Magentic-One & Magentic-UI

From Generalist Multi-Agent Systems to Human-in-the-Loop Agents

Yikun Han

October 2, 2025

ScienceNLP Lab Meeting

Magentic-One

Motivation: Agentic Systems Need Both Reasoning & Action

- Modern LLMs can reason; real-world tasks require **tools, browsing, files, code**.
- Challenges: multi-step planning, recovery from errors, long-horizon dependencies.
- **Goal**: A *generalist* system that works across domains without per-task tuning.

Background: AutoGen

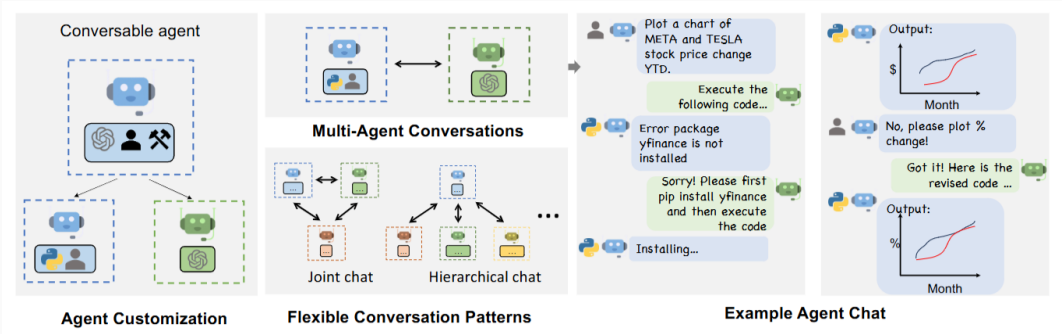


Figure 1: AutoGen enables diverse LLM-based applications using multi-agent conversations. (Left) AutoGen agents are conversable, customizable, and can be based on LLMs, tools, humans, or even a combination of them. (Top-middle) Agents can converse to solve tasks. (Right) They can form a chat, potentially with humans in the loop. (Bottom-middle) The framework supports flexible conversation patterns.

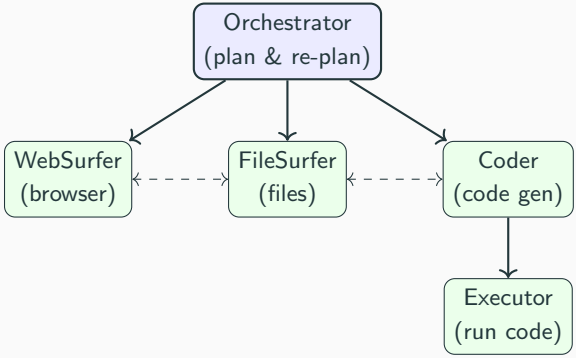
Summary

A generalist multi-agent system: an **Orchestrator** plans/re-plans and coordinates specialized agents (*WebSurfer*, *FileSurfer*, *Coder*, *Executor/Terminal*) to solve complex, open-ended tasks involving the web, files, and code.

- **Modular**: add/remove agents without prompt retuning.
- **Competitive**: GAIA, AssistantBench, WebArena.
- **Practicality**: released with *AutoGenBench* for rigorous evaluation.

System Components (High Level)

- **Orchestrator**: global planner, state tracker, error recovery.
- **WebSurfer**: navigate/search/click/extract/summarize the web.
- **FileSurfer**: read files, browse folders, synthesize from docs.
- **Coder**: write/analyze code; generate artifacts.
- **Executor/Terminal**: run code, manage env/deps.



System Components (Illustration)

Figure 2: An illustration of the Magentic-One mutli-agent team completing a complex task from the GAIA benchmark. Magentic-One's Orchestrator agent creates a plan, delegates tasks to other agents, and tracks progress towards the goal, dynamically revising the plan as needed.

Orchestrator Agent

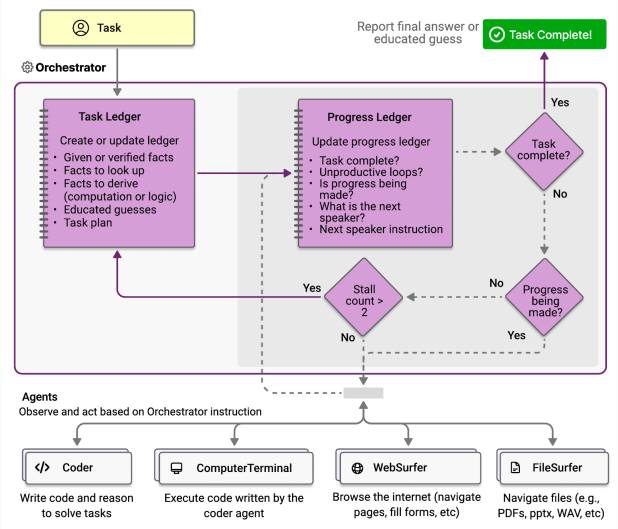


Figure 3: The Orchestrator agent that implements two loops: an outer loop and an inner loop.

