fellowship.ai

Final Demo: Autonomous Web Agents

28.03.2025

Overview

- Comparison of different agents

  - UI-TARs

  - Midscene

  - Proxy-lite

- Conclusion & Recommendation

Comparison of different agents

Starting point: The 32nd cohort's prototype

Problems: The prototype turned out both computationally expensive and ineffective.

We therefore experimented with several promising **open-source alternatives**:

- The UI-TARS model implemented in the Midscene.js chrome extension
- The UI-TARS model implemented in the UI-TARS Desktop agent
- The proxy-lite web agent

The 32nd cohort's agent

There were two types of benchmark tests carried out for the previous cohort's agent:

- Evaluation of the agent's ability to choose the correct actions in individual steps.
- Evaluation of the agent's finish and success rates in entire runs.

The 32nd cohort's agent

Evaluation of the agent's ability to choose the correct actions in individual steps:

Action	TP + FN	TP + FP	TP	Recall	Precision	F1-Score
Click	20	25	14	0.7000	0.5600	0.6222
Type	11	10	5	0.4545	0.5000	0.4761
Scroll	3	0	0	0.0000	0.0000	0.0000
Return Value	6	1	1	0.1667	1.0000	0.2858

Total predictions	Correct predictions	Wrong predictions	Accuracy
40	20	20	0.5000

The 32nd cohort's agent

Evaluation of the agent's finish and success rates in entire runs on allrecipes.com:

Finish rate: 20.00%
Success rate: 50.00%

Additional problems:

- High execution time
- The test crashed due to high computational costs

Prompt	Finish?	Success?	Steps
Provide a recipe for vegetarian lasagna with more than 100 reviews and a rating of at least 4.5 stars suitable for 6 people.	Yes	Yes	10
Find a recipe for a vegetarian lasagna that has at least a four-star rating and uses zucchini.	No	No	13
Find a recipe for a vegetarian lasagna under 600 calories per serving that has a prep time of less than 1 hour.	No	No	7
Locate a recipe for vegan chocolate chip cookies with over 60 reviews and a rating of at least 4.5 stars on Allrecipes.	No	No	4
Find a recipe for Baked Salmon that takes less than 30 minutes to prepare and has at least a 4 star rating based on user reviews.	No	No	3
Search for a popular Pasta Sauce with more than 1000 reviews and a rating above 4 stars. Create a shopping list of ingredients for this recipe.	No	No	4
Search for a vegetarian lasagna recipe that has at least a four-star rating and over 500 reviews.	No	No	1
Find a popular recipe for a chocolate chip cookie and list the ingredients and preparation steps.	No	No	1
Search for a recipe for Beef Wellington on Allrecipes that has at least 200 reviews and an average rating of 4.5 stars or higher. List the main ingredients required for the dish.	Yes	No	14
Find a high-rated recipe for vegetarian lasagna, list the key ingredients required, and include the total preparation and cook time stated on the recipe.	No	No	8

UI-TARS

- A **model** and accompanying **agent framework** developed by ByteDance Research
- The model comes in 3 different sizes: **2b**, **7b** and **72b**

We experimented with the 2b and 7b models in two different frameworks:

1. The **Midscene.js** agent
2. The **UI-TARS Desktop** agent

The 7b model only ran under usage of Hugging Face endpoints, for the 2b model either Hugging Face endpoints or a GPU was required.

Midscene.js

Midscene.js is a **Chrome extension** serving as an autonomous web agents.

The extension does not come with an inherent model; we ran it with the **UI-TARS** model.

The agent showed severe problems:

- **Excessive token consumption:** Runs frequently resulted in a ”Planning422 (Token Limit Exceeded)” error.
- **Incomplete outputs:** Token restrictions often led to truncated or partial data retrieval.

UI-TARS Desktop agent

The agent showed strong benchmarking results in a study carried out by ByteDance (Source: arxiv):

- **Element accuracy** (72b): 74.7
- **Operation F1-Score** (72b): 92.5
- **Step success rate** (72b): 68.6

These strong scores were backed up by the group's experimental runs for example in tasks of extracting equated monthly instalment (EMI) data from bank webpages.

