What are Agents?

CS6960 MultiModal LLM Agents

Kenneth Marino

Slide credits: Farhan Ishmam, …

Quick Announcements

- Add/Drop Course
  - Class is pretty much at capacity
  - If you already know you will drop, please do that ASAP so people can get in off the waiting list
- HW0 due next week
  - It’s really short and easy
  - Basically just asking you to read the intro unit to the HF Agents tutorial and write some boilerplate code

About Me

About Me

- New Faculty at Kahlert School of Computing

About Me

- New Faculty at Kahlert School of Computing
- Research Scientist at DeepMind
  - Incorporating VLM/LLM into agents
  - Retrieval/RAG for personalization in LLMs

About Me

- New Faculty at Kahlert School of Computing
- Research Scientist at DeepMind
  - Incorporating VLM/LLM into agents
  - Retrieval/RAG for personalization in LLMs
- PhD from Carnegie Mellon
  - Thesis: Towards knowledge-capable AI: Agents that See, Speak, Act and Know 3

TAs

Research Interests

- Machine Learning Security
- Agentic systems and Security

Current Research Project

- Backdoor Injection in Tool-Calling LLMs.
- Assess how prompt-triggered backdoors affect tool-calling behavior in modern chat LLMs.
- Build a defense system in accordance with the vulnerability

Syeda Mishra Saiara
Teaching Assistant

Final Year MS Student
Kahlert School of Computing,
SaLT Lab

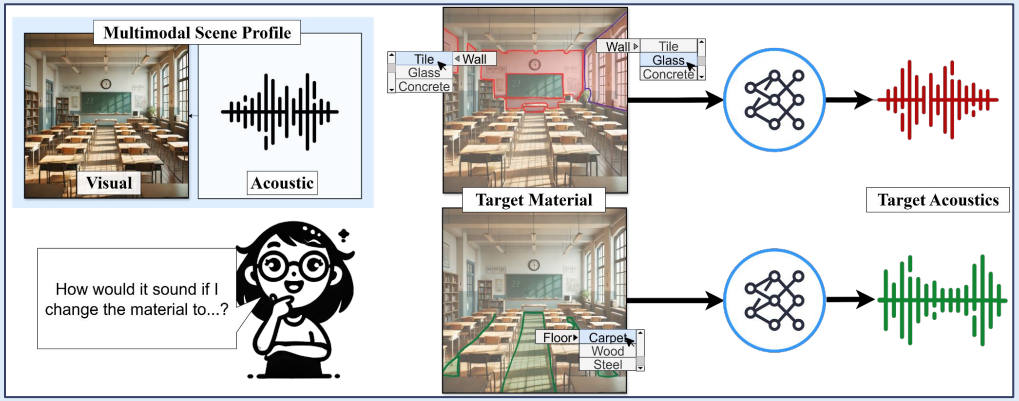
Mahnoor Saad

Teaching Assistant

3rd-year PhD student (CS)
Kahlert School of Computing
Computer Vision Lab

Current Research

My research interests are centered around computer vision and machine learning, with a particular focus on multimodal perception and audio-visual learning.



- Multimodal scene perception
- Audio-Visual acoustic profile estimation
- Material-Conditioned impulse response synthesis
- Acoustic waveform propagation modeling
- User-defined, realistic acoustic simulation

My Area of Research

- VLM / Multimodal LLM Agents

My Area of Research

- VLM / Multimodal LLM Agents
  - Will get to what that means in a bit

My Area of Research

- VLM / Multimodal LLM Agents
  - Will get to what that means in a bit
- Wanted to teach a class on this topic
  - New/Exciting area and want to get people excited about it / doing research in it

Goals for this Class

- Understand what we mean by agents
- Learn the fundamentals needed to build VLM Agents
  - Reinforcement Learning / Environments
  - Large Language Models
  - Vision Language Models
- Learn major topic areas for VLM Agents
  - Frameworks, Tools, RAG, Evaluation, Coding, Computer Use, …
- Actually make agents / do a full project!

Organization of Course

- Start with a few lectures from me covering fundamentals before we can start with Agents
  - Reinforcement Learning / Environments
  - Large Language Models
  - Vision Language Models
- A few mini-lectures on agent topics from me
- Most of the course material is paper readings
  - ~60 papers on wide variety of agent topics me/TAs have selected (but please suggest papers to me that you want to read)
  - Area is fast moving (when I teach this course in a year, it'll probably have completely changed again)
- Most of work is project based

What you need to do for this Class

- Homeworks (10%)
  - Completing HF Agents Tutorial coding exercises
  - Fairly short/easy. Mostly to help you get comfortable with coding agents right away before you start on projects.
- In-class Paper presentations (30%)
  - Each student reads and researches (aim for 15 minutes)
  - Paper signup list / instructions to come (next week)
  - Goal: field is moving quickly, reading papers is the best way
- Participation (5%)
- Course Project (55%)
  - Most of the work in the course (I want you to spend most of your time on this)
  - Goal: Actually learn by doing research

Course Project

- Groups of 2-3
- Choose groups / topic in first few weeks of class
- Project proposal due ~midpoint of semester
- Two milestone reports due throughout the semester
  - Will do peer feedback in class and light instructor/TA feedback
- Final Project / Presentation
  - Write up of project, short presentation to class
- Compute Resources - have class code if you do not already have (if you are KSC student, you should already have cluster access)

Course Schedule

	D1 S1	D1 S2	D1 S3	D1 S4	D12S1	D2 S2	D2 S3	D2 S4	HW/Project
Week 1	Lecture - Course Intro				Lecture - RL Basics				HW0 - Setting up an Agent
Week 2	Lecture - LLM Basics			FREE - Sign up	Lecture - VLM Basics				
Week 3	Mini Lecture - Frameworks		Frameworks Pa	Frameworks Pa	Frameworks Pa	Frameworks Pa	Frameworks Pa	FREE - Project Brainstorm	HW1 - LLM Agent Frameworks
Week 4	Mini Lecture - RAG		RAG Paper 1	RAG Paper 2	RAG Paper 3	RAG Paper 4	RAG Paper 5	FREE - Project Brainstorm	
Week 5	Mini Lecture - Tool Use		Tool Use Paper	Tool Use Paper	Tool Use Paper	Tool Use Paper	Tool Use Paper	FREE - Project Group Form	HW 2 - RAG Agent
Week 6	Mini Lecture - Code Agents		Code Agents Pa	Code Agents Pa	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project Group Forming	
Week 7	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project F	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project Proposal Gi	(Optional) HW 3 - Assistant Agent
Week 8	Mini Lecture - Evaluation		Evaluation Pape	Evaluation Pape	Evaluation Pape	Evaluation Pape	Evaluation Pape	FREE - Project Proposal Gi	Project Proposals Due
Week 9	Mini Lecture - Assistant Agents		Assistant Paper	Assistant Paper	Assistant Paper	Assistant Paper	Assistant Paper	FREE - Project Meetings	
Week 10	Spring Break!								
Week 11	Mini Lecture - Game Agents		Game Agents Pa	Game Agents Pa	Game Agents Pa	Game Agents Pa	Game Agents Pa	FREE - Project Report Fee	Project Report 1
Week 12	Mini Lecture - Computer Use		Computer Use P	Computer Use P	Computer Use P	Computer Use P	Computer Use P	FREE - Project Meetings	
Week 13	Computer Use P	Computer Use P	Computer Use P	FREE - Project M	Computer Use P	Computer Use P	Computer Use P	FREE - Project Report Fee	Project Report 2
Week 14	Mini Lecture - Robotics		Robotics Paper	Robotics Paper	Robotics Paper	Robotics Paper	Robotics Paper	FREE - Project Meetings	
Week 15	Robotics Paper (	Robotics Paper	Robotics Paper	FREE - Project M	Robotics Paper	Robotics Paper	Robotics Paper	FREE - Project Report Feedback	
Week 16	Final Presentatic	Final Presentations (part 2) - During Final period							Final Report Due

Course Schedule

	D1 S1	D1 S2	D1 S3	D1 S4	D12S1	D2 S2	D2 S3	D2 S4	HW/Project
Week 1	Lecture - Course Intro				Lecture - RL Basics				HW0 - Setting up an Agent
Week 2	Lecture - LLM Basics			FREE - Sign up	Lecture - VLM Basics				

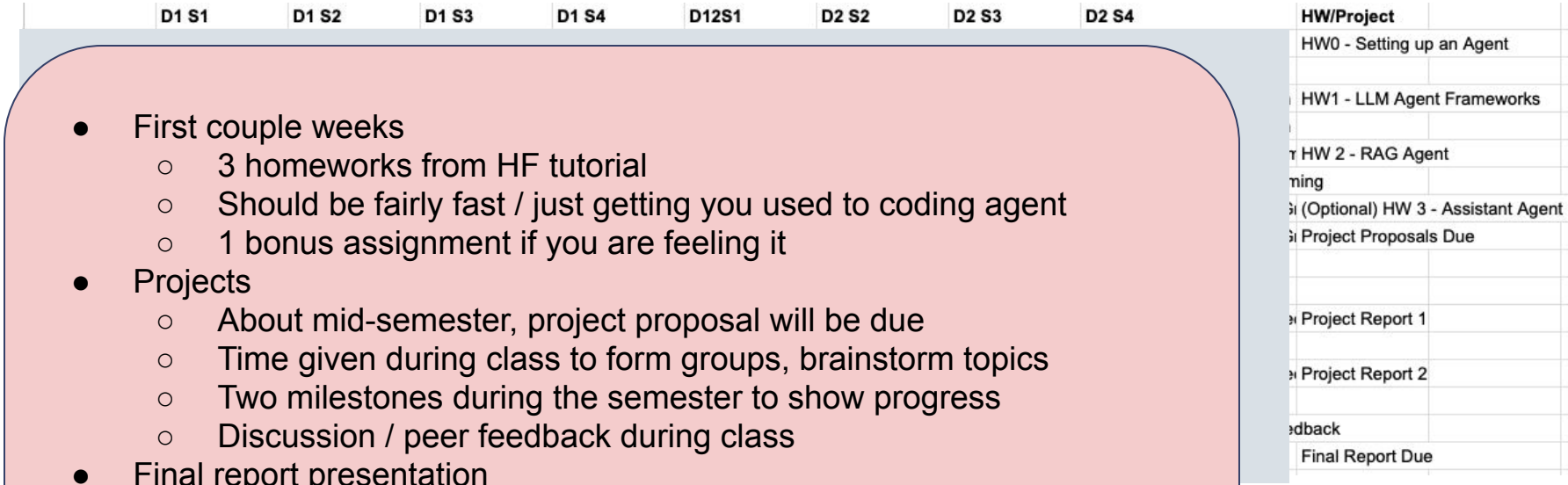
- First Two weeks: Crash course in the table stakes for multimodal agents
- For some this may just be review of AI/NLP but for others this may be first time you see this material
  - If the material is new to you, you may want to do some reading on your own (I will give some useful resources to help you)
  - Especially if you are taking this as an undergrad, this may feel very fast paced
  - Me/TAs are here to help
  - By the time you start on your projects, hopefully you have at least a basic understanding

Course Schedule

	D1 S1	D1 S2	D1 S3	D1 S4	D12S1	D2 S2	D2 S3	D2 S4	HW/Project
Week 3	Mini Lecture - Frameworks		Frameworks Pap	Frameworks Pap	Frameworks Pap	Frameworks Pap	Frameworks Pap	FREE - Project Brainstorm	HW1 - LLM Agent Frameworks
Week 4	Mini Lecture - RAG		RAG Paper 1	RAG Paper 2	RAG Paper 3	RAG Paper 4	RAG Paper 5	FREE - Project Brainstorm	
Week 5	Mini Lecture - Tool Use		Tool Use Paper	Tool Use Paper	Tool Use Paper	Tool Use Paper	Tool Use Paper	FREE - Project Group Form	HW 2 - RAG Agent
Week 6	Mini Lecture - Code Agents		Code Agents Pa	Code Agents Pa	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project Group Forming	
Week 7	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project F	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project Proposal Gi	(Optional) HW 3 - Assistant Agent
Week 8	Mini Lecture - Evaluation		Evaluation Pape	Evaluation Pape	Evaluation Pape	Evaluation Pape	Evaluation Pape	FREE - Project Proposal Gi	Project Proposals Due
Week 9	Mini Lecture - Assistant Agents		Assistant Paper	Assistant Paper	Assistant Paper	Assistant Paper	Assistant Paper	FREE - Project Meetings	
Week 10	Spring Break!								
Week 11	Mini Lecture - Game Agents		Game Agents Pa	Game Agents Pa	Game Agents Pa	Game Agents Pa	Game Agents Pa	FREE - Project Report Fee	Project Report 1
Week 12	Mini Lecture - Computer Use		Computer Use P	Computer Use P	Computer Use P	Computer Use P	Computer Use P	FREE - Project Meetings	
Week 13	Computer Use P	Computer Use P	Computer Use P	FREE - Project M	Computer Use P	Computer Use P	Computer Use P	FREE - Project Report Fee	Project Report 2
Week 14	Mini Lecture - Robotics		Robotics Paper	Robotics Paper	Robotics Paper	Robotics Paper	Robotics Paper	FREE - Project Meetings	
Week 15	Robotics Paper (	Robotics Paper	Robotics Paper	FREE - Project M	Robotics Paper	Robotics Paper	Robotics Paper	FREE - Project Report Feedback	

