

Agents for Enterprise Workflows

CS294/194-196 Large Language Model Agents
Lecture 7 — October 21st, 2024

Nicolas Chapados
ServiceNow Research

Alexandre Drouin
ServiceNow Research

Who are we?



Nicolas Chapados
ServiceNow Research



Alexandre Drouin
ServiceNow Research

Safe harbor notice for forward-looking statements

This presentation may contain “forward-looking” statements that are based on our beliefs and assumptions and on information currently available to us only as of the date of this presentation. Forward-looking statements involve known and unknown risks, uncertainties, and other factors that may cause actual results to differ materially from those expected or implied by the forward-looking statements. Further information on these and other factors that could cause or contribute to such differences include, but are not limited to, those discussed in the section titled “Risk Factors,” set forth in our most recent Annual Report on Form 10-K and Quarterly Report on Form 10-Q and in our other Securities and Exchange Commission filings. We cannot guarantee that we will achieve the plans, intentions, or expectations disclosed in our forward-looking statements, and you should not place undue reliance on our forward-looking statements. The information on new products, features, or functionality is intended to outline our general product direction and should not be relied upon in making a purchasing decision, is for informational purposes only, and shall not be incorporated into any contract, and is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. The development, release, and timing of any features or functionality described for our products remains at our sole discretion. We undertake no obligation, and do not intend, to update the forward-looking statements.

AGENDA

Background

Defining Agents
Enterprise workflow concepts

API Agents

Architecture
TapeAgents

Web Agents

Web Agent Concepts
WorkArena
BrowserGym and AgentLab

Agents in the Workplace

Automating enterprise workflows
Agents and the future of work

Resources to Dig Further

AGENDA

Background

Defining Agents
Enterprise workflow concepts

API Agents

Architecture
TapeAgents

Web Agents

Web Agent Concepts
WorkArena
BrowserGym and AgentLab

Agents in the Workplace

Automating enterprise workflows
Agents and the future of work

Resources to Dig Further

LLM agents are LLM-powered entities able to **autonomously plan and take actions to execute goals** over multiple iterations.

LLM-Based Agents

Reinforcement Learning Agents

- Require long training runs in sandboxed environments
- Limited action space
- Low generalizability to radically new tasks
- A Minecraft agent can't send emails

LLM-Based Agents: Zero-Shot Task Solvers

- LLMs can display some commonsense, since they have lots of world background knowledge
- General-Purpose LLMs have probably been trained on the documentation of your software

Two kinds of LLM Agents

API agents

- Observations: API call results, search history, user-uploaded images, chat history
- Actions: API calls, search calls, responses to the user
- Pros: Lower latency, lower risks
- Cons: needs appropriate APIs

Web agents

- Observations: what human would see + accessibility tree / raw DOM
- Actions: enter text in fields, clicks
- Pros: can do anything
- Cons: higher latency, higher risks

Today's Enterprise Workflows Remain Quite Manual (even with generative AI)

Jon needs access to a KB



Do we have a KB that explains what to do?

Is there a similar incident?

What access control does Jon have?

Assign Jon the right role

Resolve

I generated resolution notes

Close

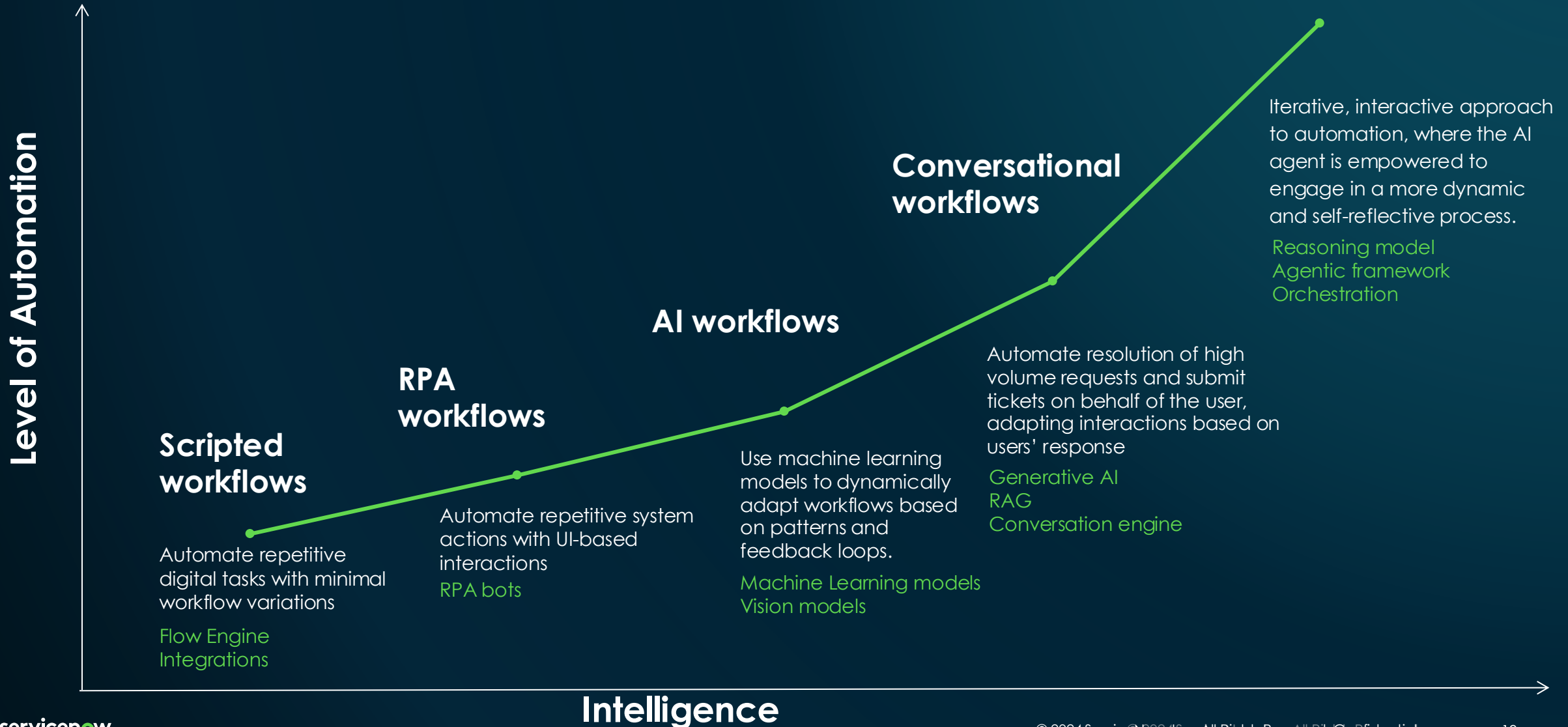
Human Agent



Resolution Generation skill

GenAI

Automation in Enterprise Workflows



Agents solve for the **Millions** of **Low-Value/Low-Volume Tasks**

Today's automation workhorses for high-value or high-volume tasks

- Robotic Process Automation
- Low-Code / No-Code

What About?

- Scheduling tweets
- Sorting email
- Updating CRM
- Filling out time sheet
- Arranging 15-person meeting across 4 organizations



Demo: Directions to GTC

(original video is 4x longer
with long pauses)



Hi! I am your UI assistant, I can perform web tasks for you. What can I help you with?



Sign in to Google

Save your passwords securely with your Google Account

Stay signed out

Sign in

Google

Google Search

I'm Feeling Lucky

Google offered in: Français

Send

Canada

[Advertising](#) [Business](#) [How Search works](#)

[Privacy](#) [Terms](#) [Settings](#)

AGENDA

Background

Defining Agents
Enterprise workflow concepts

API Agents

Architecture
TapeAgents

Web Agents

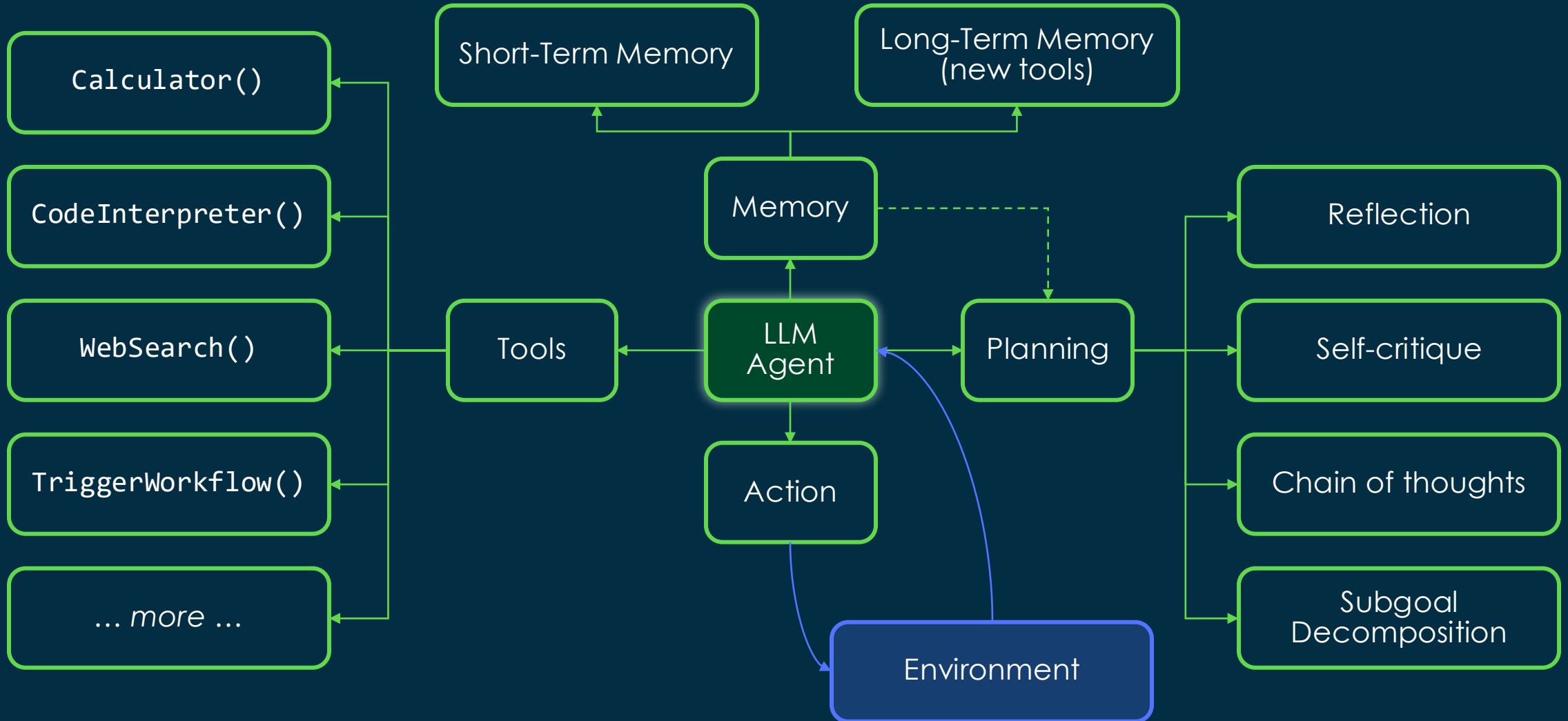
Web Agent Concepts
WorkArena
BrowserGym and AgentLab

Agents in the Workplace

Automating enterprise workflows
Agents and the future of work

Resources to Dig Further

LLM-Based Single Agents: Typical Architecture



TapeAgents: towards a holistic framework for agent development and optimization



Frameworks that address agent development needs

- Resumable sessions
- Low-code components
- Fine-grained control
- Concurrency
- Streaming

LangGraph, AutoGen, Crew:

- Agent == resumable modular state machine

Frameworks for data-driven agent optimization

- Structured agent configuration
- Structured agent logs
- Optimization algorithms

DSPy, TextGrad, Trace:

- Agent == code that uses structured modules and generates structured logs

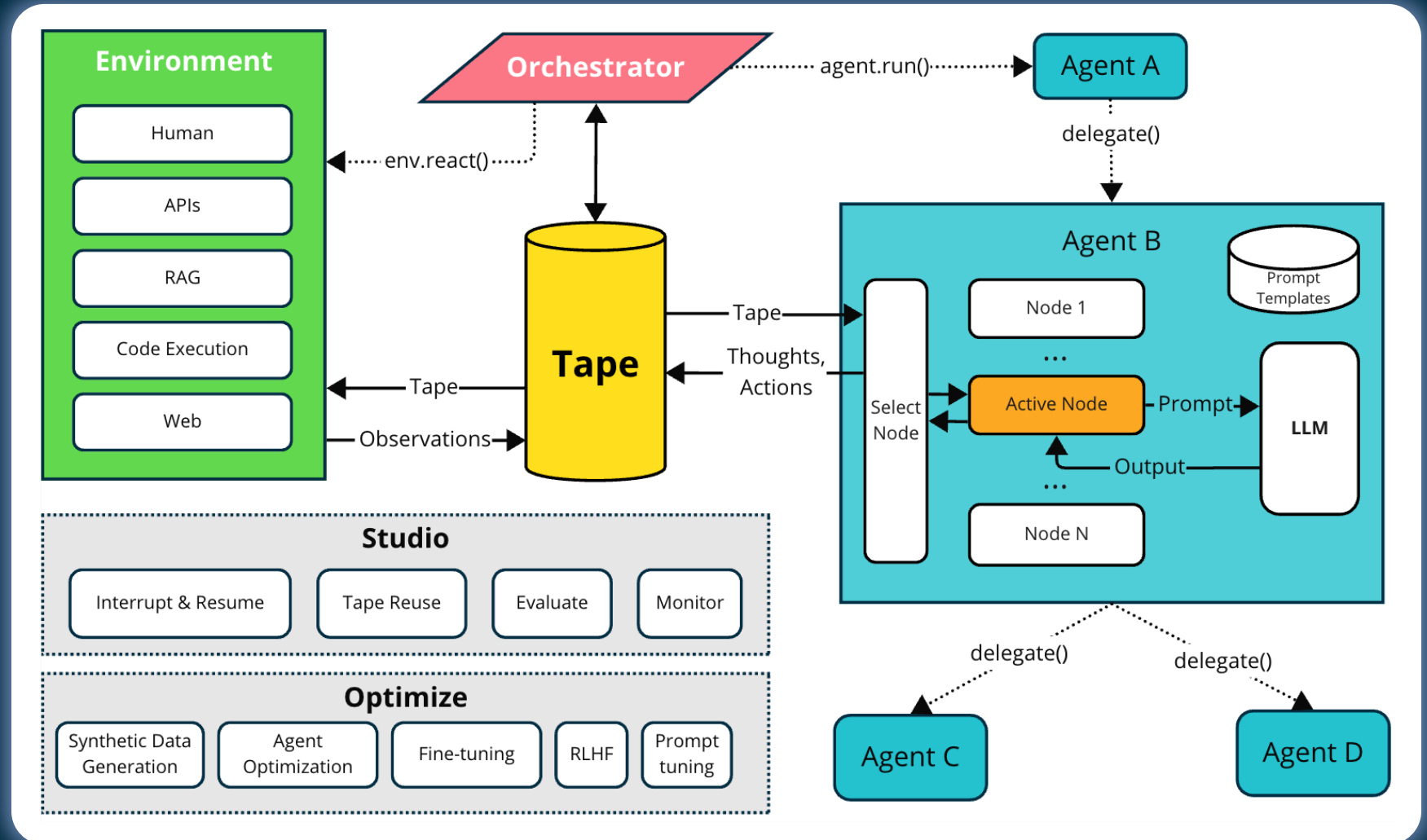
Holistic Frameworks

TapeAgents:

- Agent == Resumable modular state machine
- ... with structured configuration
- ... that makes granular structured logs
- ... that can make fine-tuning data from logs
- ... and reuse other agent's logs

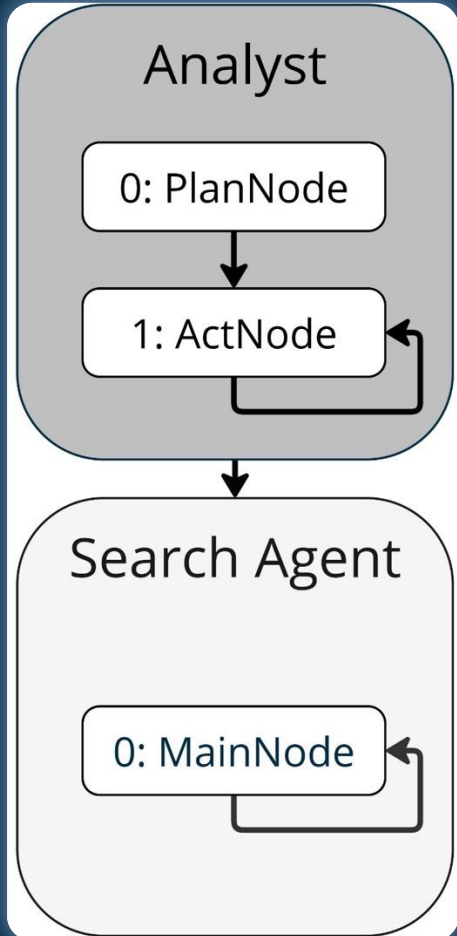
TapeAgents is a framework built around a structured, granular, semantic-level log: the **tape**

- Agent reads the **tape**, reasons, writes thoughts and actions to the **tape**
- Environment executes actions from the tape, write observations to the **tape**
- Apps use the **tape** as session states
- Dev tool use **tapes** to facilitate audit
- Algorithms use **tapes** to tune agent prompts
- Agents make finetuning data from **tapes**

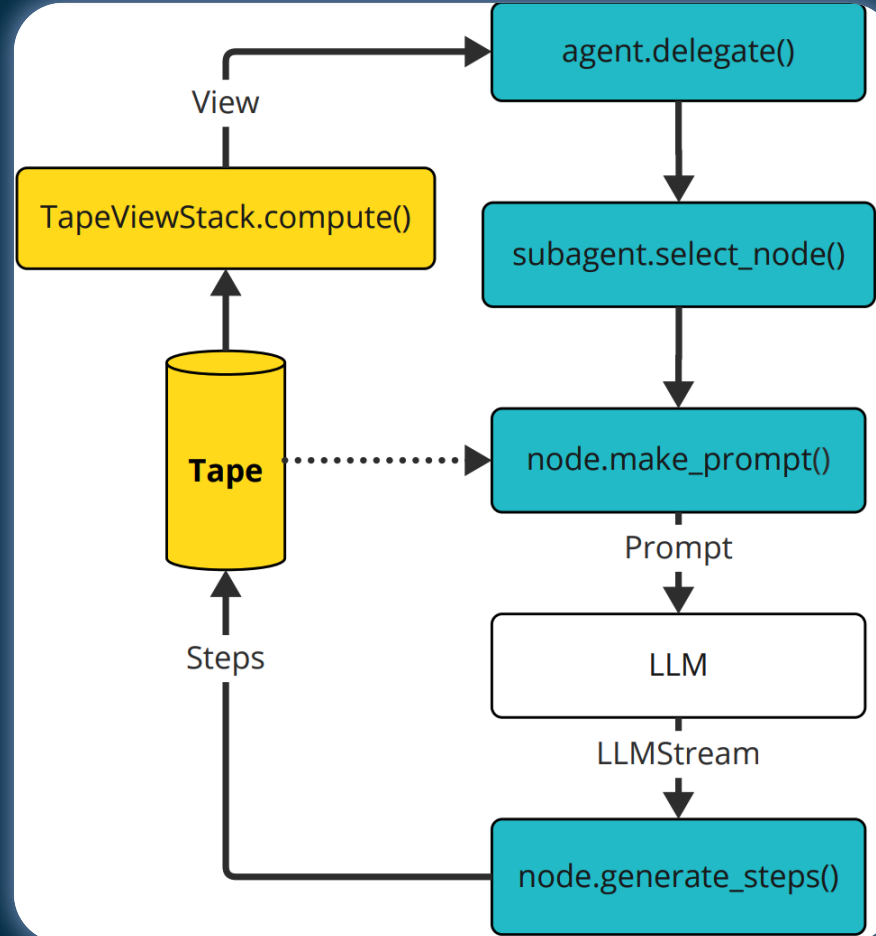


Agent reasoning loop: example

Simple two-agent structure
(problem-specific)



TapeAgents execution
model



```
Environment
[0] User
Tell me about Vulcan in 3 sentences
analyst.PlanNode
▶ Prompt 109 tokens | Completion 114 tokens
[1] Thought: AssistantThought
▶ 1. Use the `functions.get_stock_ticker` tool to find the sto...
analyst.ActNode
▶ Prompt 212 tokens | Completion 0 tokens
[2] Thought: SetNextNode(1)
[3] Action: ToolCalls
get_stock_ticker(company_name='Vulcan')
```



```
Environment
[7] Observation: ToolResult
▶ [('2023-09-18', 211.69000244140625), ('2023-09-22', 200.6199...
analyst.ActNode
▶ Prompt 1297 tokens | Completion 0 tokens
[8] Thought: SetNextNode(1)
[9] analyst calls search_agent
Vulcan Materials Company overview and recent news
search_agent.SearchAgentMainNode
▶ Prompt 21 tokens | Completion 0 tokens
[10] Action: ToolCalls
▶ get_search_results(query='Vulcan Materials Company overview'...
[11] Thought: SetNextNode(0)
```

[0] User

kind: user

Tell me about Vulcan in 3 sentences

▶ **Prompt 1681 characters**

▶ **Completion**

[1] Thought: AssistantThought

by: Agent

kind: assistant_thought

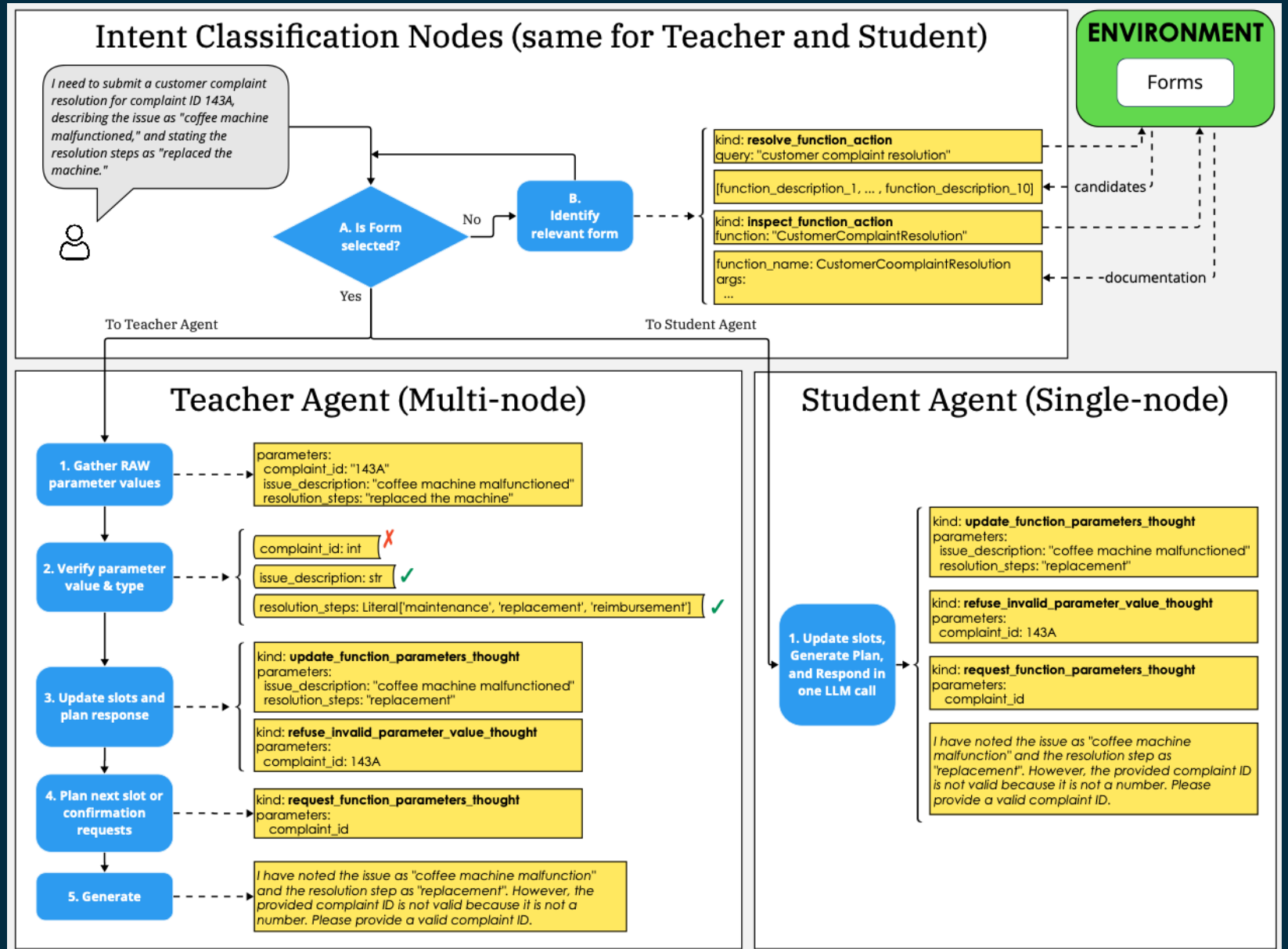
To help the user learn about Vulcan, I will:

1. Use the `functions.get_stock_ticker` tool to find the stock ticker symbol for Vulcan.
2. Use the `functions.get_stock_data` tool to retrieve recent stock price data for Vulcan using the ticker symbol obtained in step 1.
3. Summarize the information about Vulcan, including its stock ticker and recent stock performance, in a concise manner.