UI-TARS Desktop agent

Evaluation the agent's math, summarization and general reasoning skills:

Model	GSM8K	CNN Dailymail	HellaSwag
UI-TARS-7b DPO	Accuracy: 0.175	Rouge1: 0.142, Rouge2: 0.078, RougeL: 0.104, RougeLsum: 0.117	Accuracy: 0.265

Evaluating the agent's visual question answering skills on the ScienceQA dataset:

Model	Accuracy
UI-TARS-7b DPO	0.4

The proxy-lite web agent

- A **model** and **accompanying** agent framework developed by Convergence.
- The model (3.75b parameters) runs locally, requiring neither a **GPU** nor **Hugging Face endpoints.**
- The agent does not rely on **OCR**, but uses JavaScript for web navigation and scraping.

The proxy-lite web agent

Evaluation of the agent's finish and success rates in entire runs on the WebVoyager dataset:

Source: Hugging Face

Finish rate: 72.19%
Success rate: 86.55%

32nd cohort's agent:
Finish rate: 20.00%
Success rate: 50.00%

Webpage	Finish rate	Success rate	Avg. steps
Allrecipes	87.8%	95.1%	10.3
Amazon	70.0%	90.0%	7.1
Apple	82.1%	89.7%	10.7
ArXiv	60.5%	79.1%	16.0
BBC News	69.4%	77.8%	15.9
Booking	70.0%	85.0%	24.8
Cambridge Dict.	86.0%	97.7%	5.7
Coursera	82.5%	97.5%	4.7
ESPN	53.8%	87.2%	14.9
GitHub	85.0%	92.5%	10.0
Google Flights	38.5%	51.3%	34.8
Google Map	78.9%	94.7%	9.6
Google Search	71.4%	92.9%	6.0
Huggingface	68.6%	74.3%	18.4
Wolfram Alpha	78.3%	93.5%	6.1

The proxy-lite web agent

Evaluation of the agent's ability to choose the correct actions in individual steps:

Action	TP + FN	TP + FP	TP	Recall	Precision	F1-Score
Click	20	22	12	0.6000	0.5455	0.5714
Type	10	13	5	0.5000	0.3846	0.4348
Scroll	10	4	3	0.3000	0.7500	0.4286
Return Value	10	10	9	0.9000	0.9000	0.9000

Total predictions	Correct predictions	Wrong predictions	Accuracy
50	29	21	0.5800

32nd cohort's agent's F1-Score

Action	Click	Type	Scroll	Return Value	All
F1-Score	0.6222	0.4761	0.0000	0.2858	0.5000

The proxy-lite web agent

The **scrolling problem:**

The agent has the tendency to not use its scrolling tool when necessary.

This is particularly problematic when running into cookie banners that require scrolling such as e.g. on Google.com.

The agent is designed to click on the first relevant element rather than weighing up multiple options.

The proxy-lite web agent

Evaluation on the GSM8K, CNN Dailymail and HellaSwag datasets:

Model	GSM8K	CNN Dailymail	HellaSwag
UI-TARS-7b DPO	Accuracy: 0.175	Rouge1: 0.142, Rouge2: 0.078, RougeL: 0.104, RougeLsum: 0.117	Accuracy: 0.265
proxy-lite-3b	Accuracy: 0.010	Rouge1: 0.096, Rouge2: 0.047, RougeL: 0.072, RougeLsum: 0.082	Accuracy: 0.265

Evaluation on the ScienceQA dataset:

Model	Accuracy
UI-TARS-7b DPO	0.4
proxy-lite-3b	0.395

Conclusion

- **The 32nd cohort's agent:** The agent achieved low finish and success rates despite being highly computationally expensive and slow.
- **The Midscene.js agent:** The agent frequently encountered errors (e.g. due to exceeding OpenAI's token threshold) and suffered from truncated/ incomplete data extraction.
- **The UI-TARS Desktop agent:** The agent excelled in complex web navigation and data extraction but proved to be resource intensive, requiring GPU support and Hugging Face endpoints.
- **The proxy-lite agent:** This agent emerged as the most efficient choice, offering a lightweight alternative with strong performance in web-based interactions.

Recommendation

Either proceed with the proxy-lite web agent in one of two ways -

1. Use only the 3.75B model and develop a self-built web agent, possibly replicating JavaScript navigation and scraping functionalities.
2. Use the original Proxy Lite web agent from GitHub and finetune its model for financial scraping tasks.

Fine Tuning Attempt (Rahul)

1. **Dataset** – FinLang (Investopedia instructions).
2. **LoRA Config** – r=64, alpha=128, dropout=0.1, lr=5e-5, batch size 1, 5 epochs.
3. **Issue** – **Stalled** with low GPU memory (8-bit load, 128 gradient steps), output stuck on “Context” vs. “Question/Answer.