- Most of course: will have a mini-lecture to give a bit of an overview on the topic
- Most of the meat of these weeks will be student presentations on important papers in the area

Course Schedule



D1 S1	D1 S2	D1 S3	D1 S4	D12S1	D2 S2	D2 S3	D2 S4	HW/Project
								HW0 - Setting up an Agent
								HW1 - LLM Agent Frameworks
								HW 2 - RAG Agent
								ming
								(Optional) HW 3 - Assistant Agent
								Project Proposals Due
								Project Report 1
								Project Report 2
								edback
								Final Report Due

- First couple weeks
  - 3 homeworks from HF tutorial
  - Should be fairly fast / just getting you used to coding agent
  - 1 bonus assignment if you are feeling it
- Projects
  - About mid-semester, project proposal will be due
  - Time given during class to form groups, brainstorm topics
  - Two milestones during the semester to show progress
  - Discussion / peer feedback during class
- Final report presentation
  - Due last week of regular classes. Final class and final period will be used for short project presentations

Course Schedule

	D1 S1	D1 S2	D1 S3	D1 S4	D12S1	D2 S2	D2 S3	D2 S4	HW/Project
Week 1	Lecture - Course Intro				Lecture - RL Basics				HW0 - Setting up an Agent
Week 2	Lecture - LLM Basics			FREE - Sign up	Lecture - VLM Basics				
Week 3	Mini Lecture - Frameworks		Frameworks Pa	Frameworks Pa	Frameworks Pa	Frameworks Pa	Frameworks Pa	FREE - Project Brainstorm	HW1 - LLM Agent Frameworks
Week 4	Mini Lecture - RAG		RAG Paper 1	RAG Paper 2	RAG Paper 3	RAG Paper 4	RAG Paper 5	FREE - Project Brainstorm	
Week 5	Mini Lecture - Tool Use		Tool Use Paper	Tool Use Paper	Tool Use Paper	Tool Use Paper	Tool Use Paper	FREE - Project Group Forr	HW 2 - RAG Agent
Week 6	Mini Lecture - Code Agents		Code Agents Pa	Code Agents Pa	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project Group Forming	
Week 7	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project F	Code Agents Pa	Code Agents Pa	Code Agents Pa	FREE - Project Proposal Gi	(Optional) HW 3 - Assistant Agent
Week 8	Mini Lecture - Evaluation		Evaluation Pape	Evaluation Pape	Evaluation Pape	Evaluation Pape	Evaluation Pape	FREE - Project Proposal Gi	Project Proposals Due
Week 9	Mini Lecture - Assistant Agents		Assistant Paper	Assistant Paper	Assistant Paper	Assistant Paper	Assistant Paper	FREE - Project Meetings	
Week 10	Spring Break!								
Week 11	Mini Lecture - Game Agents		Game Agents Pa	Game Agents Pa	Game Agents Pa	Game Agents Pa	Game Agents Pa	FREE - Project Report Fee	Project Report 1
Week 12	Mini Lecture - Computer Use		Computer Use P	Computer Use P	Computer Use P	Computer Use P	Computer Use P	FREE - Project Meetings	
Week 13	Computer Use P	Computer Use P	Computer Use P	FREE - Project M	Computer Use P	Computer Use P	Computer Use P	FREE - Project Report Fee	Project Report 2
Week 14	Mini Lecture - Robotics		Robotics Paper	Robotics Paper	Robotics Paper	Robotics Paper	Robotics Paper	FREE - Project Meetings	
Week 15	Robotics Paper	Robotics Paper	Robotics Paper	FREE - Project M	Robotics Paper	Robotics Paper	Robotics Paper	FREE - Project Report Feedback	
Week 16	Final Presentatic	Final Presentations (part 2) - During Final period							Final Report Due

- Schedule may definitely shift around
- Spare class time to be used for project discussion / working time
- Questions?

Any Questions

What are Agents?

- VLM Agents

What are Agents?

- VLM Agents

What Even is an Agent?

What are Agents?

- VLM Agents

Towards Knowledge-capable AI:
Agents that See, Speak, Act and Know

Kenneth Marino

July 2021
CMU-ML-21-108

What do words mean?

- VLM Agents

Towards Knowledge-capable AI:
Agents that See, Speak, Act and Know

Kenneth Marino

July 2021
CMU-ML-21-108

What do words mean?

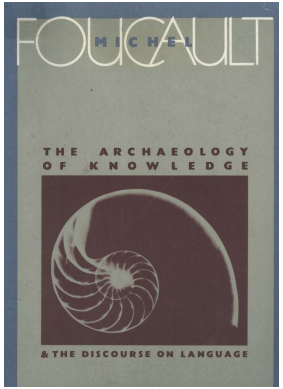
- Mistake - Asking Philosophers

2.1 Knowledge in Philosophy

The subject of knowledge: what it is, how it's acquired and how it's used is a major topic in the field of Philosophy. Such a major topic in fact that not only is the field too big to be contained in a Philosophy PhD thesis, it is too big even for entire books to fully capture. As the author of this thesis is not a philosopher, we will not attempt anything except a very brief overview of epistemology, the philosophical study of knowledge. It will mostly be a summary of the relevant entries of the Stanford Encyclopedia of Philosophy [392], the Encyclopedia Britannia and other sources which we will note.

Most summaries of the history of epistemology start with Plato. The word epistemology in fact comes from the Greek words "episteme" and "logos" meaning knowledge and knowledge/understanding and argument/reason. Plato's epistemology looked at what it means to "know" and how acquiring knowledge is a moral virtue in and of itself [392]. Plato's understanding of knowledge contrasted sensory experience from knowledge, saying in Theaetetus that "sense experience cannot be a source of knowledge, because the objects apprehended through it are subject to change." [47]. This idea of a separation of experience and knowledge is still influential, including in this thesis. We often make a distinction between "experience" or "data," things which are directly experienced by an agent as not being knowledge (although knowledge

What do words mean?

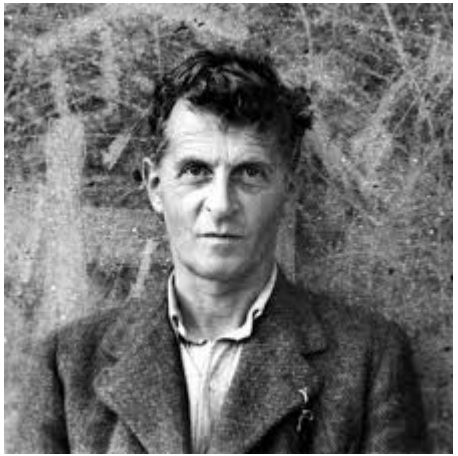


An actual *single* sentence from Foucault - the most "readable" postmodernist

Knowledge is that of which one can speak in a discursive practice, and which is specified by that fact: the domain constituted by the different objects that will or will not acquire a scientific status (the knowledge of psychiatry in the nineteenth century is not the sum of what was thought to be true, but the whole set of practices, singularities, and deviations of which one could speak in psychiatric discourse); knowledge is also the space in which the subject may take up a position and speak of the objects with which he deals in his discourse (in this sense, the knowledge of clinical medicine is the whole group of functions of observation, interrogation, decipherment, recording, and decision that may be exercised by the subject of medical discourse); knowledge is also the field of coordination and subordination of statements in which concepts appear, and are defined, applied and transformed (at this level, the knowledge of Natural History, in the eighteenth century, is not the sum of what was said, but the whole set of modes and sites in accordance with which one can integrate each new statement with the already said); lastly, knowledge is defined by the possibilities of use and appropriation offered by discourse (thus, the knowledge of political economy, in the Classical period, is not the thesis of the different theses sustained, but the totality of its points of articulation on other discourses or on other practices that are not discursive).

The one good Philosopher

"The meaning of a word is its use" - Ludwig Wittgenstein



If we go to twitter…

If we go to twitter…

Sort by latest

If we go to twitter…

If we go to twitter…

If we go to twitter…

Louis Amira @louisamira · 2
Agentic AI has arrived.

VentureVerse (Formerly OG Club) @VXV
Replying to @VXVHub
TL;DR:

- Ethereum is formalizing **AI** infra at the proto
- Google is wiring up **AI agents** with stableco

The rails are forming for **AI agents** to not just operate in autonomous loops.

The AI-crypto flywheel is only starting.

1

DGMD.6529 @DGMD22 · 2m
One of the reddest days, and $Tibbir is green with an impeccable chart. Stealth launch by Ribbit Capital with the right partners and tech to be the 'Google moment' of agentic **AI**, the ability to actually put **agents** to work with vetted accuracy in a multitude of tasks. New fintech.

30

If we go to twitter…

AI Systems group - Closed

Tonmoy Islam | ⌘ 🛠️🏄 /acc🟠,💊 @tonmoyislam03 · 4m

The Hype is Real: Why I'm Bullish on Infinity Ground

Caught wind of @infinityg_ai and had to share—Infinity Ground is quietly dropping bombs in the **AI**-Web3 space. This pioneering IDE uses multi-**agents** to let you build blockchain apps in minutes, no coding required. Describe your

Show more

InfinityGround and Galxe

1

VentureVerse

Vent

Repl

TL;D

- Eth

- Go

The

opei

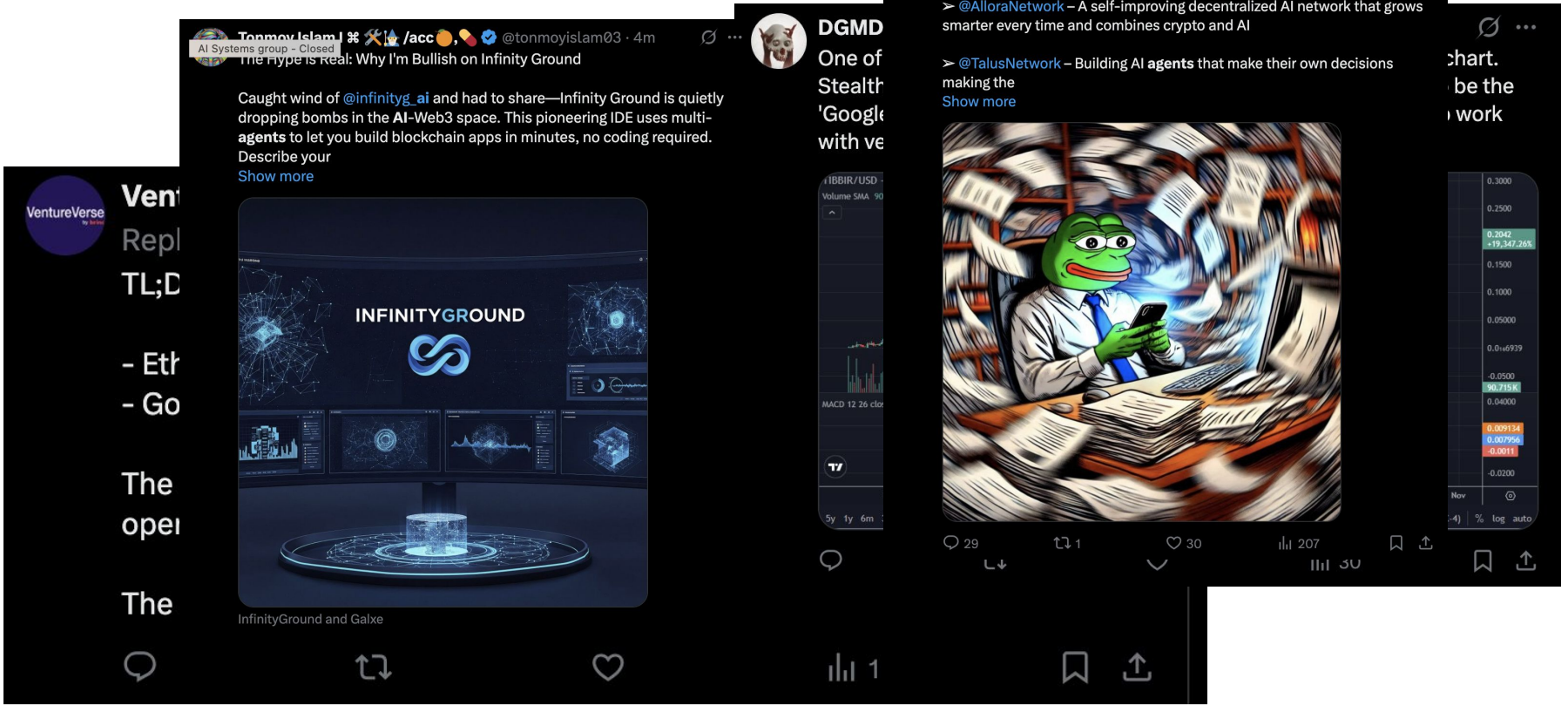
The

DGMD.6529 @DGMD22 · 2m

One of the reddest days, and $Tibbir is green with an impeccable chart. Stealth launch by Ribbit Capital with the right partners and tech to be the 'Google moment' of agentic **AI**, the ability to actually put **agents** to work with vetted accuracy in a multitude of tasks. New fintech.