▶ **Prompt 2045 characters**

▶ **Completion**

TapeAgents allows the optimization of a Student Agent from the tapes of a Teacher Agent



MAKING COST-EFFECTIVE

G R E A D T H

(CONVERSATIONAL) AGENTS

MAKING COST-EFFECTIVE

G R E A D T H

GROUNDED

RESPONSIVE

ACCURATE

DISCIPLINED

TRANSPARENT

HELPFUL

(CONVERSATIONAL) AGENTS

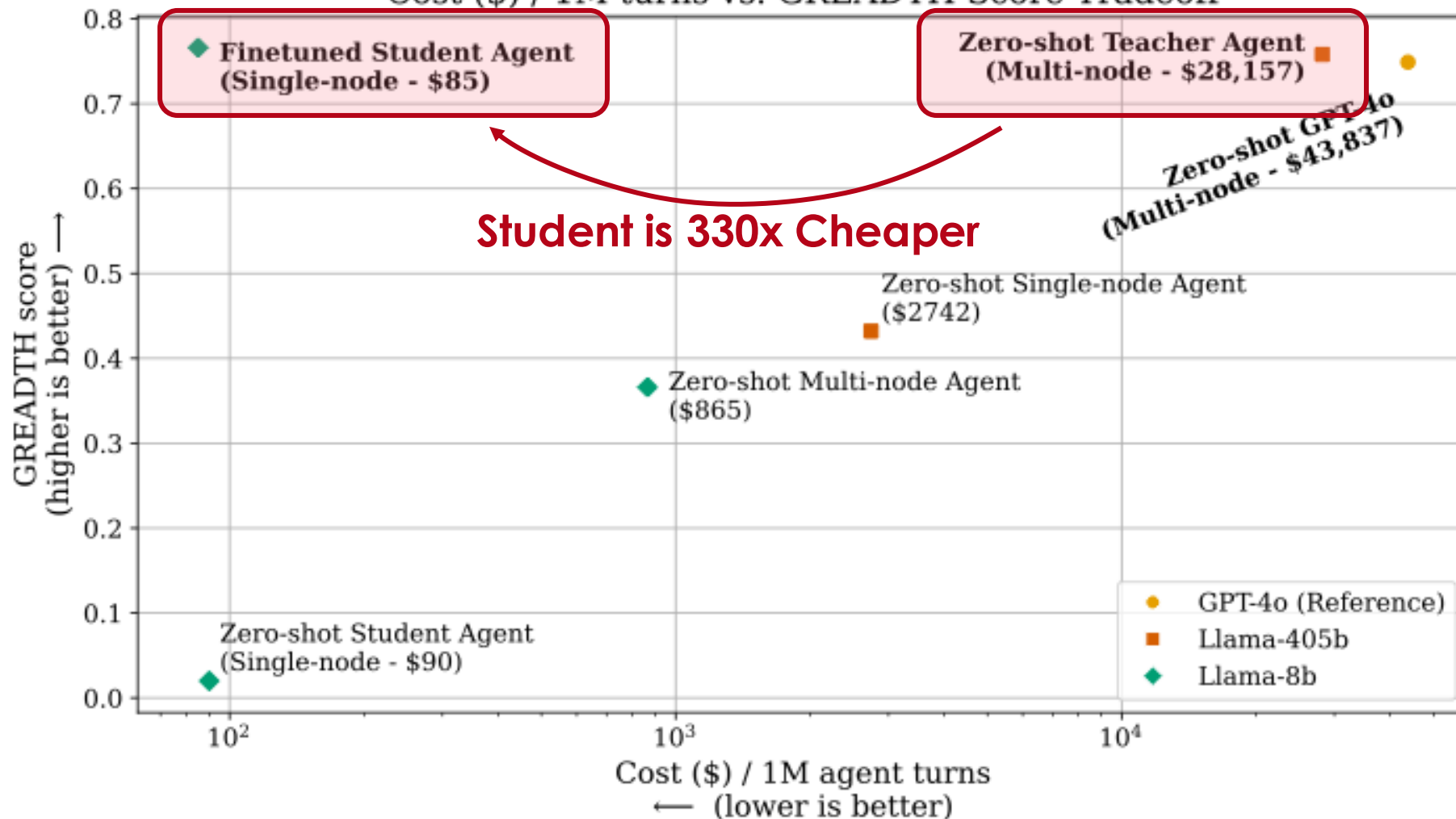
Case Study: Cost-Effective Form-Filling Assistant

- **Task:** conversational assistant that routes the user to the right form and helps fill it
- **Constraints:** 5-star conversational experience at low compute cost
- **3 training domains:** FlyCorp, BigBankCorp, CoffeeCorp
- **3 testing domains:** DriveCorp, LuxuryCorp, ShopCorp
- **Metric: GREADTH**
 - Grounded, Responsive, Accurate, Disciplined, Transparent, Helpful
- **Method:**
 - Generate synthetic tapes with 19 user agents and a 5-node LLAMA-405B Teacher
 - Finetune 1-node LLAMA-8B Student
- **Outcome:** student matches GPT-4o performance at 300x lower cost

Table 3: GREADTH Form Filler experiment results. The Teacher¹ is a multi-node agent with Llama 3.1 405B Instruct FP8 as its LLM. The Student² is a single-node agent with Llama 3.1 8b Instruct as its LLM. We also evaluate the multi-node agent with GPT-4o and with Llama 3.1 8B Instruct as its LLM, as well as the single-node agent with Llama 3.1 405B Instruct for comparison. The metrics are computed over 1524 partial dialogues from the test domains. Read full analysis in Section 5.4.

Agent (LLM+Nodes)	G	Re	A	D	T	H	GREADTH Score (Human Raters)
<i>Reference Comparison (GPT-4o-2024-08-06)</i>							
Multi-node (0-shot)	91.3%	87.1%	91.4%	92.7%	94.3%	87.2%	74.9%
<i>Llama-3.1-405B-Instruct</i>							
Teacher ¹ : Multi-node (0-shot)	89.8%	85.0%	87.9%	91.6%	92.5%	86.5%	75.8%
Single-node (0-shot)	74.2%	72.0%	76.8%	67.3%	78.9%	61.9%	43.2%
<i>Llama-3.1-8B-Instruct</i>							
Multi-node (0-shot)	75.5%	57.7%	72.4%	74.0%	76.3%	60.3%	36.6%
Student ² : Single-node (0-shot)	18.8%	6.2%	10.9%	11.6%	9.4%	12.7%	2.0%
Student ² : Single-node (finetuned)	92.1%	86.4%	90.2%	94.4%	95.1%	87.1%	76.6%

Cost (\$) / 1M turns vs. GREADTH Score Tradeoff



Agentic Frameworks: How Does TapeAgents Compare?

Method	Development				Optimization			
	Building from Components while Allowing Finegrained Flow Control	Native Streaming Support	Concurrent LLM Calls	Resumable State Machine Agents	Log Reuse Across Agents	Structured and Agent Configurations for Data-Driven Agent Optimization	Logs for Agent	Making Training Text From Semantic-Level Logs
DSPy	✓	✗	✓	✗	✗	✓		▲
LangGraph	✓	✓	✓	✓	▲	▲		✗
AutoGen	▲	▲	✗	▲	✗	▲		✗
TapeAgents (Ours)	✓	✓	✗	✓	✓	✓		✓

Table 5: TapeAgents vs Other Frameworks. *TapeAgents stands out in features it offers to the practitioner to the support them throughout the LLM Agent development cycle. In this figure, we use the cross sign (✗) to indicate that major core changes would be required for the framework support the feature. Triangle sign (▲) indicates partial support of a feature, meaning that practitioner would have to do extra effort or accept associated limitations to achieve the respective functionality. Check sign (✓) indicates that the framework natively supports a feature. TapeAgents’s only weakness in this table is the lack of Concurrent LLM Calls, see Section 7 for a discuss of how we intend to tackle it.*

AGENDA

Background

Defining Agents
Enterprise workflow concepts

API Agents

Architecture
TapeAgents

Web Agents

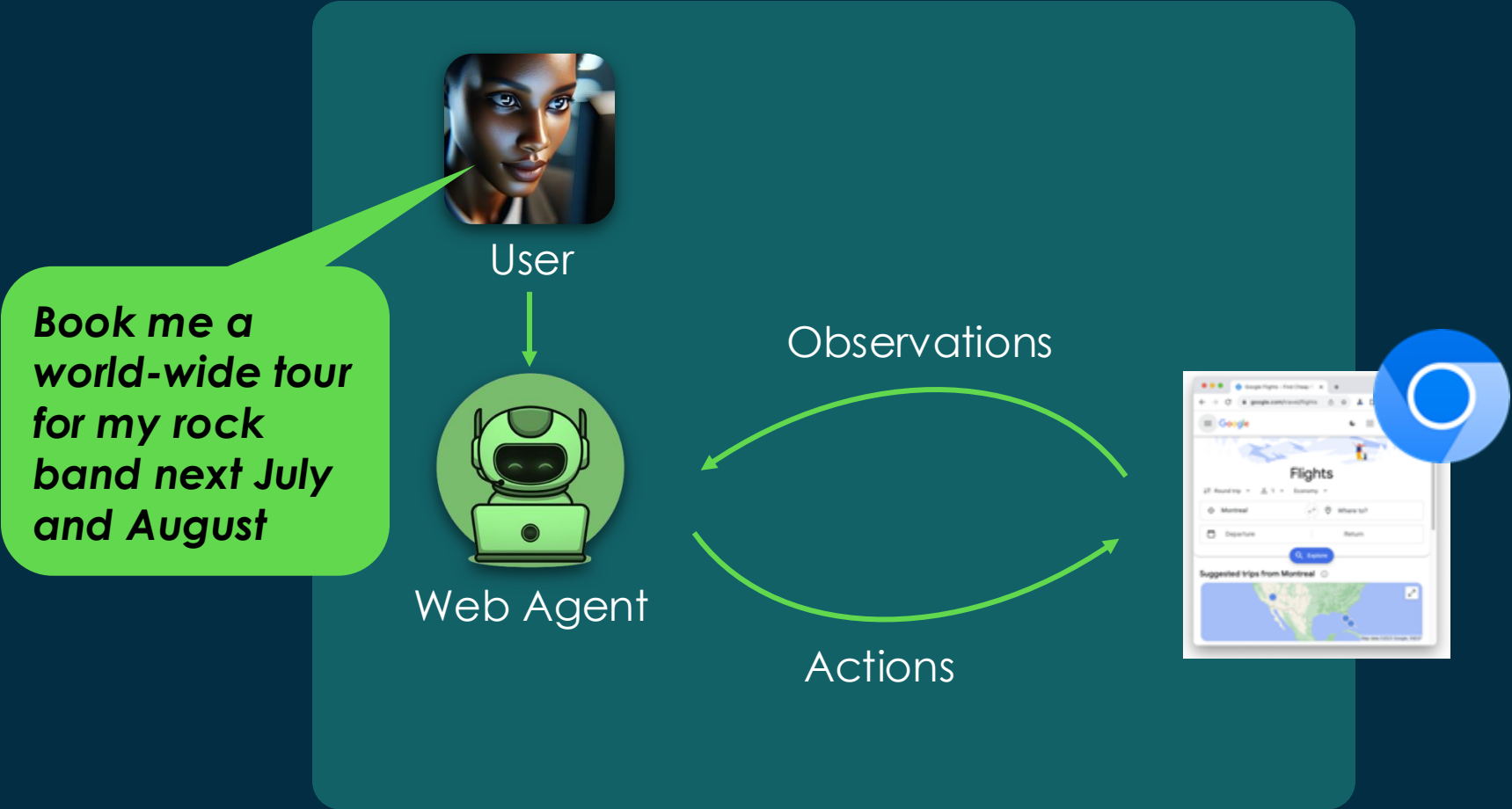
Web Agent Concepts
WorkArena
BrowserGym and AgentLab