Design Principles

- **Generalist by default**: no per-task wiring for common domains.
- **Modular**: plug-in agents; minimal coupling.
- **Evaluable**: AutoGenBench with repetition/isolation controls.
- **Safety-aware**: least privilege, logging, containment.

Orchestration & Collaboration Patterns

- **Plan** → **Act** → **Observe** → **Re-plan** loop.
- **Hand-offs**: outputs from WebSurfer → Coder; artifacts → FileSurfer.
- **Retry & Recovery**: detect stalls, adjust subgoals, switch agents.
- **State**: scratchpad, partial results, tool provenance.

Benchmarks & Result

Dataset	Category	Magentic-One (GPT-4o)	Magentic-One (GPT-4o, o1)	Best Baseline [75, 71]
GAIA [29]	Level 1	46.24	**54.84**	53.76
	Level 2	28.3	32.7	**37.11**
	Level 3	18.75	22.92	**26.53**
AssistantBench [71]	Easy	69.9	73.4	**81**
	Medium	35.6	**47.1**	44.6
	Hard	**16.9**	14.8	13.3
WebArena [79]	Reddit	53.77	–	**65.1**
	Shopping	33.16	–	**36.9**
	CMS	**29.1**	–	24.7
	Gitlab	27.78	–	**39.4**
	Maps	**34.86**	–	33.9
	Cross Site	14.6	–	–

Figure 4: Performance comparison between Magentic-One (GPT-4o), Magentic-One (GPT-4o, o1) and the best baseline for each benchmark's test set.

Benchmarks & Result

Dataset	Category	Magentic-One (GPT-4o)	Magentic-One (GPT-4o, o1)	Best Baseline [75, 71]
GAIA [29]	Level 1	46.24	**54.84**	53.76
	Level 2	28.3	32.7	**37.11**
	Level 3	18.75	22.92	**26.53**
AssistantBench [71]	Easy	69.9	73.4	**81**
	Medium	35.6	**47.1**	44.6
	Hard	**16.9**	14.8	13.3
WebArena [79]	Reddit	53.77	–	**65.1**
	Shopping	33.16	–	**36.9**
	CMS	**29.1**	–	24.7
	Gitlab	27.78	–	**39.4**
	Maps	**34.86**	–	33.9
	Cross Site	14.6	–	–

Of the cities within the United States where U.S. presidents were born, which two are the farthest apart from the westernmost to the easternmost going east, giving the city names only? Give them to me in alphabetical order, in a comma-separated list.

Which supermarkets within 2 blocks of Lincoln Park in Chicago have ready-to-eat salad for under $15?

Tell me the count of comments that have received more downvotes than upvotes for the user who made the latest post on the Showerthoughts forum.

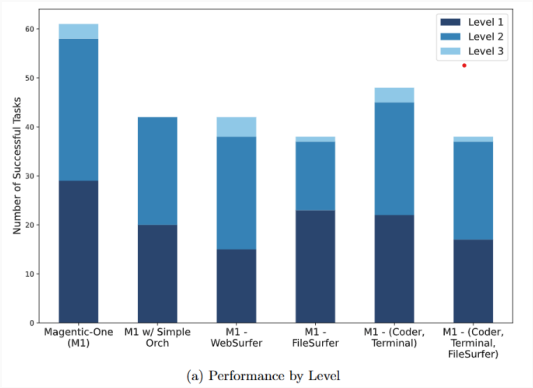
Figure 5: Performance comparison between Magentic-One (GPT-4o), Magentic-One (GPT-4o, o1) and the best baseline for each benchmark's test set.