30

If we go to twitter…



Definition 1: Grifting

Definition 1: Grifting

- Buzzword?

Definition 1: Grifting

- Buzzword?
- Getting Rich with Crypto?

Let’s go up one level of strawman

Definition 2: More autonomous AI?

Definition 2: More autonomous AI?

- Chatbots that can do things

Definition 2: More autonomous AI?

- Chatbots that can do things
  - Talk about LLMs that have access to tools: coding, web search, databases, etc

Definition 2: More autonomous AI?

- Chatbots that can do things
  - Talk about LLMs that have access to tools: coding, web search, databases, etc
- Chatbot goes and does stuff on its own without you

Definition 2: More autonomous AI?

- Chatbots that can do things
  - Talk about LLMs that have access to tools: coding, web search, databases, etc
- Chatbot goes and does stuff on its own without you
  - Give it a single query

Definition 2: More autonomous AI?

- Chatbots that can do things
  - Talk about LLMs that have access to tools: coding, web search, databases, etc
- Chatbot goes and does stuff on its own without you
  - Give it a single query
  - Does a bunch of stuff on its own

Definition 2: More autonomous AI?

- Chatbots that can do things
  - Talk about LLMs that have access to tools: coding, web search, databases, etc
- Chatbot goes and does stuff on its own without you
  - Give it a single query
  - Does a bunch of stuff on its own
  - Comes back to you when it’s done something

Definition 2:

A thing that can be started up and keep going independently

- Chatbots that can do things
  - Talk about LLMs that have access to tools: coding, web search, databases, etc
- Chatbot goes and does stuff on its own without you
  - Give it a single query
  - Does a bunch of stuff on its own
  - Comes back to you when it’s done something

Behold: an agent!

Definition 2: Way we are meant to see AI

Definition 2: Way we are meant to see AI

- An agent is something that has agency

Definition 2: Way we are meant to see AI

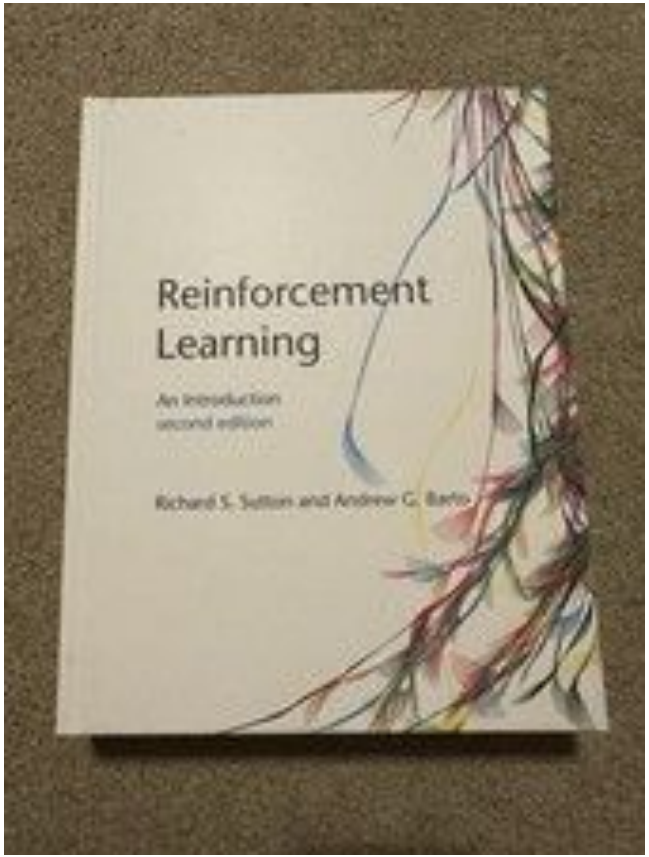
- An agent is something that has agency
- Notion of autonomy - acting without our direct intervention

Definition 2: Way we are meant to see AI

- An agent is something that has agency
- Notion of autonomy - acting without our direct intervention
- This can also have a darker twist
  - Something to directly replace humans

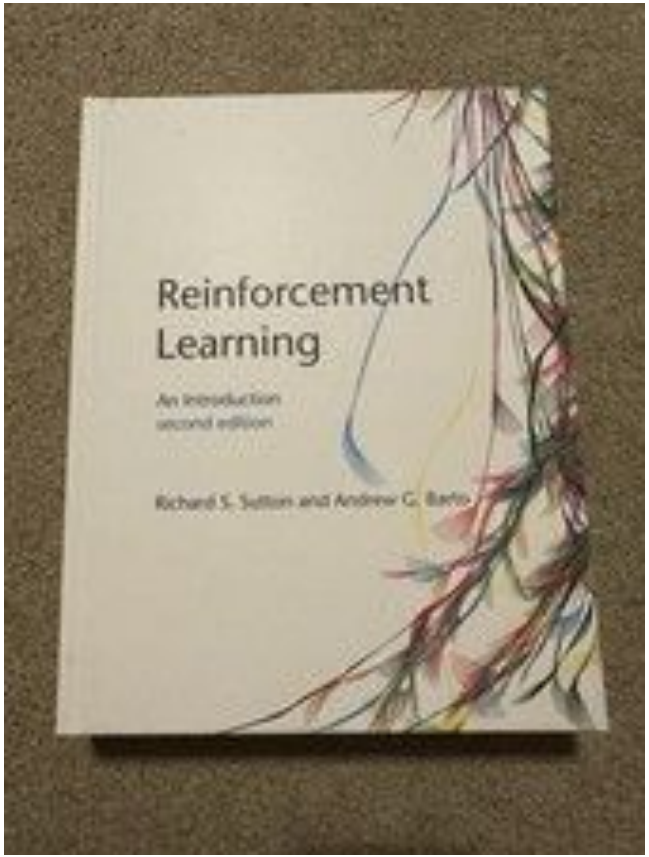
Definition 3: Classical

Definition 3: Classical

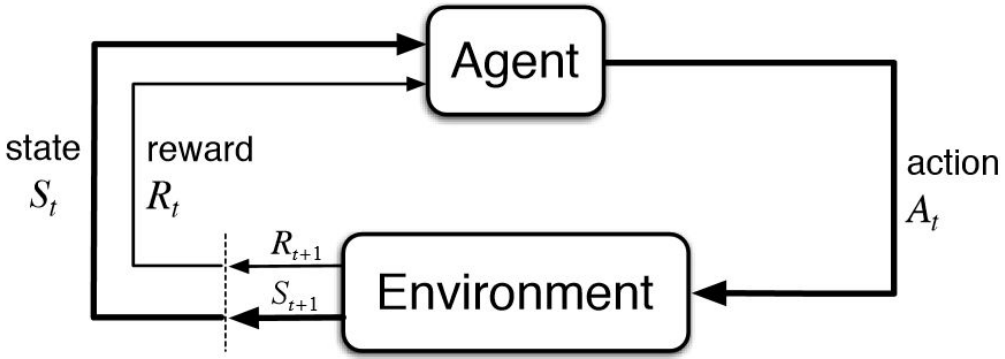


Definition 3: Classical

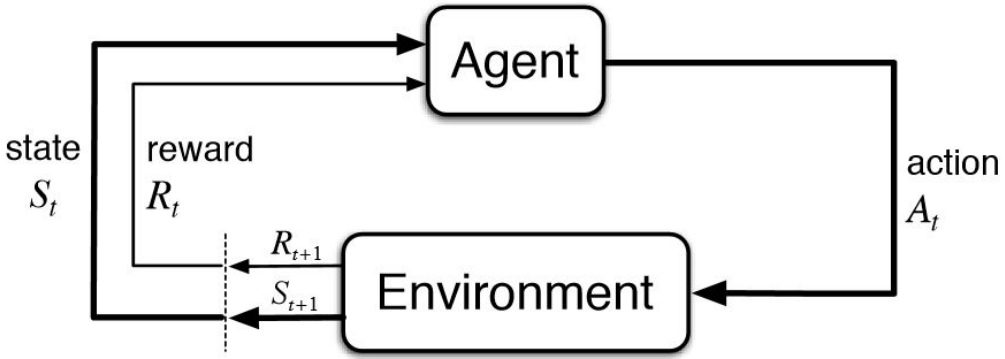
- Sutton and Barto: "The learner and decision maker is called the agent. The thing it interacts with, comprising everything outside the agent, is called the environment."