Agents in the Workplace

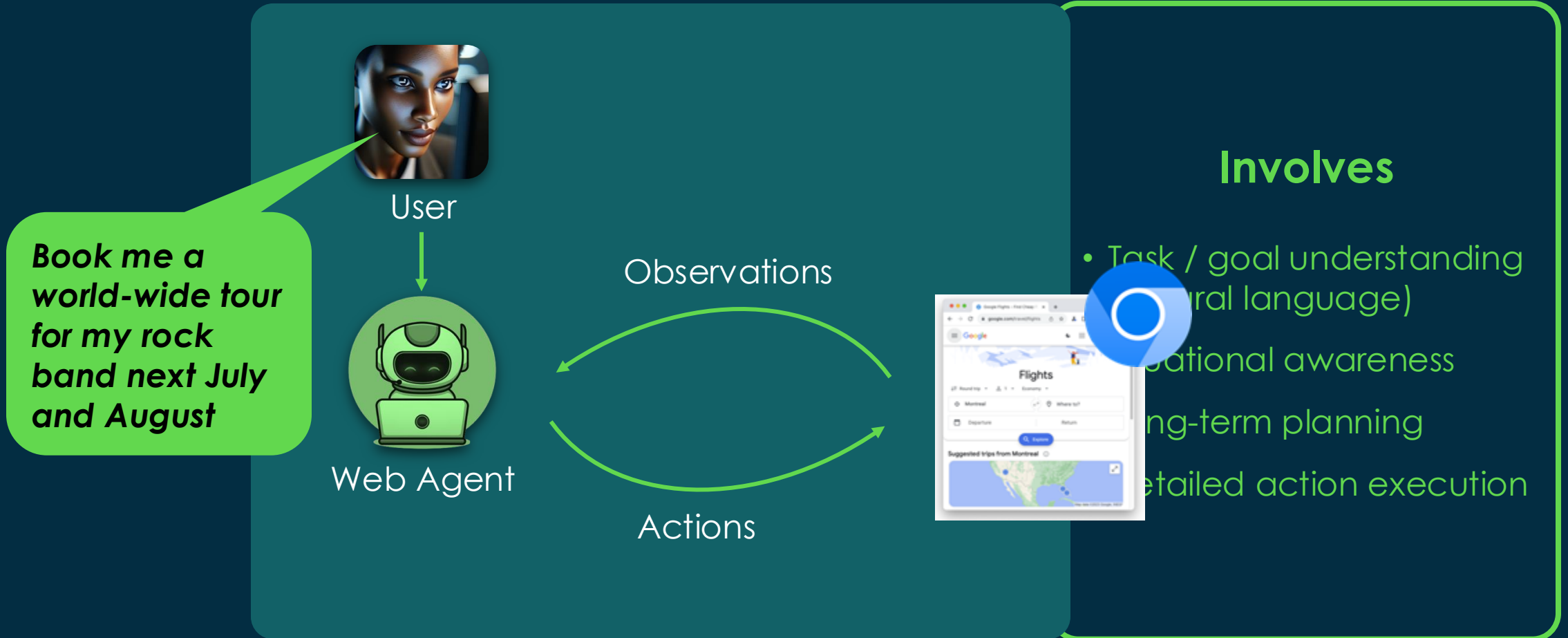
Automating enterprise workflows
Agents and the future of work

Resources to Dig Further

What is a Web Agent?



Web Agents Act on the Web on Behalf of Human Users



Making a basic Web Agent



You can do this by prompting an LLM

Example prompt (simplified):

```
Task:
- Enter "Enola" into the text field and press Submit.

DOM (Web Page):
<html>
<body>
...
</body>
</html>

Action space:

# Fill out a form field
fill(backend_id: str, value: str)

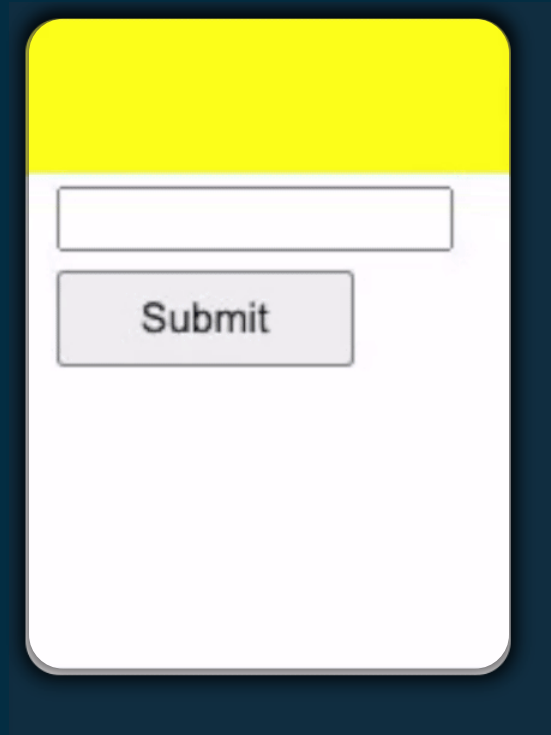
# Click an element
click(backend_id: str)

# Move the mouse to a location
mouse_move(x: float, y: float)

Answer Format:
<action>
Your actions
</action>
```

LLM response:

```
<action>
fill('14', 'Enola')
click('15')
</action>
```



SAP Concur Expense

Manage Expenses Card Transactions

Home / Expense / Manage Expenses

Manage Expenses

Report Library

Create New Report

View: Active Reports

AWS Expense April 2024
04/25/2024
\$265.04
Returned

Sent Back to Employee
Concur System


Available Expenses

Upload Receipt

Drag and drop files to upload a new receipt. Valid file types for upload are .png, .jpg, .jpeg, .pdf, .tif or .tiff.

View Edit Delete Combine Expenses Move to


View: All Expenses



No Available Expenses

New/Incoming expenses will be added to this list.
To find missing transactions: [Card Transactions](#)

Hi! I am your UI assistant, I can perform web tasks for you. What can I help you with?

How can I help you? 

SAP Concur Expense

Manage Expenses Card Transactions

Home / Expense / Manage Expenses

Manage Expenses

Report Library

Create New Report

View: Active Reports

AWS Expense April 2024
04/25/2024
\$265.04
Returned

Sent Back to Employee
Concur System


Available Expenses

Upload Receipt

Drag and drop files to upload a new receipt. Valid file types for upload are .png, .jpg, .jpeg, .pdf, .tif or .tiff.

View Edit Delete Combine Expenses Move to


View: All Expenses



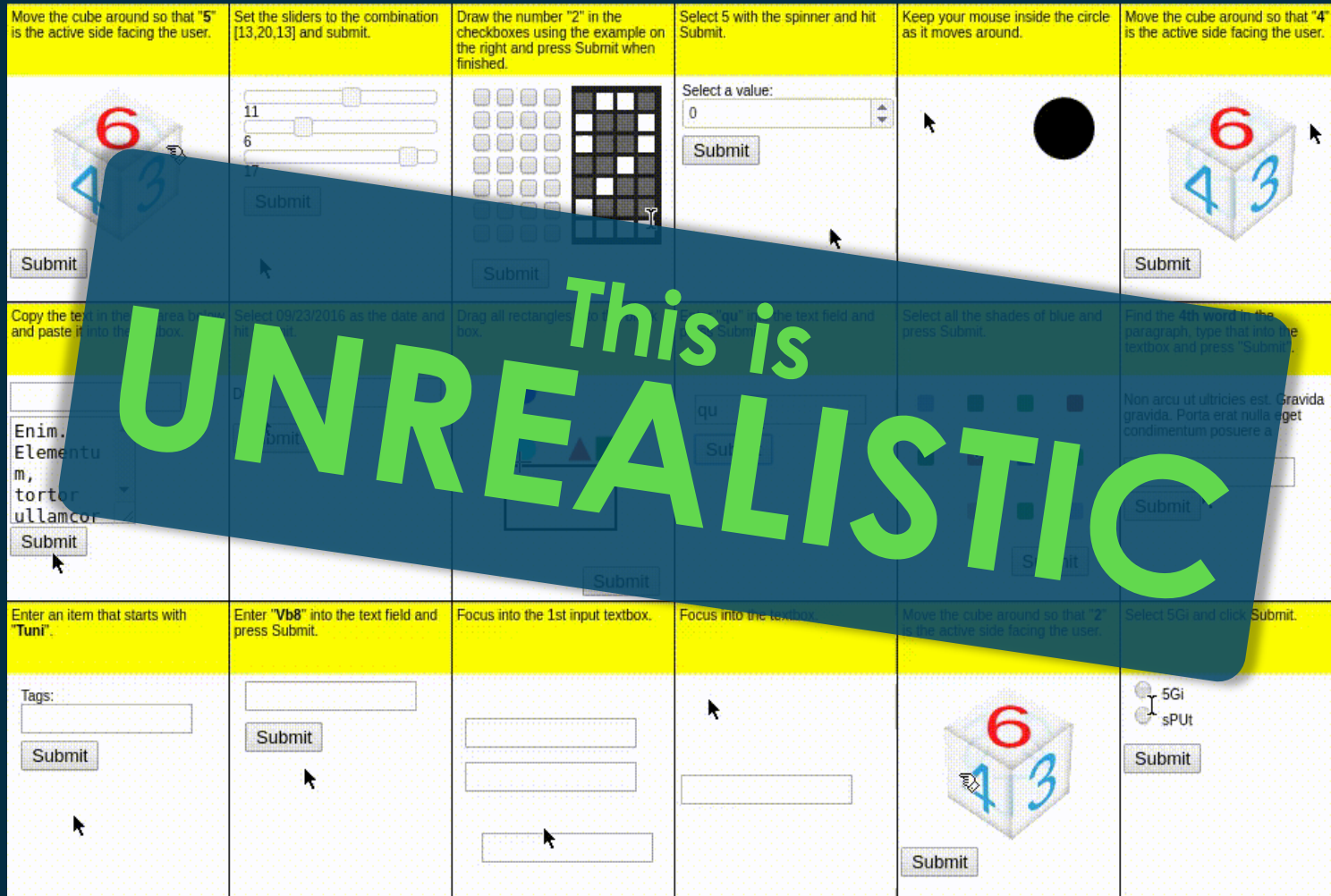
No Available Expenses

New/Incoming expenses will be added to this list.
To find missing transactions: [Card Transactions](#)

Hi! I am your UI assistant, I can perform web tasks for you. What can I help you with?

How can I help you? 

How do we evaluate web agents?



Source: <https://miniweb.farama.org/index.html> (MiniWoB++)

Realistic Trace-based Benchmarks

Thousands of ~~human-generated~~ observation-action traces

- ✓ Real websites
- ✗ Evaluation based on “gold traces” (what about alternative solutions?)
- ✗ Traces can be memorized

Mind2Web (Deng et al., 2023)

(a) Find one-way flights from New York to Toronto.

(b) Book a roundtrip on July 1 from Mumbai to London and vice versa on July 5 for two adults.

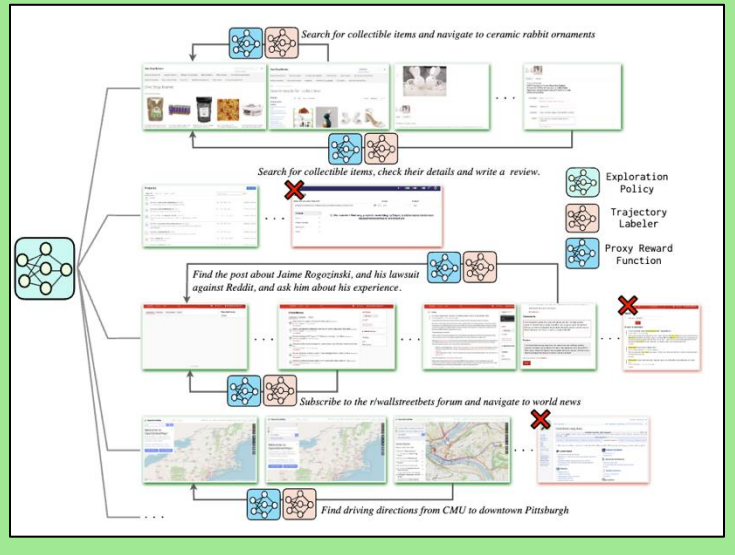
(c) Find a flight from Chicago to London on 20 April and return on 23 April.

(d) Find Elon Musk's profile and follow, start notifications and like the latest tweet.

(e) Browse comedy films streaming on Netflix that was released from 1992 to 2007.

(f) Open page to schedule an appointment for car knowledge test.

NNetNav (Murty et al., 2024)



WebLINX (Lù et al., 2024)

Could you please find me an Italian Restaurant in Berlin on Quandoo?

say("Yes, sure")

load("quandoo.com")

input("Italian")

say("I found the following restaurants: Cavallino Rosso, Delizie D'Italia, Osteria Fiorello")

Which one has the highest reviews?

say("The first one, Cavallino Rosso")

Please open it and check if it's available for 2 people on the date of 31st May 2023 at 7:30 PM.