Benchmarks & Result

Method	GAIA	AssistantBench (EM)	AssistantBench (accuracy)	WebArena
omne v0.1 (GPT-4o, o1)	40.53±5.6	–	–	–
Trase Agent v0.2 (GPT-4o, o1, Gemini)	39.53±5.5	–	–	–
Multi Agent (NA)	38.87±5.5	–	–	–
das agent v0.4 (GPT-4o)	38.21±5.5	–	–	–
Sibyl (GPT-4o) [56]	34.55±5.4	–	–	–
HF Agents (GPT-4o)	33.33±5.3	–	–	–
FRIDAY (GPT-4T) [61]	24.25±4.8	–	–	–
GPT-4 + plugins [29]	14.60±4.0	–	–	–
SPA → CB (Claude) [71]	–	13.8±5.0	26.4±6.4	–
SPA → CB (GPT-4T) [71]	–	9.9±4.3	25.2±6.3	–
Infogent (GPT-4o)	–	5.5±3.3	14.5±5.1	–
Jace.AI (NA)	–	–	–	**57.1±3.4**
WebPilot (GPT-4o) [75]	–	–	–	**37.2±3.3**
AWM (GPT-4) [57]	–	–	–	35.5±3.3
SteP (GPT-4) [49]	–	–	–	33.5±3.2
BrowserGym (GPT-4o) [10]	–	–	–	23.5±2.9
GPT-4	6.67±2.8[29]	6.1 ±3.5[71]	16.5 ±5.4[71]	14.9±2.4[79]
Human	92.00±3.1	–	–	78.2±2.8
Magentic-One (GPT-4o)	32.33±5.3	11.0 ±4.6	25.3 ±6.3	32.8±3.2
Magentic-One (GPT-4o, o1)	38.00±5.5	13.3 ±4.9	27.7 ±6.5	*

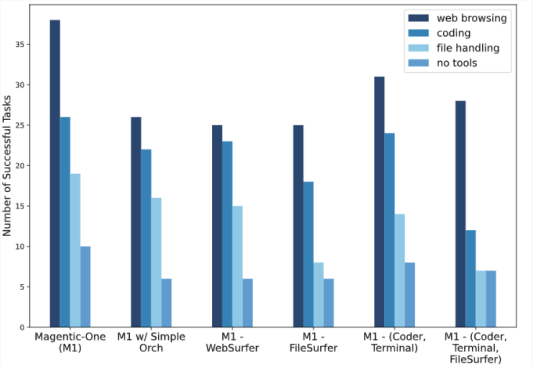
Figure 6: Performance of Magentic-One compared to relevant baselines on the test sets of GAIA, WebArena and AssistantBench.

Ablation Studies



(a) Performance by Level

Figure 7: The performance of different ablations of Magentic-One on the GAIA validation set broken down by difficulty level.



(b) Performance by Capabilities

Figure 8: Ablation results broken down by required capabilities.

Error Analysis

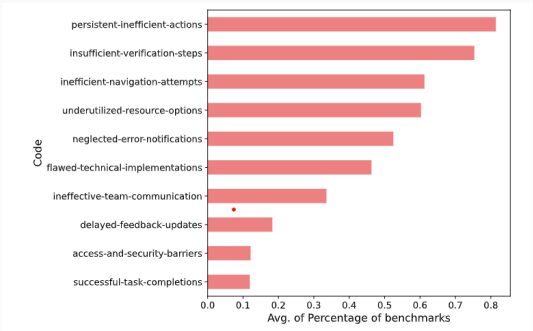


Figure 9: Distribution of error codes obtained by the automated analysis of MagenticOne's behavior as observed in the logs of the validation examples across all benchmarks studied.

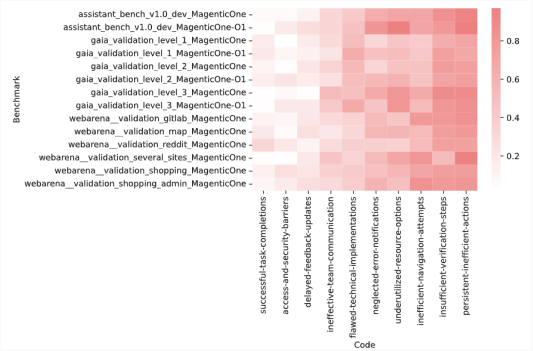


Figure 10: Heatmap of the error codes obtained by the automated analysis of Magentic-One's behavior as observed in the logs of the validation examples across all benchmarks studied.

Capabilities in Practice

- **Web**: search, navigate, extract tables, follow links, download.
- **Files**: read PDFs/CSVs/Markdown; summarize & synthesize.
- **Code**: prototype scripts, data processing, light automation.
- **Long tasks**: decompose into subgoals; checkpoint progress.

Failure Modes & Limitations

- **Misplanning / Drift**: wrong subgoals; fix with re-planning.
- **Tool brittleness**: dynamic web UIs, auth walls.
- **Hallucinated affordances**: attempting actions not supported.
- **Safety & Security**: unintended side effects; require sandboxing & oversight.