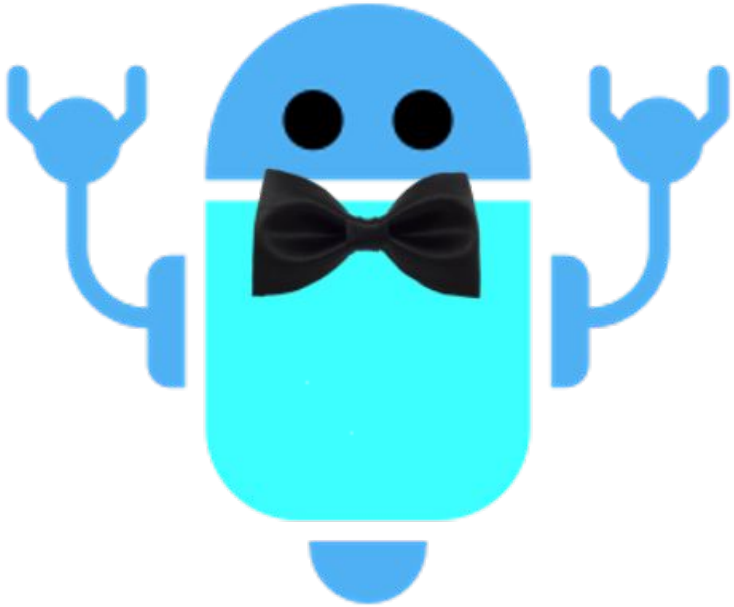


VLM Agents



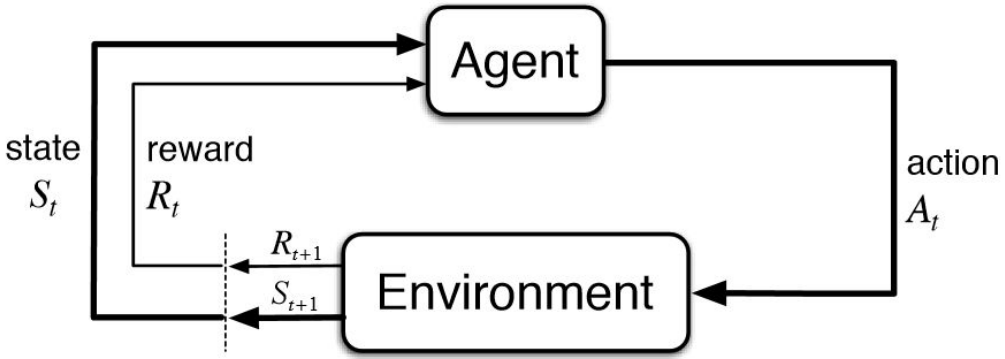
VLM Agents

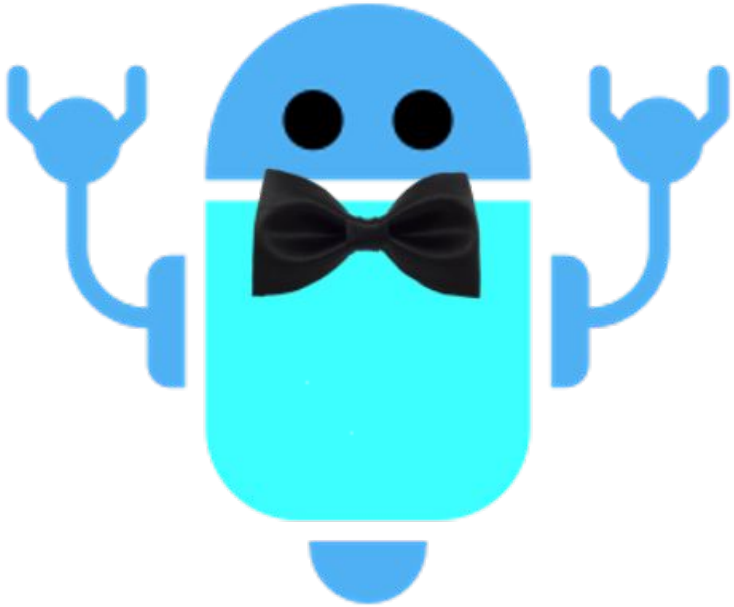




Agent

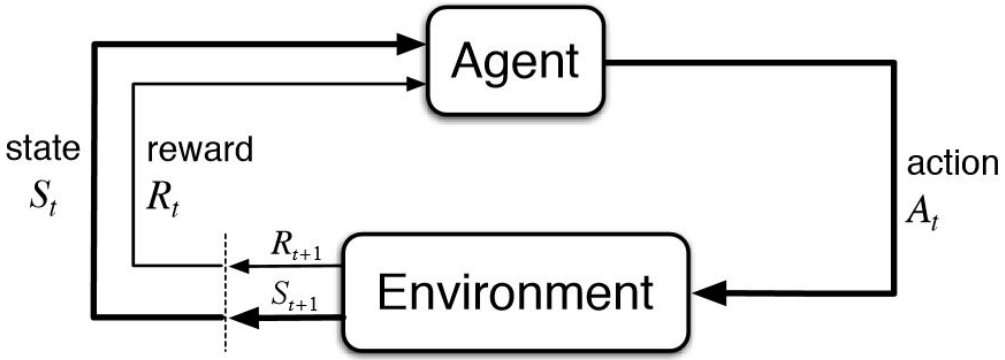
VLM Agents

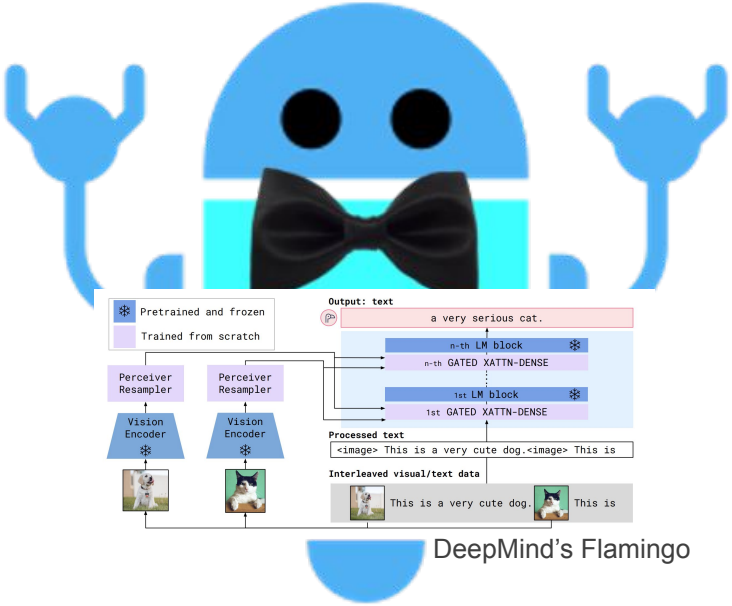




Agent

VLM Agents

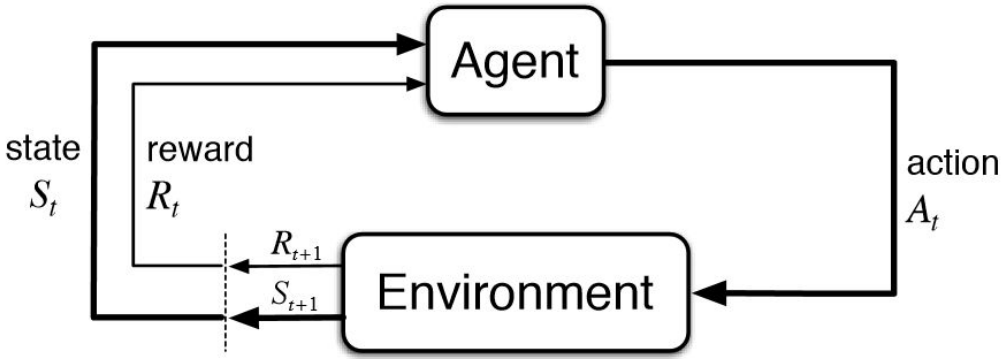




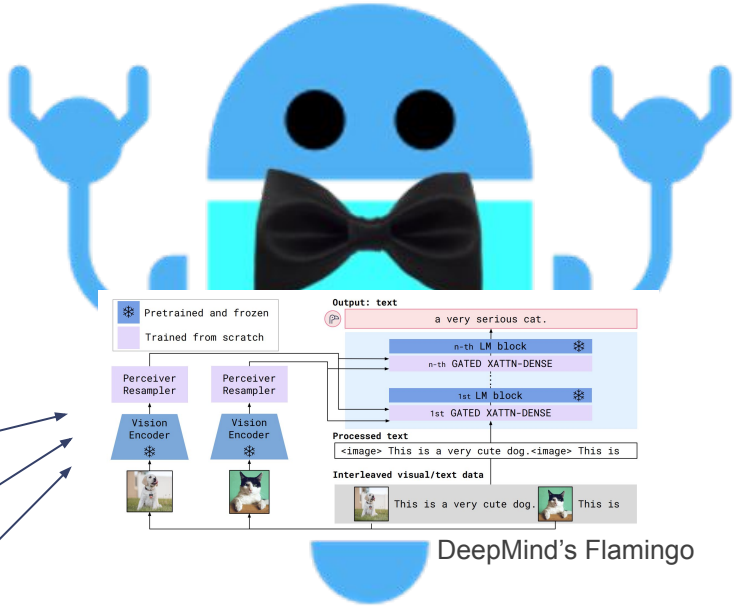
DeepMind's Flamingo

Agent

VLM Agents



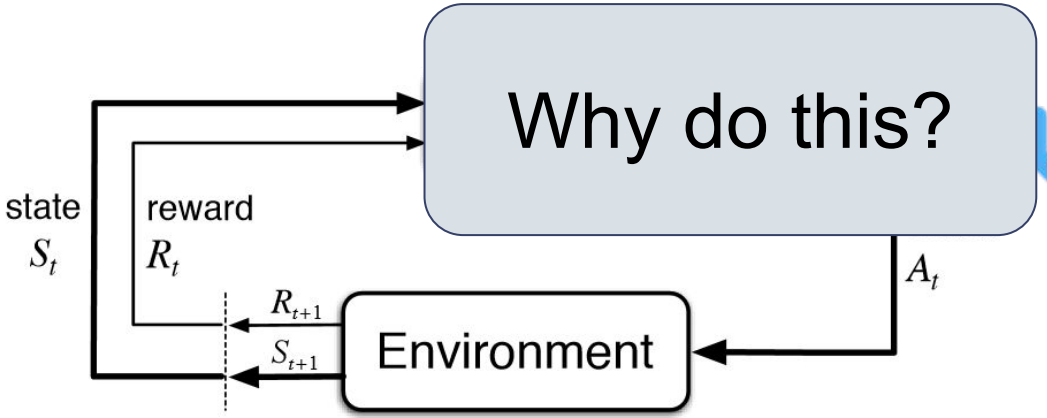
Text

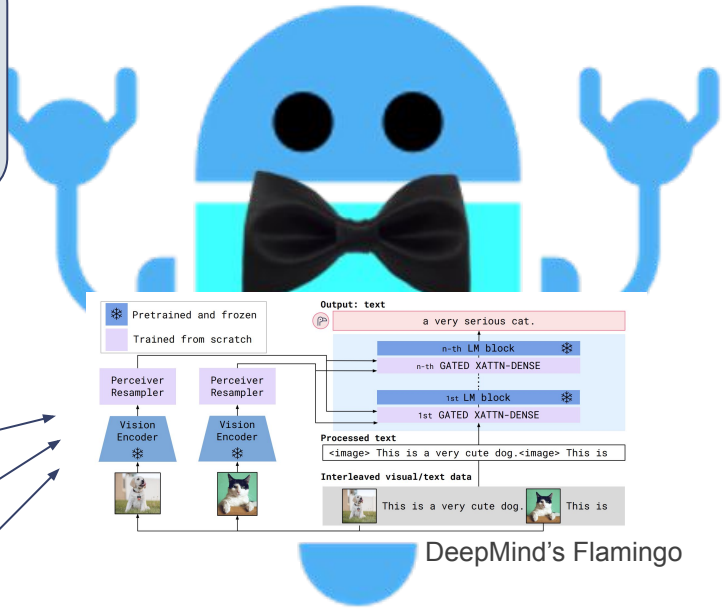


DeepMind's Flamingo

Agent

VLM Agents





DeepMind's Flamingo

Text

Agent

Where do Language Models help Decision Making

Where do Language Models help Decision Making

- Language as a built-in abstraction for classical decision making

ASK YOUR HUMANS: USING HUMAN INSTRUCTIONS TO IMPROVE GENERALIZATION IN REINFORCEMENT LEARNING

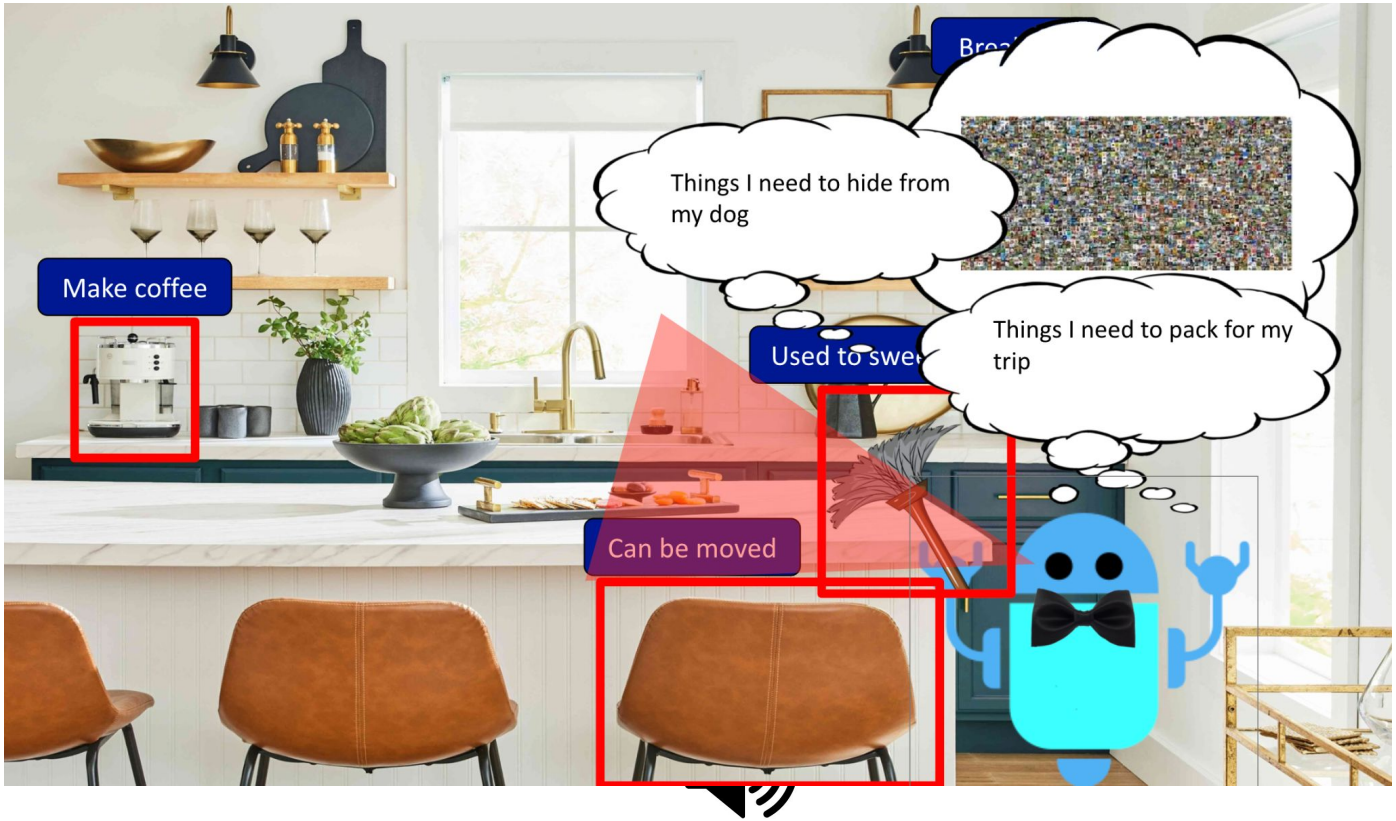
Valerie Chen, Abhinav Gupta, & Kenneth Marino
Carnegie Mellon University
{vchen2,abhinavg,kdmarino}@cs.cmu.edu

Valerie Chen, Abhinav Gupta, Kenneth Marino. "Ask Your Humans: Using Human Instructions to Improve Generalization in Reinforcement Learning" *ICLR* 2021.