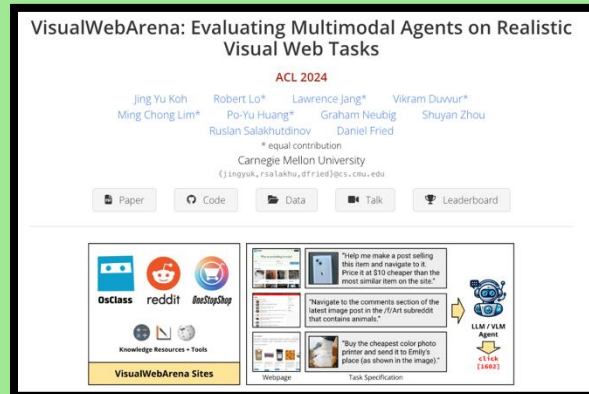
Realistic Live Environment Benchmarks

Evaluate end result rather than sequence of actions (e.g., database state)

✓ Agnostic to action trace

✓ Low memorization risk (no traces)

Sandboxed environments



Tasks performed on **locally hosted** server

✓ High bandwidth (for parallel experiments)

✗ Limited to open-source software

✗ Complex local setup (e.g., Docker)

Tasks performed on a **remote** server

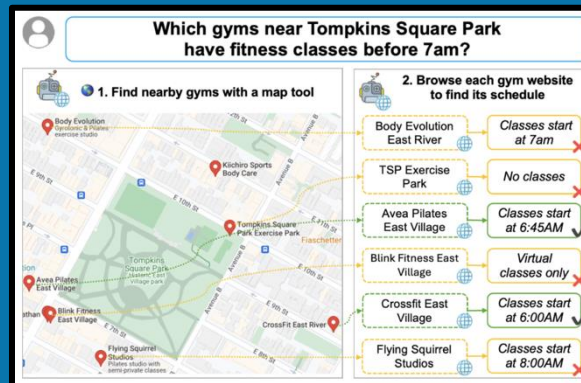
✓ More realistic (supports any website, latency)

✓ No need for complex local setup

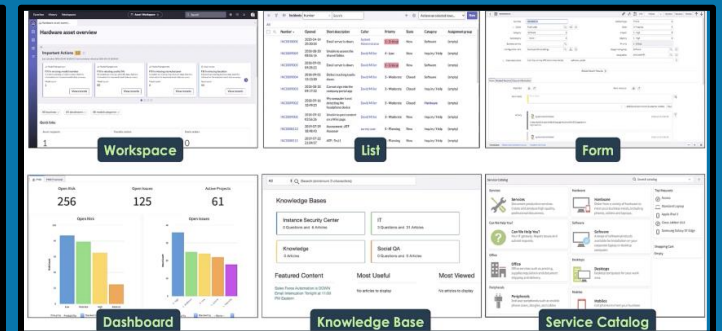
✗ Can be unreliable (network issues)

Open Web Environments

AssistantBench (Yoran et al., 2024)



WorkArena (Drouin, Gasse et al., 2024)





WorkArena



`pip install browsergym-workarena`

An open-source benchmark of ~600 work-related tasks built on the ServiceNow platform

The image displays five different views of the ServiceNow user interface:

- Workspace:** A dashboard with 'Important Actions' and 'Quick Links'.
- List:** A table view of incidents with columns for Number, Opened, Short description, Caller, Priority, State, Category, and Assignment group.
- Dashboard:** A dashboard with three charts: 'Open Risk' (256), 'Open Issues' (125), and 'Active Projects' (61).
- Knowledge Base:** A page with 'Knowledge Bases' and 'Featured Content'.
- Service Catalog:** A page with 'Office' and 'Peripherals' items.

The screenshot shows the ServiceNow Developer site with the following text:

servicenow Developer Learn Reference Guides Connect

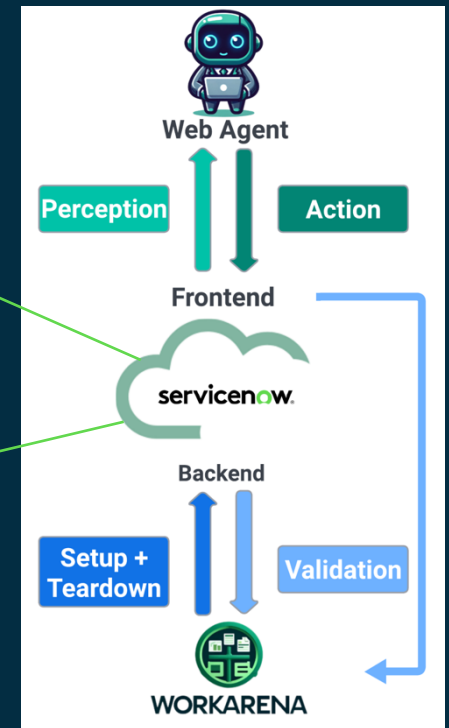
Your instance is ready! **Free**

Your instance URL: <https://...service-now.com>

Username: admin
Current password: [redacted]

Keep your new instance active by developing on the instance or logging into the Developer Site. If your instance is inactive for 10 days, it will be reclaimed and released for other developers to use.

Return to the Developer Site Open Instance



Tasks span basic UI interactions and complex realistic workflows

Open Web

Service Catalog

Search catalog

**Services****Services**

Document production services. Create and produce high-quality, professional documents.

Hardware**Hardware**

Order from a variety of hardware to meet your business needs, including phones, tablets and laptops.

Shopping Cart

Empty

Can We Help You?**Can We Help You?**

Your IT gateway. Report issues and submit requests.

Software**Software**

A range of software products available for installation on your corporate laptop or desktop computer.

Office**Office**

Office services such as printing, supplies requisition and document shipping and delivery.

Desktops**Desktops**

Desktop computers for your work area.

Peripherals**Peripherals**

End user peripherals such as mobile phone cases, dongles, and cables

Mobiles**Mobiles**

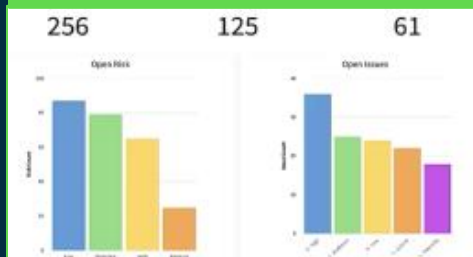
Cell phones to meet your business needs.

WorkArena++ Towards Realistic Enterprise Workflows

1 Knowledge base



2 Dashboard



3 Service Catalog



Example: The agent is assigned a ticket and instruction: "Please solve this."

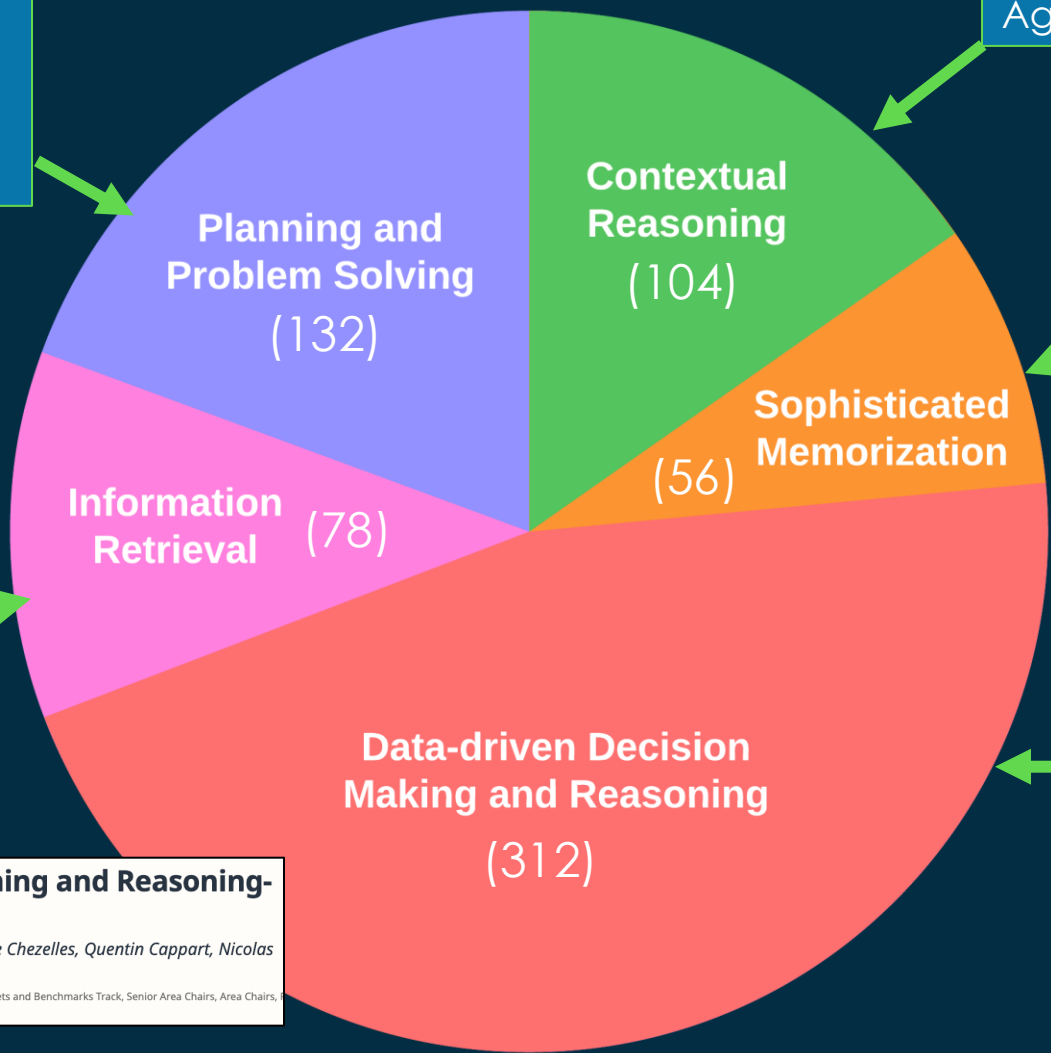
WorkArena++ Towards Realistic Enterprise Workflows

Overview of tasks

Solve a series of enterprise decision-making problems:

- Workload balancing
- Scheduling with constraints
- Assigning work to experts

Some tasks are purposely infeasible. Agent must detect this.



Read dashboards and act:

- Restock IT asset inventory

Search for information in lists, forms, and KBs:

- Find if a user's laptop is under warranty

Navigate the platform to gather multiple bits of information and then solve a task:

- Offboard a user

Read dashboard, make calculations, take action

Budget management: choose where to invest based on expected return

Expense management

WorkArena++: Towards Compositional Planning and Reasoning-based Common Knowledge Work Tasks

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier de Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, Alexandre Drouin

Published: 26 Sept 2024, Last Modified: 26 Sept 2024 | NeurIPS 2024 Track Datasets and Benchmarks Poster | Datasets and Benchmarks Track, Senior Area Chairs, Area Chairs, Authors | Revisions | BibTeX | CC BY 4.0

WorkArena++ is far from being solved

Realistic Workflows

Task Category (task count)	Agent Curriculum (full benchmark)					Human
	GPT-3.5	GPT-4o	GPT-4o-v	Llama3	Mixtral	
<i>Success rate (higher is better)</i>						
WorkArena L3 (235)						93.9 ±3.4
Contextual Understanding (32)						87.5 ±11.7
Data-driven Decision-Making (55)						100.0 ±0.0
Planning and Problem Solving (44)						87.5 ±11.7
Information Retrieval (56)						100.0 ±0.0
Sophisticated Memorization (48)						91.7 ±8.0
WorkArena L2 (235)						93.9 ±3.4
Contextual Understanding (32)						100.0 ±0.0
Data-driven Decision-Making (55)						84.6 ±10.0
Planning and Problem Solving (44)						100.0 ±0.0
Information Retrieval (56)						100.0 ±0.0
Sophisticated Memorization (48)						91.7 ±8.0
WorkArena L1 (33 × 10 seeds)						
MiniWoB (125 × 5 seeds)						
WebArena (812)						

WorkArena++ is far from being solved

Realistic Workflows

Task Category (task count)	Agent Curriculum (full benchmark)					Human
	GPT-3.5	GPT-4o	GPT-4o-v	Llama3	Mixtral	
WorkArena L3 (235)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	93.9 ±3.4
Contextual Understanding (32)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	87.5 ±11.7
Data-driven Decision-Making (55)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	100.0 ±0.0
Planning and Problem Solving (44)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	87.5 ±11.7
Information Retrieval (56)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	100.0 ±0.0
Sophisticated Memorization (48)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	91.7 ±8.0
WorkArena L2 (235)	0.0 ±0.0	3.0 ±1.1	3.8 ±1.3	0.0 ±0.0	0.0 ±0.0	93.9 ±3.4
Contextual Understanding (32)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	100.0 ±0.0
Data-driven Decision-Making (55)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	84.6 ±10.0
Planning and Problem Solving (44)	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	0.0 ±0.0	100.0 ±0.0
Information Retrieval (56)	0.0 ±0.0	0.0 ±0.0	3.6 ±2.5	0.0 ±0.0	0.0 ±0.0	100.0 ±0.0
Sophisticated Memorization (48)	0.0 ±0.0	14.6 ±5.1	14.6 ±5.1	0.0 ±0.0	0.0 ±0.0	91.7 ±8.0
WorkArena L1 (33 × 10 seeds)	6.1 ±1.3	42.7 ±2.7	41.8 ±2.7	17.9 ±2.1	12.4 ±1.8	
MiniWoB (125 × 5 seeds)	43.4 ±1.6	71.3 ±1.5	72.5 ±1.5	68.2 ±1.2	62.4 ±1.6	
WebArena (812)	6.7 ±0.9	23.5 ±1.5	24.0 ±1.5	11.0 ±1.1	12.6 ±0.5	

What explains this?