Safety & Governance Practices (Recommended)

- **Least privilege**: scoped credentials, read-only by default.
- **Containment**: containers/VMs for Executor tasks.
- **Auditability**: detailed logs; replay traces.
- **Guardrails**: domain whitelists, action confirmations on risky ops.

Extensibility & Tooling

- Implemented atop **AutoGen**; installable as `autogen-ext[magentic-one]`.
- Add new agents (APIs, domain tools) with minimal prompt surgery.
- Evaluate with **AutoGenBench**; isolate side effects; repeated trials.

Magentic-UI

Why Human-in-the-Loop?

- Autonomy helps, but **humans still better** at judgment, risk checks.
- HITL can ↑reliability, ↑safety, and ↑sample efficiency.
- **Magentic-UI**: a web UI + multi-agent backend that operationalizes HITL.

What is Magentic-UI?

- Open-source research prototype.
- Built on a flexible multi-agent stack **derived from Magentic-One**.
- Supports web browsing, file ops, code gen/exec; extensible via **MCP**.

What is Magentic-UI?

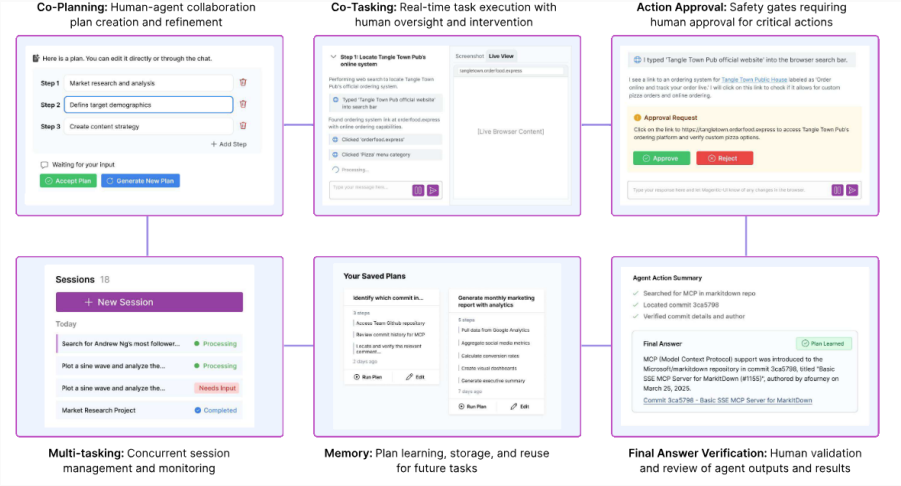


Figure 11: agentic-UI is an open-source research prototype of a human-centered agent that is meant to help researchers study open questions on human-in-the-loop approaches and oversight.

HITL Mechanisms in Magentic-UI

- **Co-planning**: user edits the plan before execution.
- **Co-tasking & multi-tasking**: parallel goals with human arbitration.
- **Action guards**: confirm/deny sensitive actions.
- **Long-term memory**: reuse context across sessions.

Building on Magentic-One

- Reuses the **Orchestrator + Specialists** pattern.
- Adds UI hooks for plan steps, tool outputs, and confirmations.
- Emphasizes **transparency**: show actions, states, artifacts.

When to Prefer HITL

- High-risk actions (payments, emails, system changes).
- Ambiguous specs or noisy environments.
- Novel tasks where autonomy struggles; HITL $\Rightarrow$ faster iteration.

Discussion

Key Takeaways

- **Magentic-One**: generalist, modular, evaluated, practical.
- **Magentic-UI**: human-centered controls on top of the same stack.
- For research groups, combine autonomy with **HITL** to move fast safely.

References

References i

A. Fourney et al.
Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks.
arXiv:2411.04468, 2024.
https://arxiv.org/abs/2411.04468

Microsoft Research Blog.
Magentic-One: A Generalist Multi-Agent System for Solving Complex Tasks.
2024.
https://www.microsoft.com/en-us/research/articles/magentic-one-a-generalist-multi-agent-system-for-solving-complex-tasks/

References ii

AutoGen Docs.
Magentic-One user guide & API.
https://microsoft.github.io/autogen/stable//user-guide/agentchat-user-guide/magentic-one.html

H. Mozannar et al.
Magentic-UI: Towards Human-in-the-loop Agentic Systems.
arXiv:2507.22358, 2025.
https://arxiv.org/abs/2507.22358

References iii

Microsoft Research Blog.
Magentic-UI, an experimental human-centered web agent.
2025.
https://www.microsoft.com/en-us/research/blog/magentic-ui-an-experimental-human-centered-web-agent/

Microsoft GitHub.
Magentic-UI repository.
https://github.com/microsoft/magentic-ui