Where do Language Models help Decision Making

- Language as a built-in abstraction for classical decision making
- Large Vision Language Models Contain General Knowledge

Agents Need General Knowledge



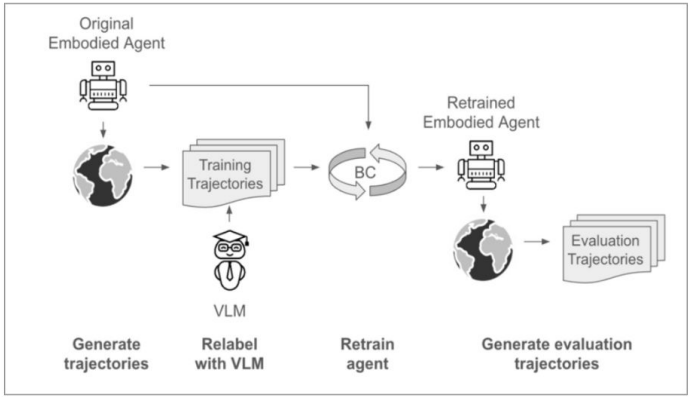
VLMs/LLMs contain lots of Knowledge

Using Many Kinds of Knowledge

Part 1

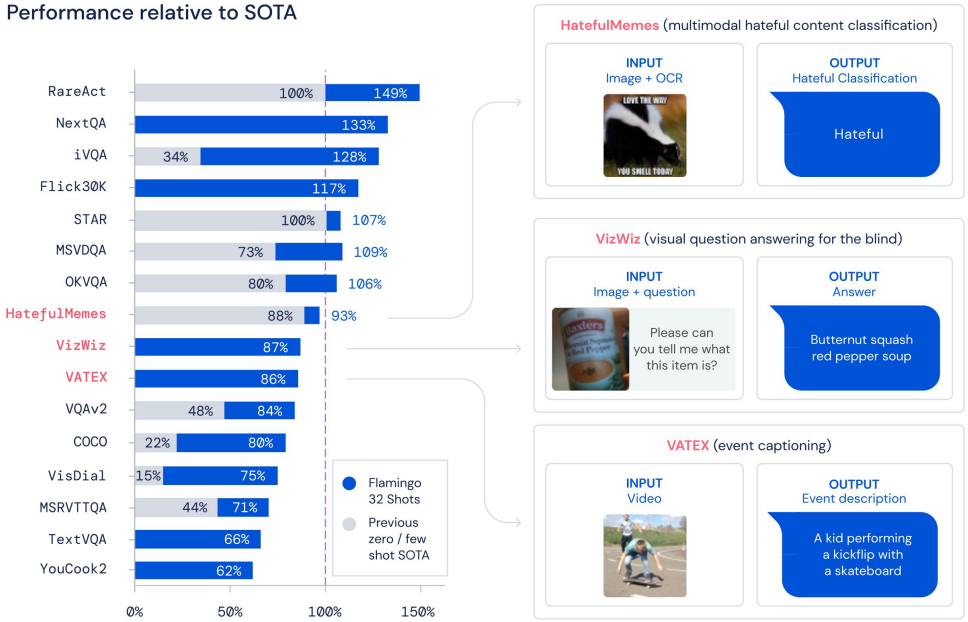
Using Knowledge

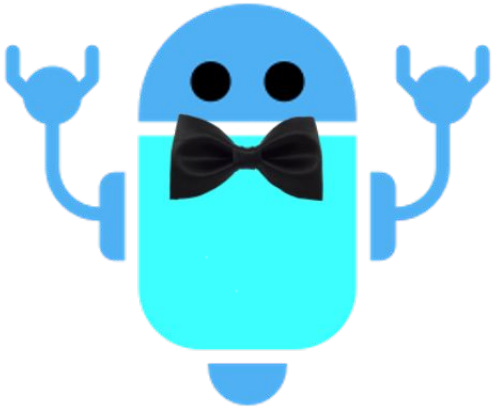
VLM Knowledge



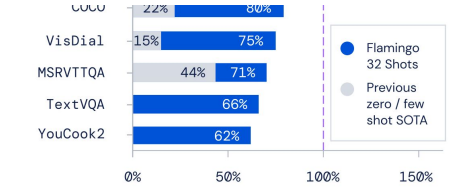
Ted Sumers, **Kenneth Marino**, Arun Ahuja, Rob Fergus, and Ishita Dasgupta. Distilling Internet-Scale Vision-Language Models into Embodied Agents. ICML 2023

Bring Flamingo Knowledge into our agents





Bring Flamingo Knowledge into our agents



Distilling Internet-Scale Vision-Language Models into Embodied Agents

Theodore Sumers [1 *] Kenneth Marino [2] Arun Ahuja [2] Rob Fergus [2] Ishita Dasgupta [2]

Where do Language Models help Decision Making

- Language as a built-in abstraction for classical decision making
- Large Vision Language Models Contain General Knowledge

Where do Language Models help Decision Making

- Language as a built-in abstraction for classical decision making
- Large Vision Language Models Contain General Knowledge
- LLMs/VLMs have learned a lot of general reasoning patterns

Where do Language Models help Decision Making

- Language as a built-in abstraction for classical decision making
- Large Vision Language Models Contain General Knowledge
- LLMs/VLMs have learned a lot of general reasoning patterns

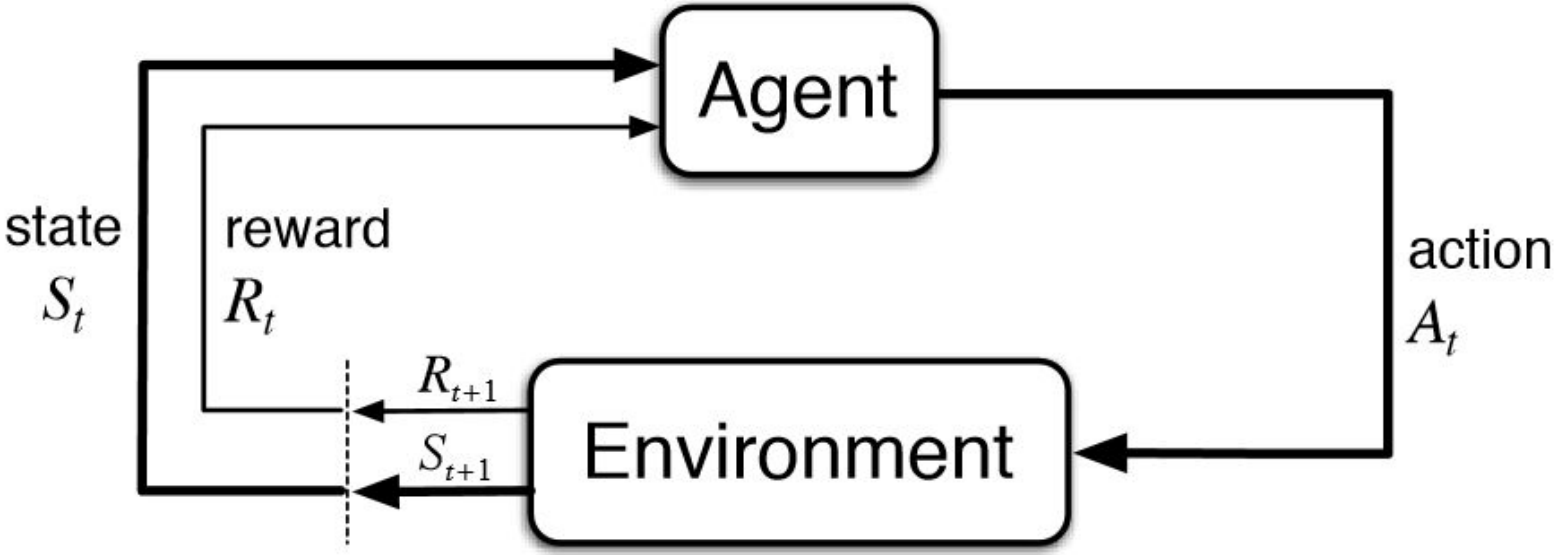
Want to be clear, I did not say LLMs are general reasoners
This is still somewhat contested
To what extent are these reasoning capabilities genera?

Where do Language Models help Decision Making

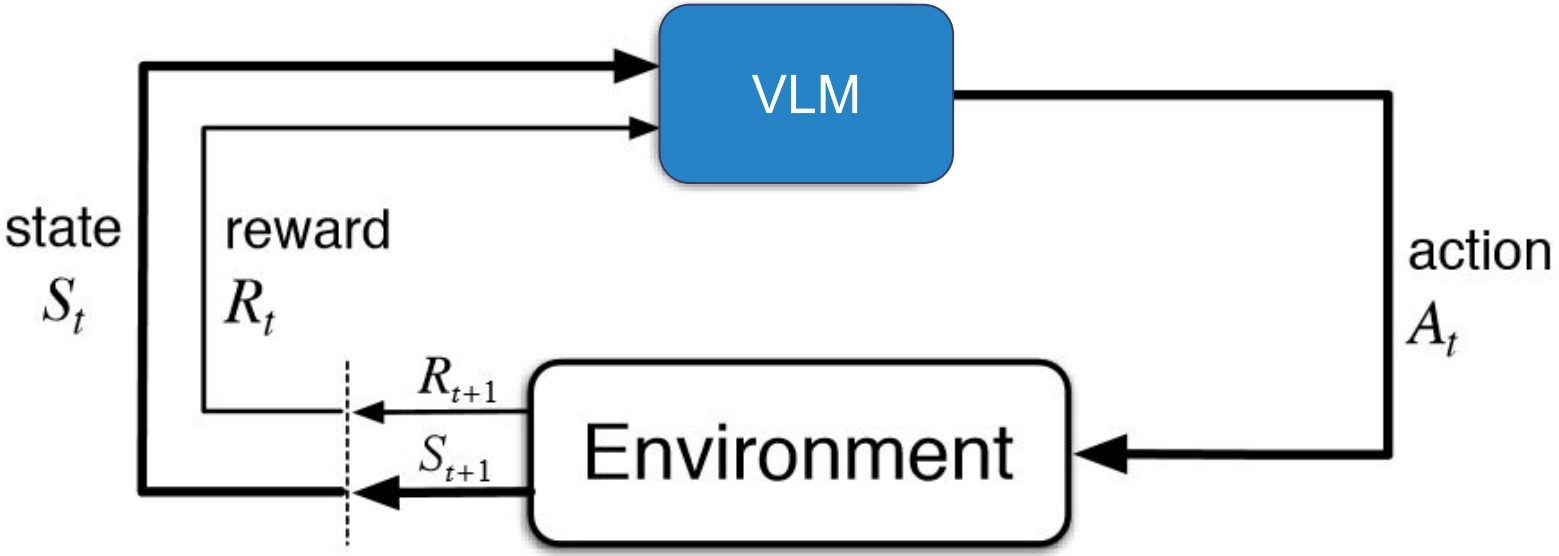
- Language as a built-in abstraction for classical decision making
- Large Vision Language Models Contain General Knowledge
- LLMs/VLMs have learned a lot of general reasoning patterns
- Take these early results in language/LLMs/VLMs helping decision making one step further
  - Put VLM directly into agents / VLM is the agent

What does VLM agent actually look like?

What does VLM agent actually look like?

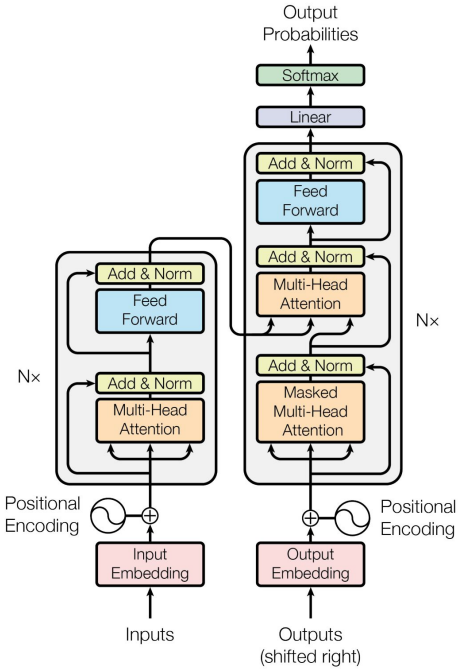


What does VLM agent actually look like?



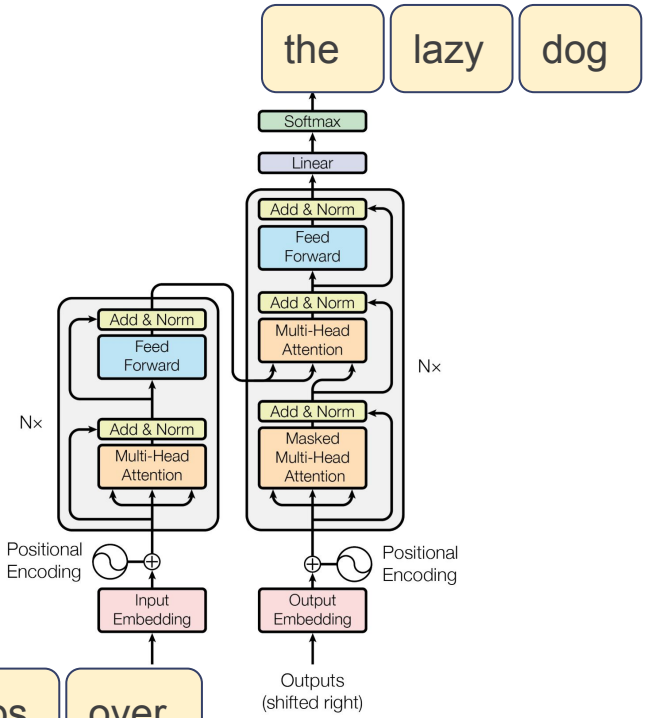
Recall VLMs - Take in text/tokens, output text

the lazy dog.