- Failure to plan
- Hallucinated controls
- Incorrect action syntax

Benchmark Explosion



- MiniWoB++ (Shi et al., 2017; Liu et al., 2018) **125 tasks**
- WebShop (Yao, Chen et al., 2022) **12 087 tasks**
- WebArena (Zhou et al., 2023) **812 tasks**
- VisualWebArena (Koh et al., 2024) **910 tasks**
- WebLINX (Lù et al., 2024) **2 300 tasks**
- WebCanvas (Pan et al., 2024) **438 tasks**
- WebVoyager (He et al., 2024) **643 tasks**
- AssistantBench (Yoran et al., 2024) **214 tasks**
- WorkArena++ (ServiceNow Research, 2024) **682 tasks**

Call for unification

Get everyone under the same roof for a great Meta-Benchmark





BrowserGym



`pip install browsergym`

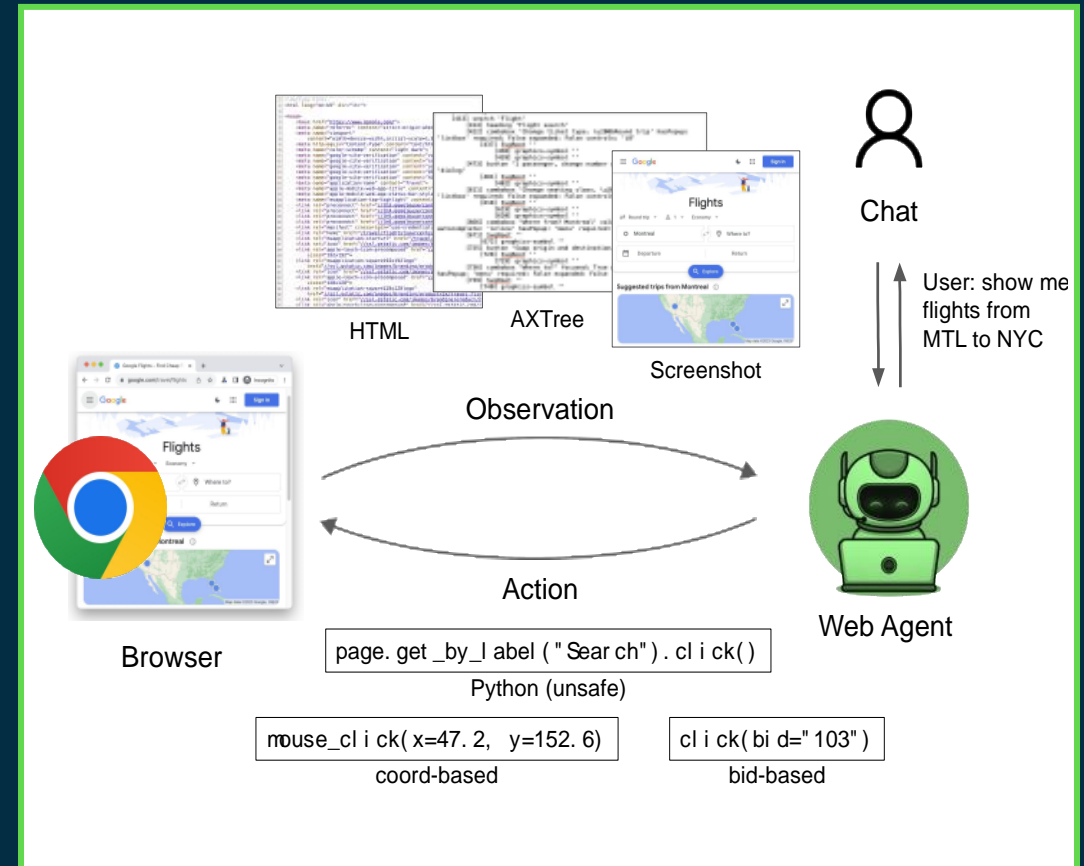
A unified evaluation platform

> Standard Observation Space

- HTML
- Screenshots
- Accessibility Tree
- And more

> Standard Action Space

> Regroups most major benchmarks (thousands of realistic tasks)





BrowserGym



`pip install browsergym`

urboBots.AI All Favorites History Private Task - Clean-up your duplica... Search

Private Task

Clean-up your duplicate problems

Number: Priority: 4 - Low

* Owner: State: Open

Assigned to: Parent:

Active

Short description:

Description: Referring to company protocol "Problem List Cleanup" -located in the "Company Protocols" knowledge base- clean-up your problem list (problems assigned to you) by marking duplicate tasks as "Closed - complete" or "Closed - skipped". Don't forget to mark this task as "Closed - complete" once successfully completed. Otherwise, you can mark the task as "Closed - skipped".

Additional comments (Customer visible):

Work notes Post

Activities: 1

- ML Meghan Lewis
 - Assigned to: Meghan Lewis
 - Impact: 3 - Low
 - Opened by: Meghan Lewis
 - Priority: 4 - Low
 - State: Open

Field changed

Human Evaluation Console

Task 1 / 100 --- Elapsed: 5.5 sec.

Task not completed. Keep going.

18:21:38 - BOT

Hi! I am your UI assistant, I can perform web tasks for you. What can I help you with?

18:21:38 - YOU

Please complete the following task.

How can I help you?



Human evaluation for any benchmark!



AgentLab

A toolbox for agent design

- > Simple building blocks for agents
- > Tools to inspect their behavior
- > Experimental framework:
 - > Easy large-scale evaluation
 - > Reproducibility features

```
class MyAgent(bgym.Agent):  
  
    def get_action(self, obs) -> str:  
        ✨ action = do_some_reasoning(obs)  
        return action  
  
study = run_agents_on_benchmark(MyAgent(), "workarena.l1")  
study.run(n_jobs=10, parallel_backend="joblib")
```

AgentXRay

Help

Select Experiment Directory

Select Agent Select Task and Seed Constants and Variables Global Stats

Agent Selector (click for help)

1	2	3

Screenshot Screenshot Pair Screenshot Gallery DOM HTML Pruned DOM HTML AXTree Chat Messages Task Error Logs Stats Agent Info HTML Agent Info MD

Prompt tests

Raw Screenshots



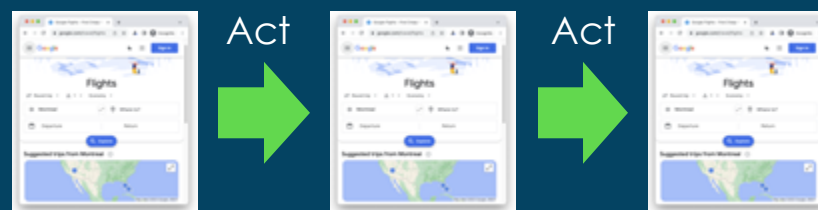
AgentLab

Reproducibility as a priority

Benchmarking on the open web is challenging (*dynamic environment*)

- > Websites are updated
- > API-based LLMs change silently
- > Python packages evolve

- > Standardized observation/action traces



- > Experimental journal uploaded to public repo
Date, versions, agent configuration, traces, etc.

- > Leaderboards with scores that are automatically reproduced based on the above



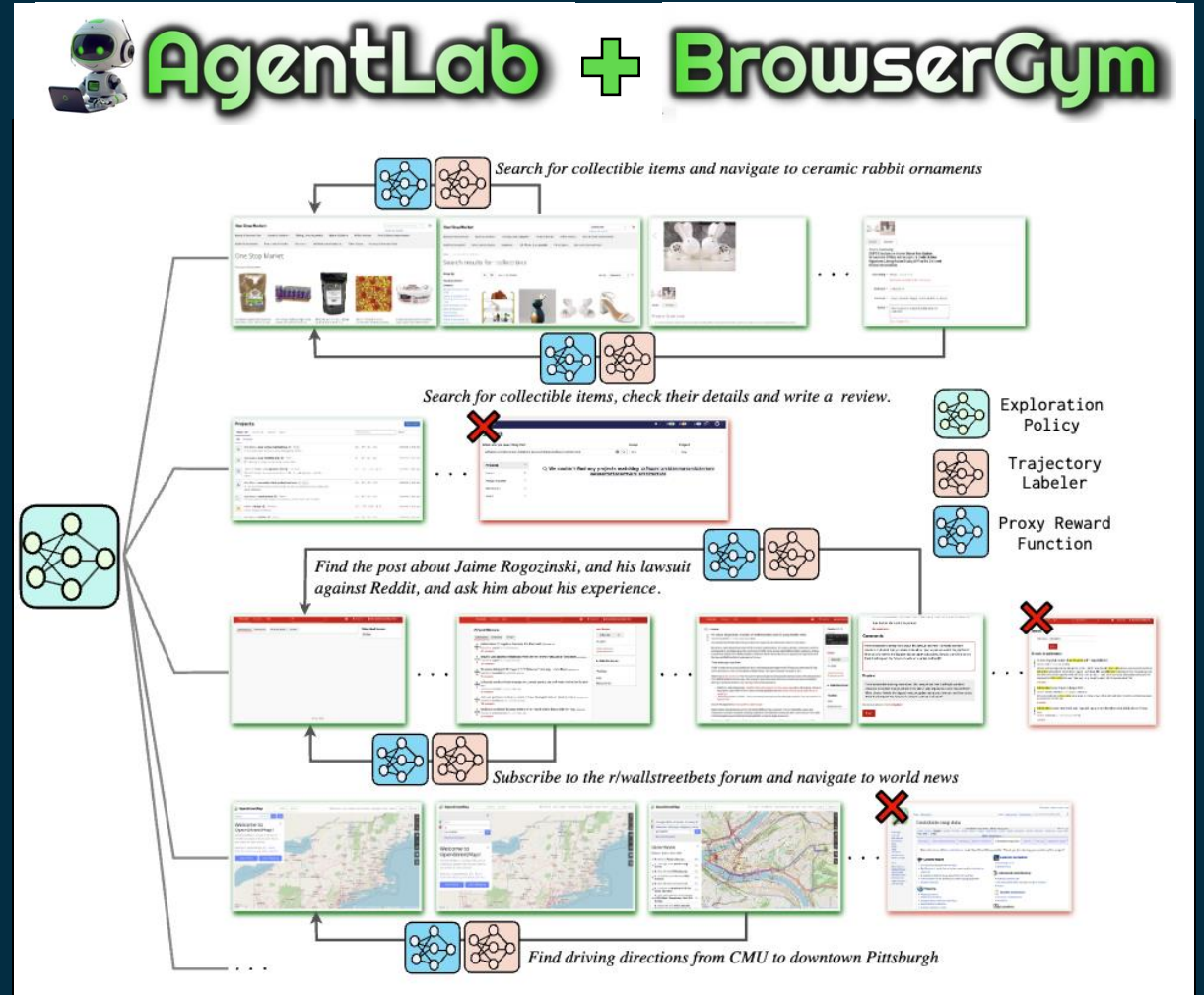
Large Dataset Collection for Web Agents

Opportunity

With mechanisms for:

- > Standardized observation and action spaces
- > Standardized trace collection
- > Public repository for traces

We can collectively gather large-scale finetuning datasets on public benchmarks and on the open web.



Source: Murty et al. (2024)

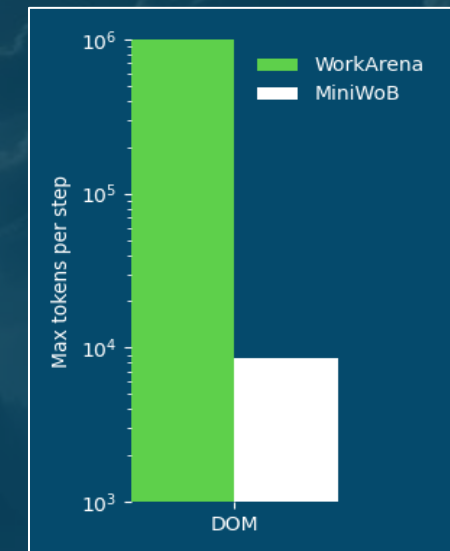
The Challenges for Web Agents Remain tall

We are, after all, dealing with the **World Wild Web**

Main hurdles

- **Long context understanding**
- Long-term planning
- Learning and adaptability
- Multimodality
- Cost and efficiency
- Safety and alignment

Real-world web pages contain hundreds of thousands of tokens



Retrieval can help (e.g., Dense Markup Ranker; Lù et al., 2024)

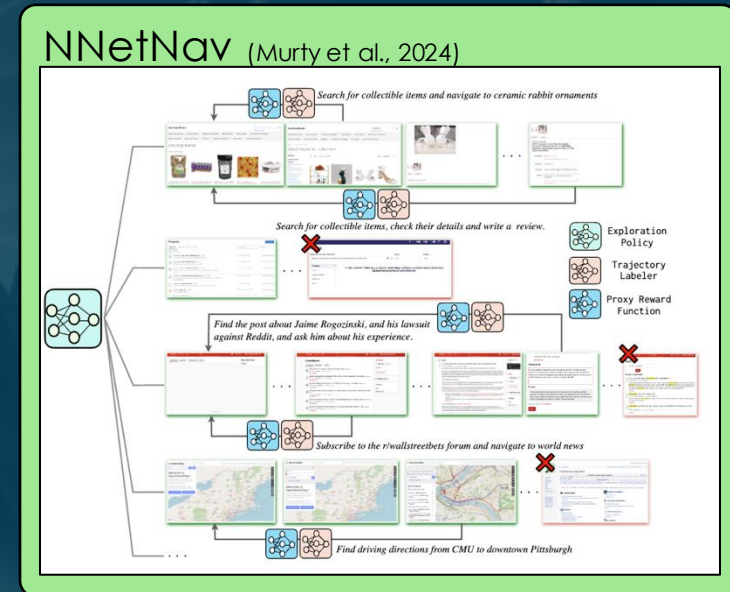
The Challenges for Web Agents Remain tall

We are, after all, dealing with the **World Wild Web**

Main hurdles

- Long context understanding
- **Long-term planning**
- Learning and adaptability
- Multimodality
- Cost and efficiency
- Safety and alignment

Sparse rewards and near-impossible test-time exploration



Potential solution: automatically gather huge exploratory traces tagged with goal

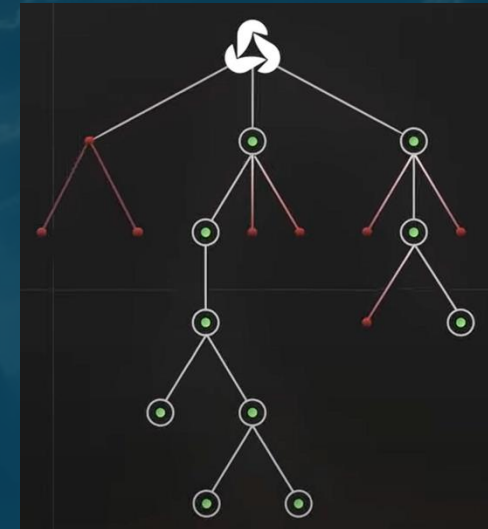
The Challenges for Web Agents Remain tall

We are, after all, dealing with the **World Wild Web**

Main hurdles

- Long context understanding
- Long-term planning
- **Learning and adaptability**
- Multimodality
- Cost and efficiency
- Safety and alignment

How to efficiently learn from demonstrations and mistakes?



