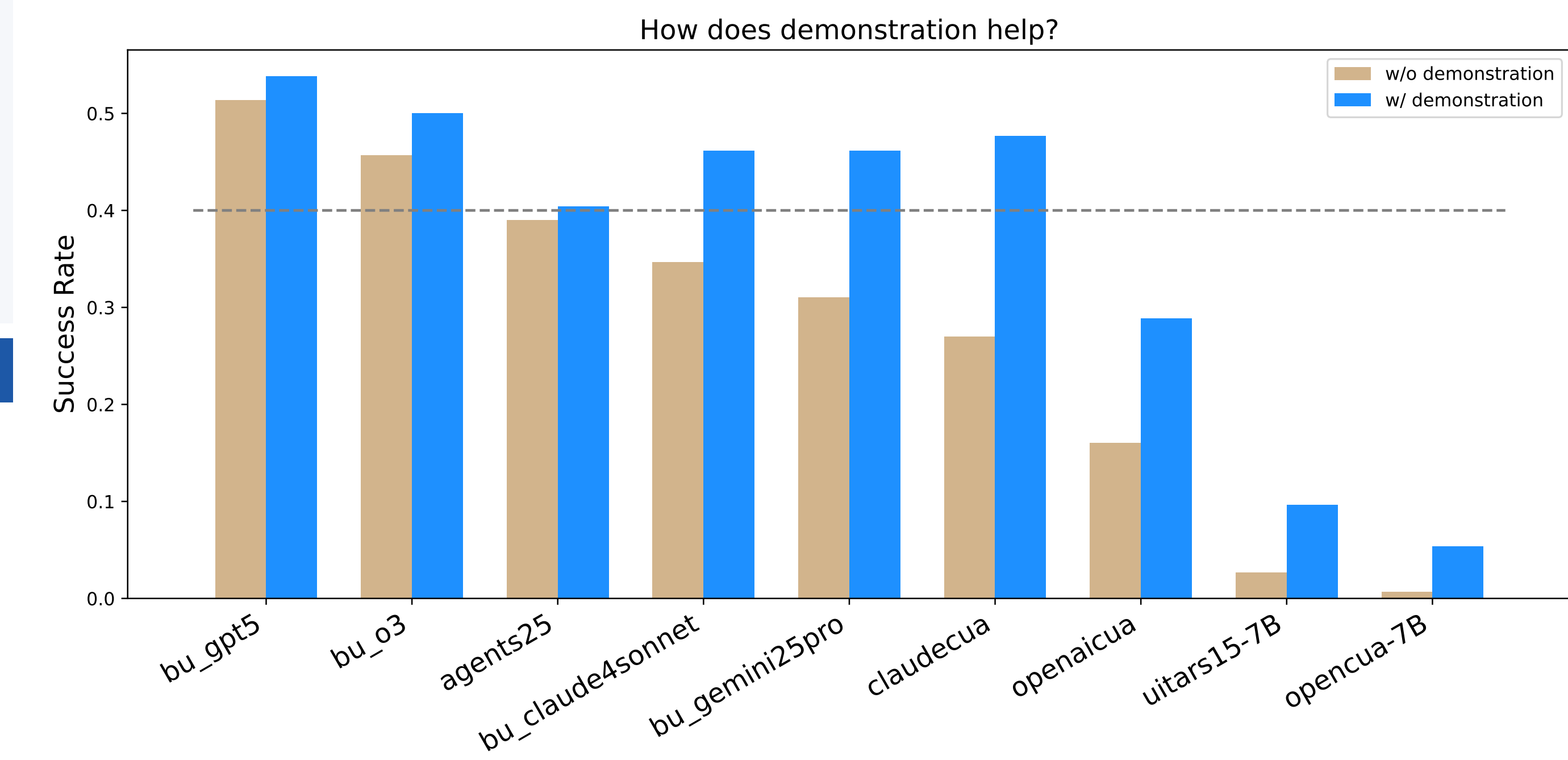
- **Open- vs. closed-source gap:** Open-source CUAs fall behind closed-source CUAs.
- **Browser-use > computer-use** in success rate (richer observation + stronger planning)
- **Computer-use = lower latency** but lower success – useful for latency-sensitive settings.
- **Domain Gap:** work well on one environment does not mean on the other.
- **Demos are effective:** ↑ 50% success rate, ↓13% time, ↓16% cost.

Generalization Gap: OSWorld → SCUBA



Significant drop from OSWorld to SCUBA. Main failure causes: **planning** and **grounding**.

Effectiveness of Demonstration



Sample demonstration.

- **TASK:**Create a Queue with the name "Shoe Case Support". Only send email to members under the distribution list "support.shoe@papaltd.com". Add the Object "Customer". Include "Role: Customer Support, North America" as the Member.
- **Experience:**
 - Log in and open Setup → Navigate to Users > Queues. → Click New Queue.
 - Configure the queue: Queue Label: Shoe Case Support. → Queue Email: support.shoe@papaltd.com.
 - Add supported object: Select Customer and click Add to Selected Objects.
 - Add members: In Queue Members, choose Role, select Role: Customer Support, North America, and click Add to Selected Members.
 - Save the queue.

One more thing ...

- Stay tuned for SCUBA-Pro
- Cross-platforms to coveri common Sales workflows
- Containerized Environment for reproducibility