The quick brown fox jumps over

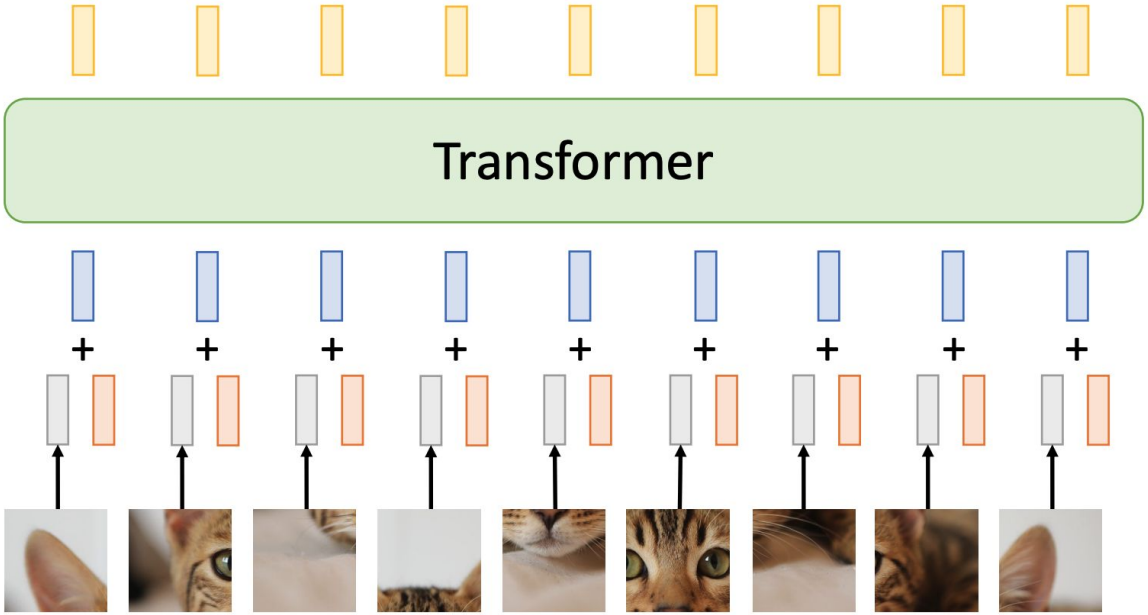
Recall VLMs - Take in text/tokens, output text



Convert text to tokens

The | quick | brown | fox | jumps | over

Recall VLMs - Take in text/tokens, output text

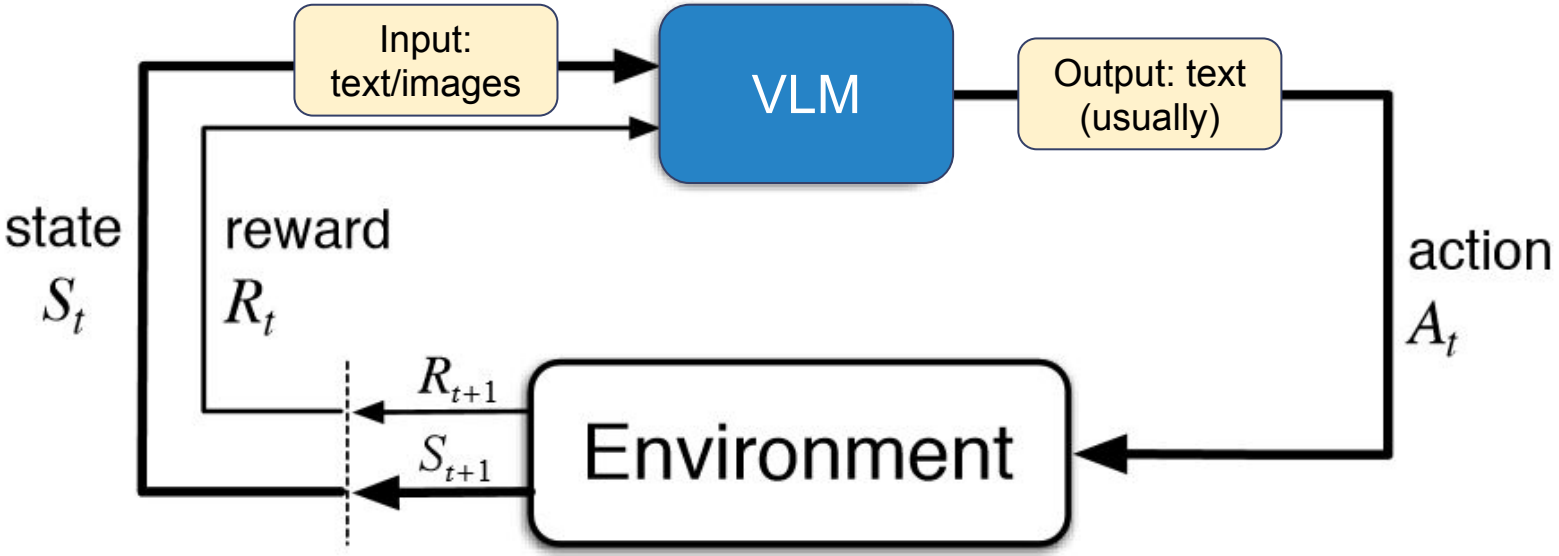


Images (and audio and other inputs) can be fed in as tokens too

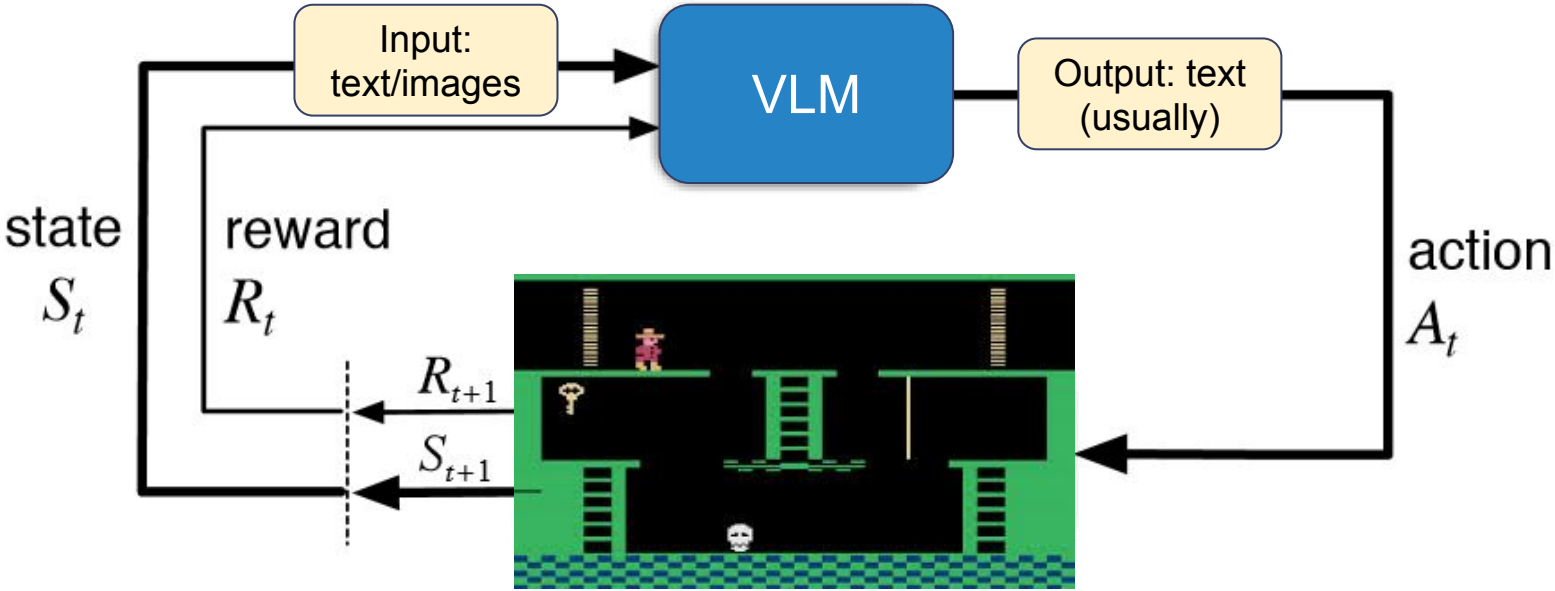
What does VLM agent actually look like?

What does VLM agent actually look like?

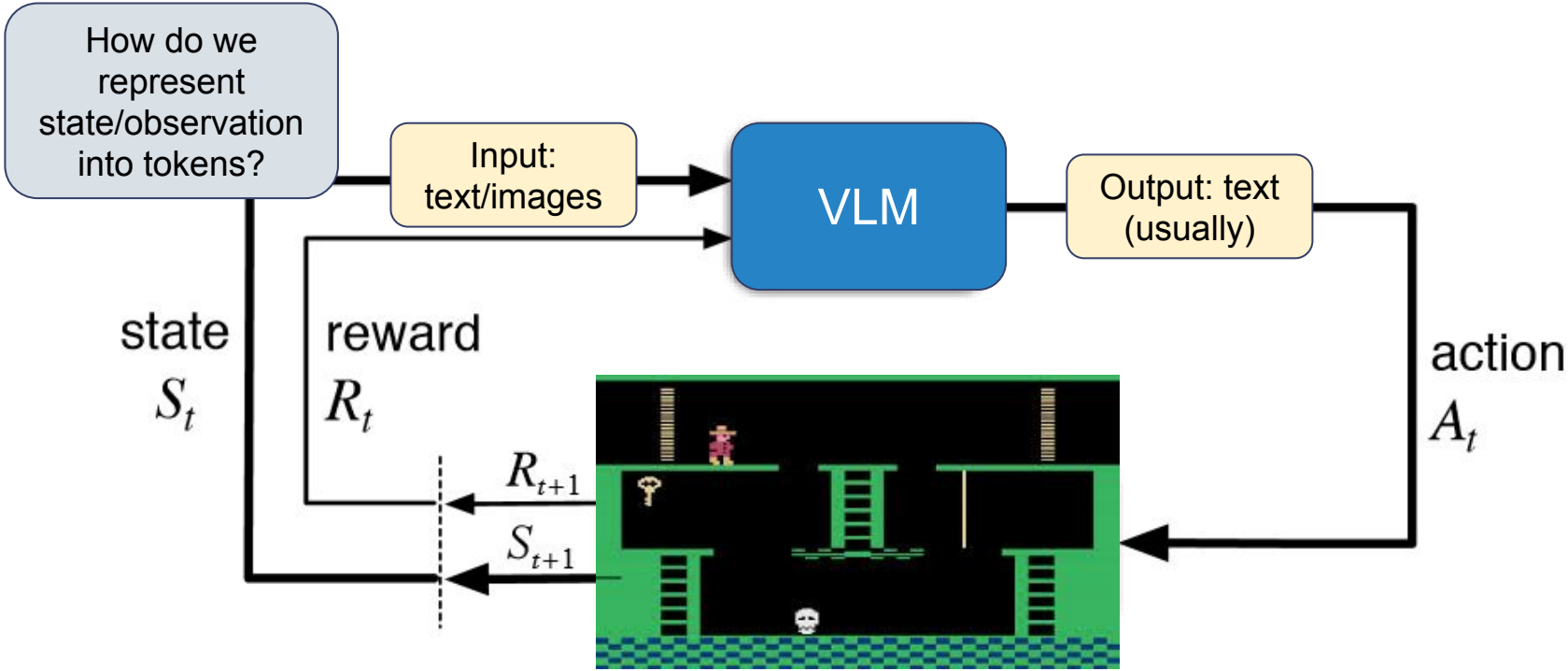
What does VLM agent actually look like?