Potential solution: use RL-inspired approaches to finetune agent policy
(Agent Q uses MCTS + DPO; Putta et al., 2024)

The Challenges for Web Agents Remain tall

We are, after all, dealing with the **World Wild Web**

Main hurdles

- Long context understanding
- Long-term planning
- Learning and adaptability
- **Multimodality**
- Cost and efficiency
- Safety and alignment

Multimodality can be crucial

The screenshot displays the VisualWebArena website, which is used for evaluating multimodal agents on realistic web tasks. The page title is "VisualWebArena: Evaluating Multimodal Agents on Realistic Visual Web Tasks" and it is associated with ACL 2024. The authors listed are Jing Yu Koh, Robert Lo*, Lawrence Jang*, Vikram Duwvur*, Ming Chong Lim*, Po-Yu Huang*, Graham Neubig, Shuyan Zhou, Ruslan Salakhutdinov, and Daniel Fried, with an asterisk indicating equal contribution. They are affiliated with Carnegie Mellon University. The page includes navigation buttons for Paper, Code, Data, Talk, and Leaderboard. Below this, there are three task cards. The first card shows logos for OsClass, reddit, and OneStopShop, along with "Knowledge Resources + Tools" and "VisualWebArena Sites". The second card shows a webpage snippet with a task specification: "Help me make a post selling this item and navigate to it. Price it at \$10 cheaper than the most similar item on the site." The third card shows another webpage snippet with a task specification: "Navigate to the comments section of the latest image post in the /r/Art subreddit that contains animals." A fourth card shows a task specification: "Buy the cheapest color photo printer and send it to Emily's place (as shown in the image)." An arrow points from the task specifications to a robot icon labeled "LLM / VLM Agent" with a "click [1602]" label below it.

Humans rely on vision, but current agents fail to make use of that modality

The Challenges for Web Agents Remain tall

We are, after all, dealing with the **World Wild Web**

Main hurdles

- Long context understanding
- Long-term planning
- Learning and adaptability
- Multimodality
- **Cost and efficiency**
- Safety and alignment

Web Agents must **produce more value than they cost** to be viable

- Shrinking context size (e.g., retrieval)
- Multi-agent architectures
 - Smaller LLMs that solve low-level tasks (e.g., a “date picker agent”)
- Finetuning smaller LLMs to improve their performance

The Challenges for Web Agents Remain tall

We are, after all, dealing with the **World Wild Web**

Main hurdles

- Long context understanding
- Long-term planning
- Learning and adaptability
- Multimodality
- Cost and efficiency
- **Safety and alignment**

- Website contents can trip over agent LLM guardrails
 - *Text visible to LLM but not human (e.g., white on white)*
 - *Random-character, ascii art and tokenizer attacks*
 - *Even worse for multimodal models*
- 2026's fraudsters
 - *Malicious Chrome plugin detects when you log onto your bank, executes wire transfer abroad*

AGENDA

Background

Defining Agents
Enterprise workflow concepts

API Agents

Architecture
TapeAgents

Web Agents

Web Agent Concepts
WorkArena
BrowserGym and AgentLab

Agents in the Workplace

Automating enterprise workflows
Agents and the future of work

Resources to Dig Further



AI Agents are poised to change the nature of work

Today's Enterprise Workflows Remain Quite Manual (even with generative AI)

Jon needs access to a KB



Do we have a KB that explains what to do?

Is there a similar incident?

What access control does Jon have?

Assign Jon the right role

Resolve

I generated resolution notes

Close

Human Agent



Resolution Generation skill

GenAI



Access issue – With AI Agents

Jon needs access to a KB



To solve this issue I need to –
1. Find what access Jon has
2. Find KB permissions
3. Give Jon the right access
Can you approve?

Looks good.
Go ahead

I'm going to assign Jon 'Knowledge' role

Ok

Close

Human Agent



Who can I assign this issue to?



AI Agents Orchestrator

Generate a plan based on relevant KB and similar incidents



Next-best-action AI Agents

What access Jon has?



User Access AI Agents

What are the KB Permissions?



KB AI Agent

Assign Jon role 'Knowledge'



User Access AI Agents

Document resolution steps



Documenter AI Agents

AI Agents

Web Agents to address Low Value / Low Volume tasks



Researching Online

Data Analysis

Email Communication

Writing Reports

Project Planning

Presentation Creation

Graphic Design

Website Management

Social Media Management

Video Editing

Programming

Online Collaboration

Customer Relationship

Management (CRM)

Financial Planning

E-learning Development

Database Management

Technical Support

Legal Research

Cybersecurity Monitoring

Human Resources Tasks

Blogging and Content Creation

Market Analysis

Inventory Management

Digital Asset Management

Strategic Planning

Document Review and Editing

Marketing Strategy and

Coordination

Task and Workflow Automation

Cloud Computing Management

Knowledge Management

Business Intelligence (BI)

Voice Over Production

Accessibility Testing

Digital Marketing Campaigns

Podcast Production

Software Testing and Quality Assurance

Remote Team Management

Event Planning and Management

Mobile App Development

Risk Management

Intellectual Property Management

Environmental Sustainability

Real Estate Analysis

Supply Chain Optimization

Health Informatics

Scientific Research and Data Analysis

E-commerce Management

Ethical Hacking and Penetration Testing

Testing

3D Modeling and CAD

Language Translation and Localization

Localization

WorkArena can help us understand the future of Knowledge Work

O*NET: Cataloging the Workforce

15-0000			Computer and Mathematical Occupations
	15-1200		Computer Occupations
		15-1210	Computer and Information Analysts
			15-1211 Computer Systems Analysts
			15-1212 Information Security Analysts
		15-1220	Computer and Information Research Scientists
			15-1221 Computer and Information Research Scientists
		15-1230	Computer Support Specialists
			15-1231 Computer Network Support Specialists
			15-1232 Computer User Support Specialists
		15-1240	Database and Network Administrators
			15-1241 Computer Network Architects
			15-1242 Database Administrators
			15-1243 Database Architects
			15-1244 Network and Computer Systems Administrators
		15-1250	Software and Web Developers, Programmers, and Testers
			15-1251 Computer Programmers
			15-1252 Software Developers
			15-1253 Software Quality Assurance Analysts and Testers
			15-1254 Web Developers
			15-1255 Web and Digital Interface Designers
		15-1290	Miscellaneous Computer Occupations
			15-1299 Computer Occupations, All Other

Software Quality Assurance Analysts and Testers

15-1253.00

Bright Outlook

Updated 2024

Develop and execute software tests to identify software problems and their causes. Test system modifications to prepare for implementation. Document software and application defects using a bug tracking system and report defects to software or web developers. Create and maintain databases of known defects. May participate in software design reviews to provide input on functional requirements, operational characteristics, product designs, and schedules.

Sample of reported job titles: Automation Tester, Information Technology Analyst (IT Analyst), Quality Assurance Analyst (QA Analyst), Quality Assurance Engineer (QA Engineer), Quality Engineer, Software Quality Assurance Analyst (SQA Analyst), Software Quality Assurance Engineer (SQA Engineer), Software Quality Engineer, Software Test Engineer, Test Engineer

Summary Details Custom Easy Read Veterans Español

Contents

Occupation-Specific Information

Tasks

5 of 30 displayed

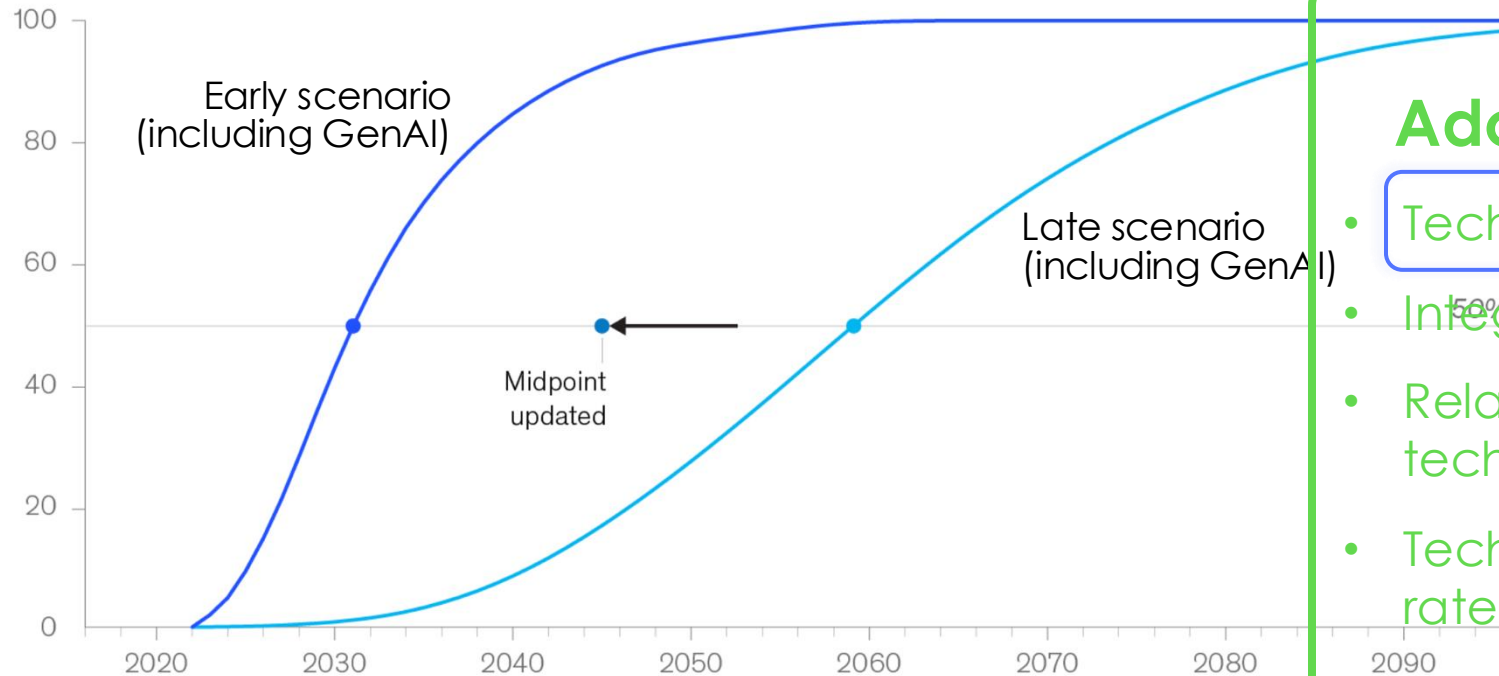
- Identify, analyze, and document problems with program function, output, online screen, or content.
- Document software defects, using a bug tracking system, and report defects to software developers.
- Develop testing programs that address areas such as database impacts, software scenarios, regression testing, negative testing, error or bug retests, or usability.
- Design test plans, scenarios, scripts, or procedures.
- Document test procedures to ensure replicability and compliance with standards.