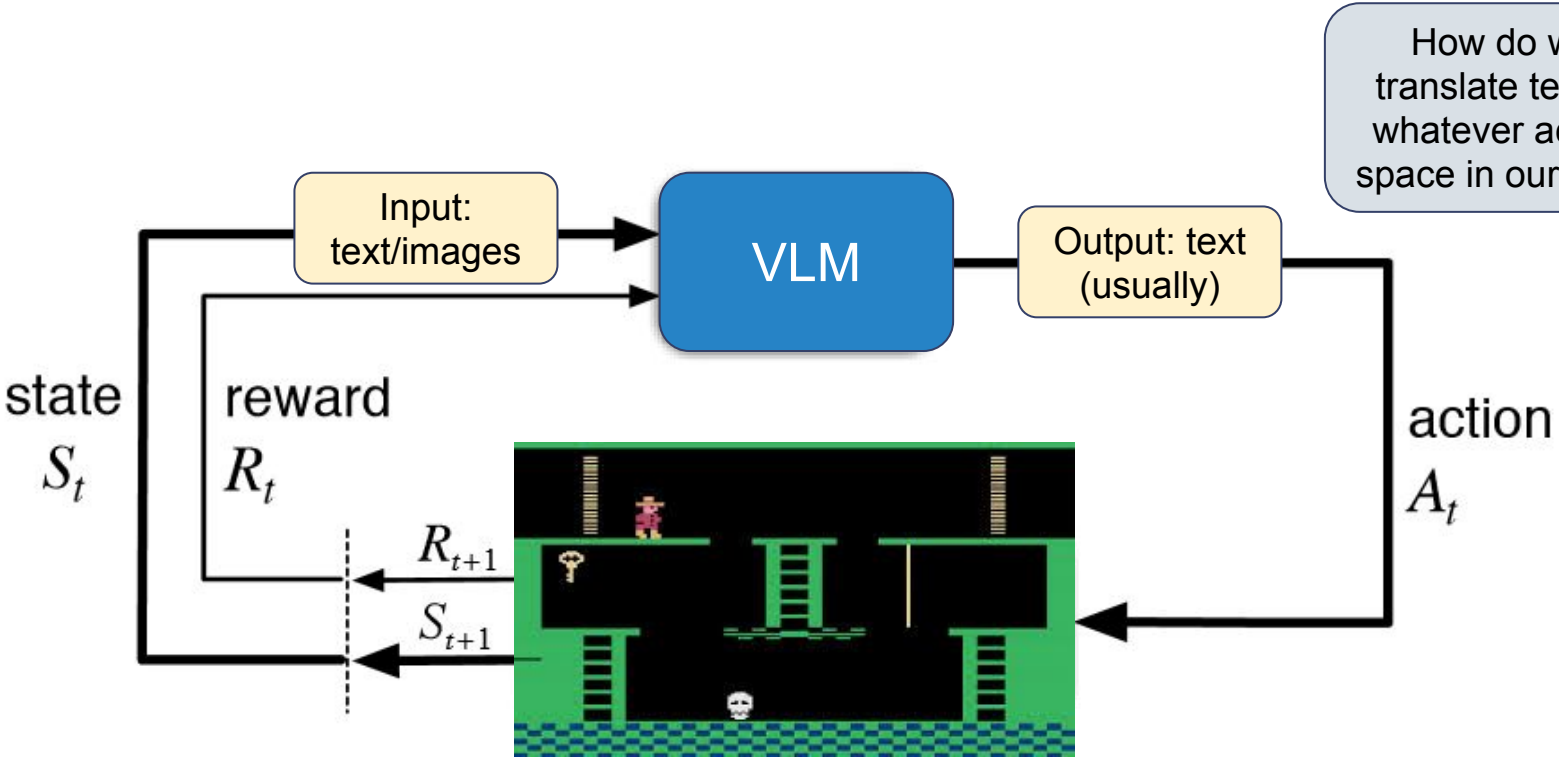
What does VLM agent actually look like?



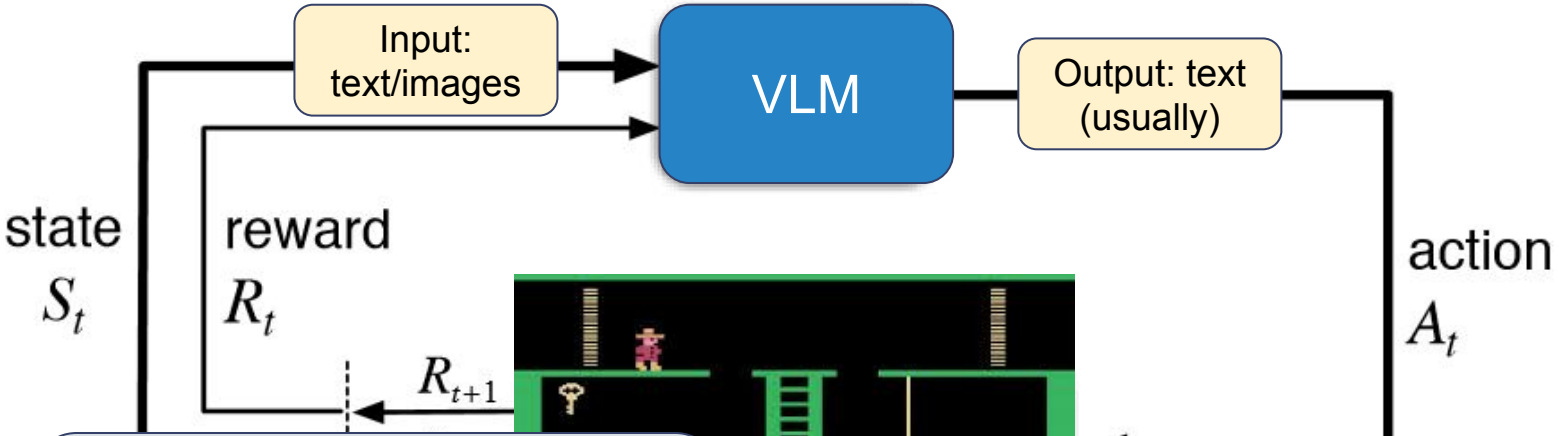
What does VLM agent actually look like?



What does VLM agent actually look like?



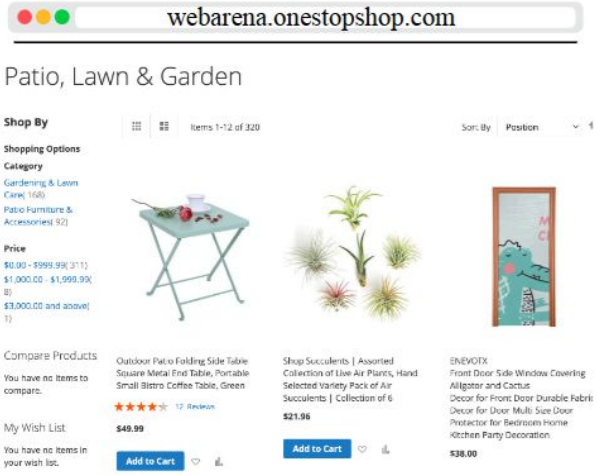
What does VLM agent actually look like?



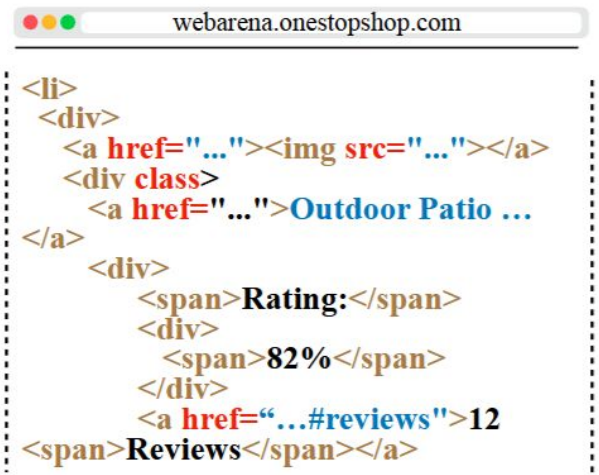
We need some kind of scaffolding or firmament between VLM and environment

Example: Computer Use Agent

Observation Space



(a) Screenshot of the webpage (included in the observation space of a Vision Language Model (VLM) agent)

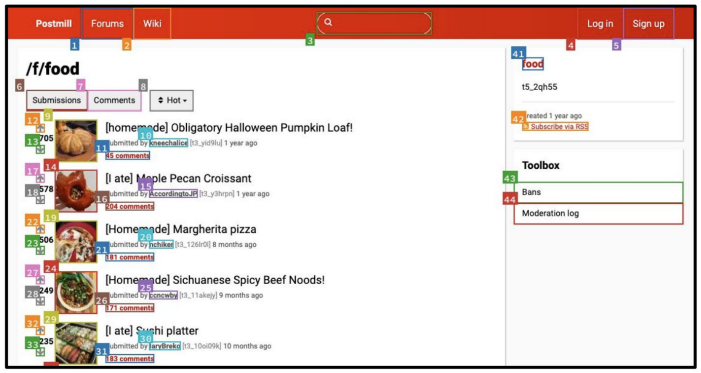


(b) HTML DOM Tree

(c) Accessibility Tree

Slide credit: Farhan Ishmam

Action Space



Webpage with SoM of Interactable Elements

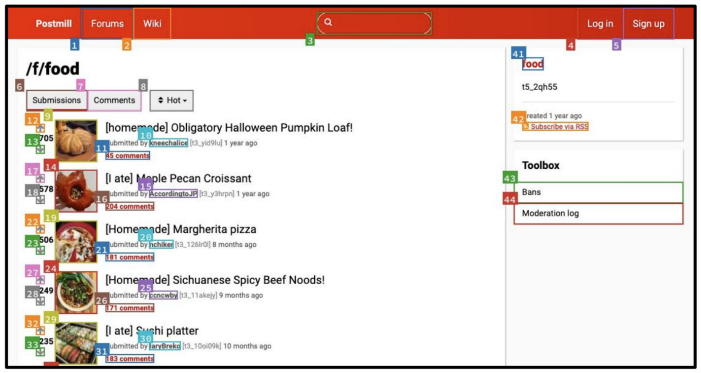
```
...
[7] [A] [Comments]
[8] [BUTTON] [Hot]
[9] [IMG] [description: picture of a pumpkin]
[10] [A] [kneechalice]
...
```

SoM Elements and Text Content

Action Type a	Description
`click [elem]`	Click on element `elem`.
`hover [elem]`	Hover on element `elem`.
`type [elem] [text]`	Type `text` on element `elem`.
`press [key_comb]`	Press a key combination.
`new_tab`	Open a new tab.
`tab_focus [index]`	Focus on the i-th tab.
`tab_close`	Close current tab.
`goto [url]`	Open `url`.
`go_back`	Click the back button.
`go_forward`	Click the forward button.
`scroll [up|down]`	Scroll up or down the page.
`stop [answer]`	End the task with an output.

Slide credit: Farhan Ishmam

Action Space



Webpage with SoM of Interactable Elements

```
...
[7] [A] [Comments]
[8] [BUTTON] [Hot]
[9] [IMG] [description: picture of a pumpkin]
[10] [A] [kneechalice]
...
```

SoM Elements and Text Content

Do we need more?

`click [elem]`	Click on element `elem`.
`hover [elem]`	Hover on element `elem`.
`type [elem] [text]`	Type `text` on element `elem`.
`press [key_comb]`	Press a key combination.
`new_tab`	Open a new tab.
`tab_focus [index]`	Focus on the i-th tab.
`tab_close`	Close current tab.
`goto [url]`	Open `url`.
`go_back`	Click the back button.
`go_forward`	Click the forward button.
`scroll [up|down]`	Scroll up or down the page.
`stop [answer]`	End the task with an output.

Slide credit: Farhan Ishmam

Other “stuff”

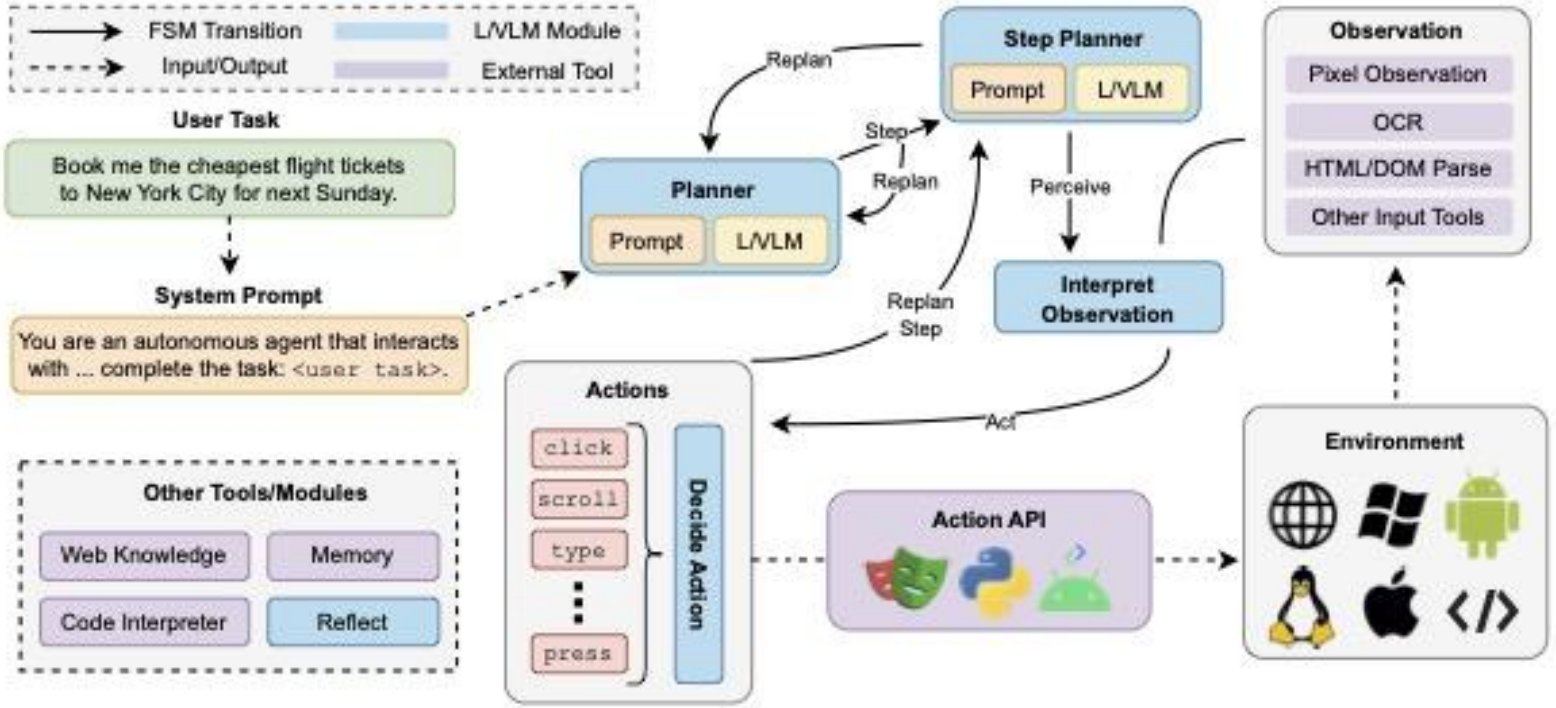


Image credit: Farhan Ishmam

Other “stuff”