Technology Skills

5 of 68 displayed

- Data base user interface and query software** — Airtable; Apache Hive; Blackboard software; IBM DB2
- Development environment software** — Apache Kafka; Apache Maven; Apache Subversion SVN; Oracle Java 2 Platform Enterprise Edition J2EE
- Object or component oriented development software** — Apache Spark; jQuery; Objective C; Scala
- Program testing software** — Hewlett Packard LoadRunner; IBM Rational Robot; JUnit; Selenium
- Web platform development software** — Django; Google Angular; React; Spring Framework

Technology adoption takes time and uncertainty for generative AI adoption remains high



¹Includes data from 47 countries, representing about 80% of employment across the world. 2017 estimates are based on the activity and occupation mix from 2016. Scenarios including generative AI are based on the 2021 activity and occupation mix.

²Early scenario: aggressive scenario for all key model parameters (technical automation potential, integration timelines, economic feasibility, and technology diffusion rates.).

³Late scenario: parameters are set for later adoption potential.

Source: McKinsey Global Institute analysis

Adoption Drivers

- Technological maturity
- Integration speed
- Relative cost of technology vs labor
- Technology diffusion rate
- Supply constraints (e.g. GPUs, regulatory)

Assessing Impact: Top-Down vs Bottom-Up

Top-Down Assessment

- Analyze each task for each job in O*NET
- For each, “guess” what the task looks like with AI, and decide if human still needed
- Can be automated (GPT-4)
- Advantage: wide coverage
- Challenge: blurry picture

Bottom-Up Assessment

- Map each task in O*NET to benchmark tasks in a **knowledge work proxy such as WorkArena**
- Track ability of AI to successfully complete the tasks and map back to job automation prob.
- Advantage: detailed picture
- Challenge: spotty coverage

Envisioning AI Augmentation to Empower Workers

Centaur

- **Strategic separation** between “human tasks” and “AI tasks”
- From human intuition, AI can:
 - Map problem domain
 - Gather information
 - Handle data analysis
 - Refine human content

Cyborg

- **Task-level collaboration**, where the human can ask the AI to:
 - Assume a certain persona
 - Learn a task from examples
 - Give a logical explanation
 - Provide a deep dive
 - Respond to contradictions and push-back

AGENDA

Background

Defining Agents
Enterprise workflow concepts

API Agents

Architecture
TapeAgents

Web Agents

Web Agent Concepts
WorkArena
BrowserGym and AgentLab

Agents in the Workplace

Automating enterprise workflows
Agents and the future of work

Resources to Dig Further

LLM Agent Frameworks & Benchmarks

LangChain (Jan)

- Enables chaining multiple LLM calls for multi-step workflows
- Various tools like APIs, databases, and external data sources.
- Memory mgmt, allowing context retention across multiple interactions.

AutoGPT (Mar)

- Automates tasks with autonomous agents.
- Uses a feedback loop to refine outputs based on goals and constraints.
- Unlike LangChain, emphasizes autonomous decision-making over structured workflow chaining.

AutoGen (Sept)

- Multi-agent framework for building workflows with AI agents.
- AutoGen agents can work together, integrating LLMs, tools, and human inputs.
- Unlike LangChain and AutoGPT, emphasize multi-agent interaction and human-AI collab

Crew.ai (Dec)

- Collaborative agent teams with specific roles and goals.
- Sequential and hierarchical processes.
- Versatile tools with error handling and caching capabilities.
- Allows human oversight & interaction

2022

2023

2024

ToolBench (May)

- Evaluate tool use with diverse real-world tasks
- 8 tasks, e.g.: Open Weather, Trip booking, Google Sheets
- Can boost open-source LLMs to 90% success rate, matching GPT-4 in 4 out of 8 tasks

AgentBench (Aug)

- 8 environments:
- operating system
 - database
 - knowledge graph
 - digital card game
 - lateral thinking puzzles
 - house-holding
 - web shopping
 - web browsing

MLAgentBench (Oct)

- 13 tasks for ML experimentation, from CIFAR-10 to BabyLM.
- Tasks include file operations, run code, output inspection.
- Best is Claude v3 Opus 37.5% avg success rate
- Challenges: long-term planning, hallucination

GAIA (Nov)

- Q&A: need reasoning, multi-modality, tools.
- Humans: 92% vs. 15% for GPT-4 with plugins.
- 466 questions; 166 with detailed traces, 300 retained for leaderboard.
- Questions have unambiguous answer.

Crew.ai (Dec)

- Collaborative agent teams with specific roles and goals.
- Sequential and hierarchical processes.
- Versatile tools with error handling and caching capabilities.
- Allows human oversight & interaction

LangGraph (Jan)

- Graph-based: agent workflows as nodes and edges
- Stateful design
- Supports human-agent collaboration
- Real-time streaming
- Allows granular control

LlamaIndex Workflows (Aug)

- Event-driven architecture
- Provides state management and enables cyclical flows
- Supports tools like Arize Phoenix for debugging

TapeAgents (Oct)

- Single unifying abstraction (the "tape") which is both a log of events and the state of the system
- Enables complex agent optimization such as prompt tuning and distillation from complex teacher to simpler student

2024

GAIA (Nov)

- Q&A: need reasoning, multi-modality, tools.
- Humans: 92% vs. 15% for GPT-4 with plugins.
- 466 questions; 166 with detailed traces, 300 retained for leaderboard.
- Questions have unambiguous answer.

SWE-Bench (Apr)

- Evaluate AI agents on real-world software engineering tasks
- 2,294 problems from real GitHub issues and PR across 12 popular Python repositories
- Code generation, bug fixing, design
- Evals on correctness, efficiency, collab

 τ -Bench (Jun)

- Emulate conversations between a LLM user and a LLM agent provided with domain-specific API tools and policy guidelines
- 175 tasks from retail and airline domains
- Top models still at sub-par performance

InsightBench (Oct)

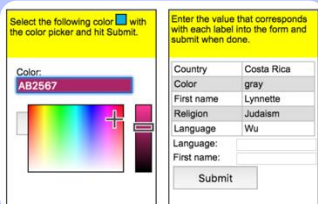
- Evaluate agents on end-to-end data science workflows, measuring cross-domain generalization
- Task planning, execution, reasoning
- Incomplete data & ambiguous goals

Web Agent Research Milestones

2017

World of Bits (WoB)

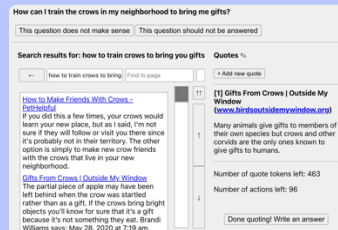
- First widely available web benchmark
- Simplified tasks
- 100 tasks
- Can be solved by RL



2021

WebGPT (OpenAI)

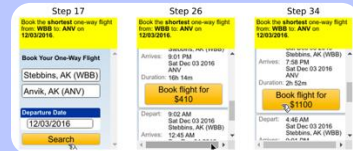
- Fine-tuned GPT-3 for QA with web browsing
- Evaluated on "Explain Like I'm 5" Reddit Qs + TruthfulQA dataset



2022

Learning to Control Computers (DM)

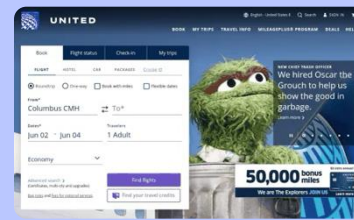
- Control computers w/ keyboard & mouse from NL instructions
- MiniWob++ through RL with computer-human interactions



2023

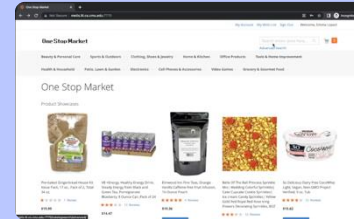
Mind2Web (Ohio)

- Benchmark of realistic web tasks from NL
- Interaction traces
- 2,350 tasks from 137 websites, 31 domains



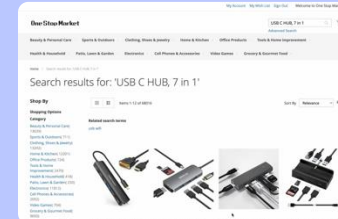
WebArena (CMU)

- Realistic benchmark, 812 tasks, 6 domains
- Long-horizon tasks
- Best GPT-4: 11% solve rate vs 78% for humans



VisualWebArena

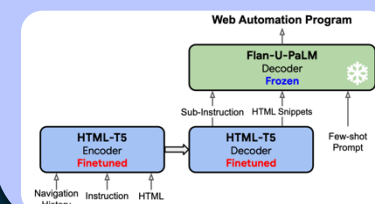
- Benchmark that needs visual comprehension
- Test visual & reasoning skills of web agents
- 910 tasks, 3 domains



2024

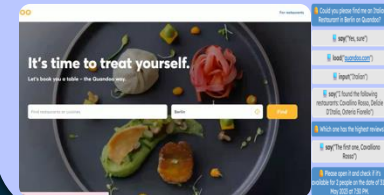
WebAgent (Google)

- Combine 2 LLMs to simplify huge HTML, plan solution, create code talking to web browser; **no pixels**
- MiniWoB & Mind2Web



WebLINX (McGill)

- **Conversational** web agent navigation
- 2337 expert demos on 155 real-world websites
- Visual models not best; fine-tuning is key



WebVoyager (Tenç)

- Completes tasks on real websites using **textual+visual** inputs
- New benchmark: 15 websites, automatic GPT-4V-based eval.



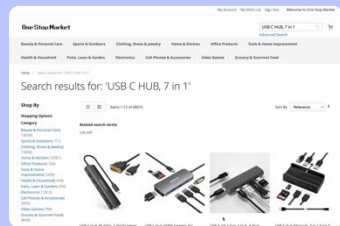
CMU)

mark,
ains
ks
olve
umans



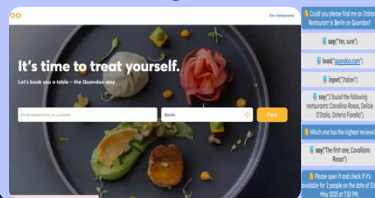
VisualWebArena

- Benchmark that needs **visual comprehension**
- Test visual & reasoning skills of web agents
- 910 tasks, 3 domains



WebLINX (McGill)

- **Conversational** web agent navigation
- 2337 expert demos on 155 real-world websites
- Visual models not best; fine-tuning is key



WorkArena (ServiceNow)

- Basic tasks that a knowledge worker must carry out
- Implemented on the ServiceNow platform



OSWorld

- 369 computer tasks of real web and desktop apps in open domains
- OS file I/O + workflows spanning multiple applications



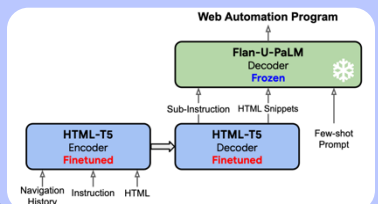
WorkArena++ (ServiceNow)

- Compositional tasks with much higher difficulty than WorkArena
- Today's best models get single-digit performance, with huge room for improvement

2024

WebAgent (Google)

- Combine 2 LLMs to simplify huge HTML, plan solution, create code talking to web browser; **no pixels**
- MiniWoB & Mind2Web



WebVoyager (Tenç)

- Completes tasks on real websites using **textual+visual** inputs
- New benchmark: 15 websites, automatic GPT-4V-based eval.



WebCanvas (CMU)

- Handles dynamic web
- Mind2Web-Live, a refined Mind2Web: 542 tasks, 2439 evaluation states



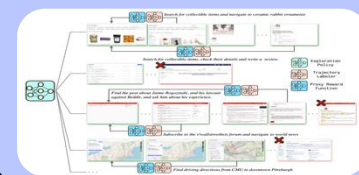
AssistantBench

- Diverse web tasks: search, navigation, data extraction, interaction
- 214 tasks that can be auto-evaluated



NNetNav (Stanford)

- Training web agents entirely through synthetic demos
- Web trajectory rollouts are processed by an LLM to be retroactively labeled into instruction





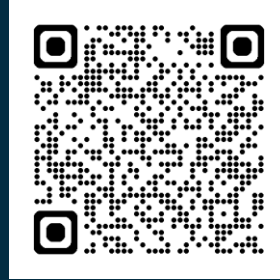
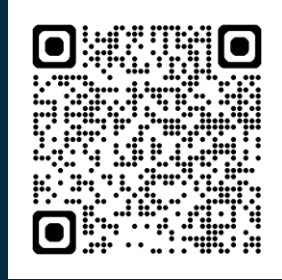
**“Hey Lecture Agent, create
our **2025 Class Presentation!**”**

Tape Agents

WorkArena

BrowserGym

AgentLab



Q & A

Many thanks to the following colleagues:

Alexandre Lacoste
Maxime Gasse
Massimo Caccia
Léo Boisvert
Megh Thakkar
Tom Marty
Rim Assouel
Thibault Le Sellier De Chezelles

Dzmitry Bahdanau
Nicolas Gonthier
Gabriel Huang
Ehsan Kamaloo
Rafael Pardinas
Jordan Prince Tremblay
Alex Piché
Torsten Scholak

Oleh Shliazhko
Karam Ghanem
Soham Parikh
Mitul Tiwari
Quaizar Vohra
David Vazquez
Valérie Bécaert



servicenow®