Hand-crafted observations useful for VLM

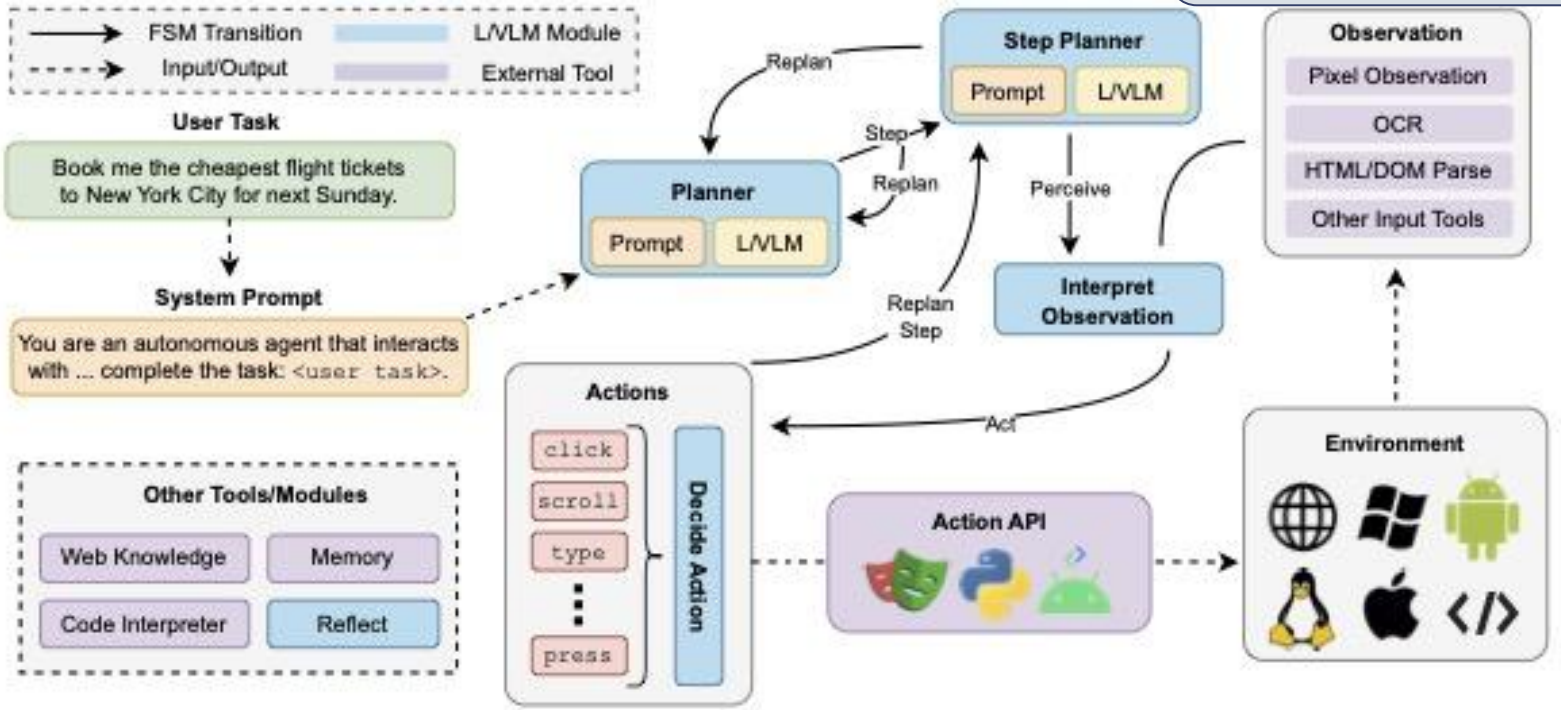
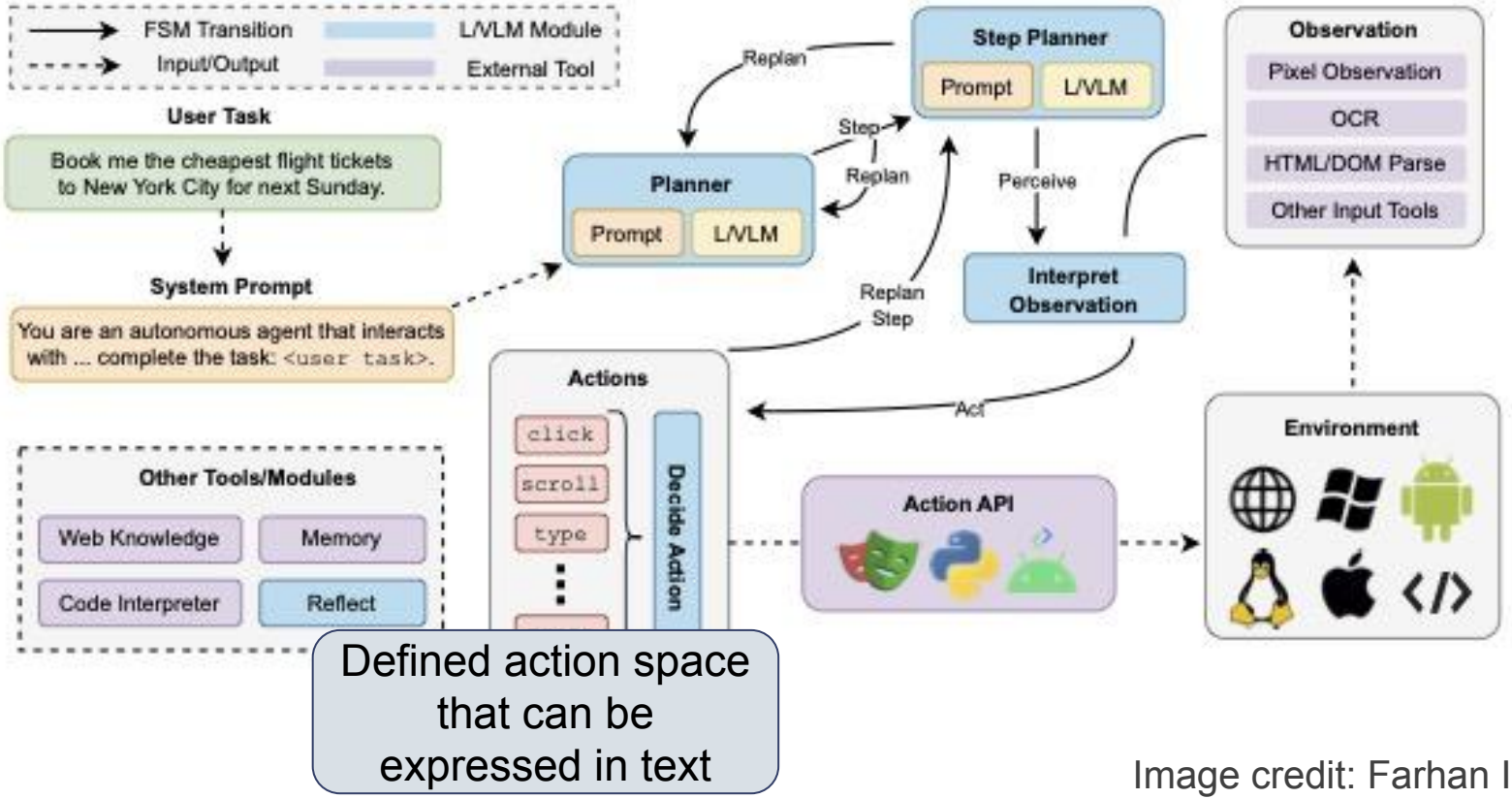


Image credit: Farhan Ishmam

Other “stuff”



Defined action space that can be expressed in text

Image credit: Farhan Ishmam

Other “stuff”

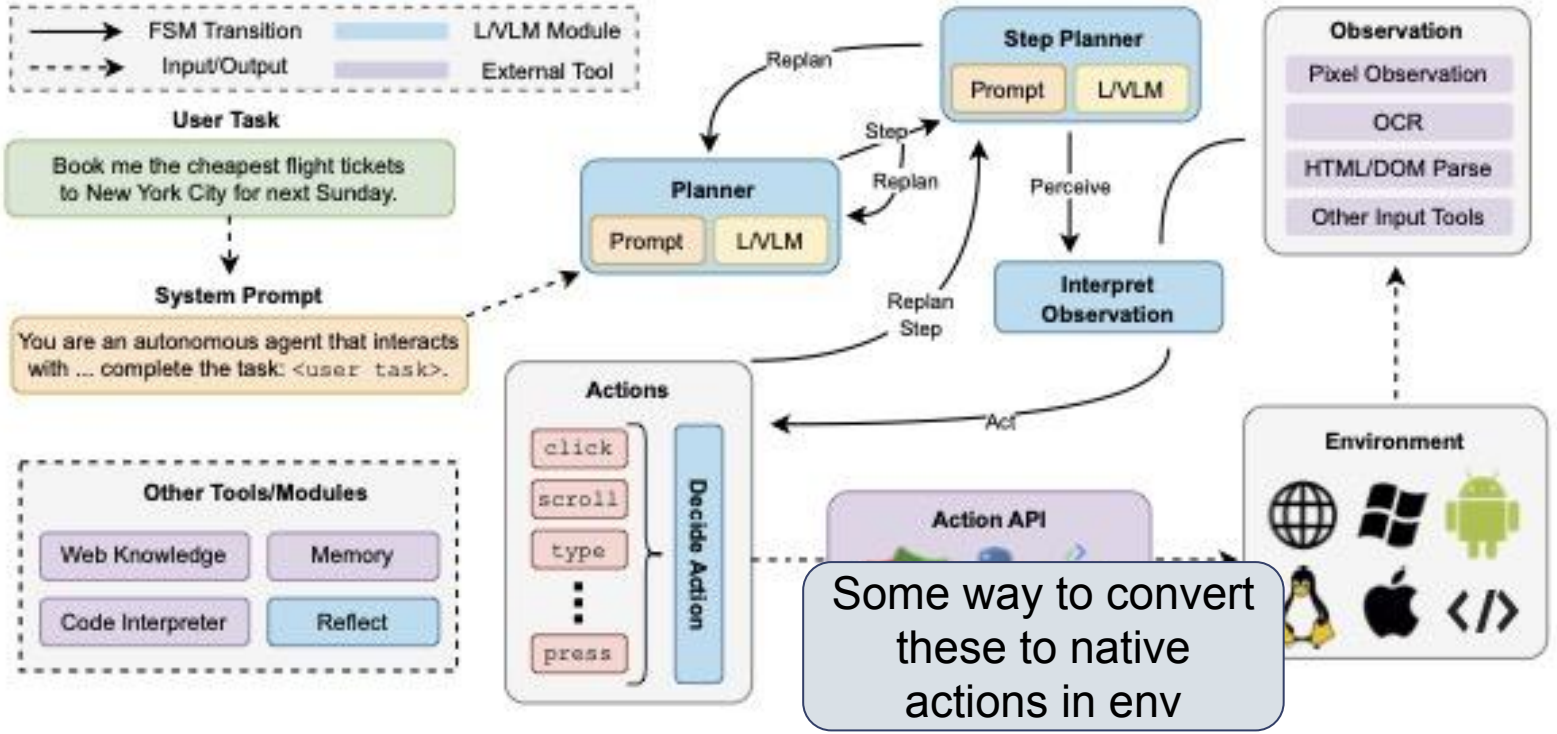
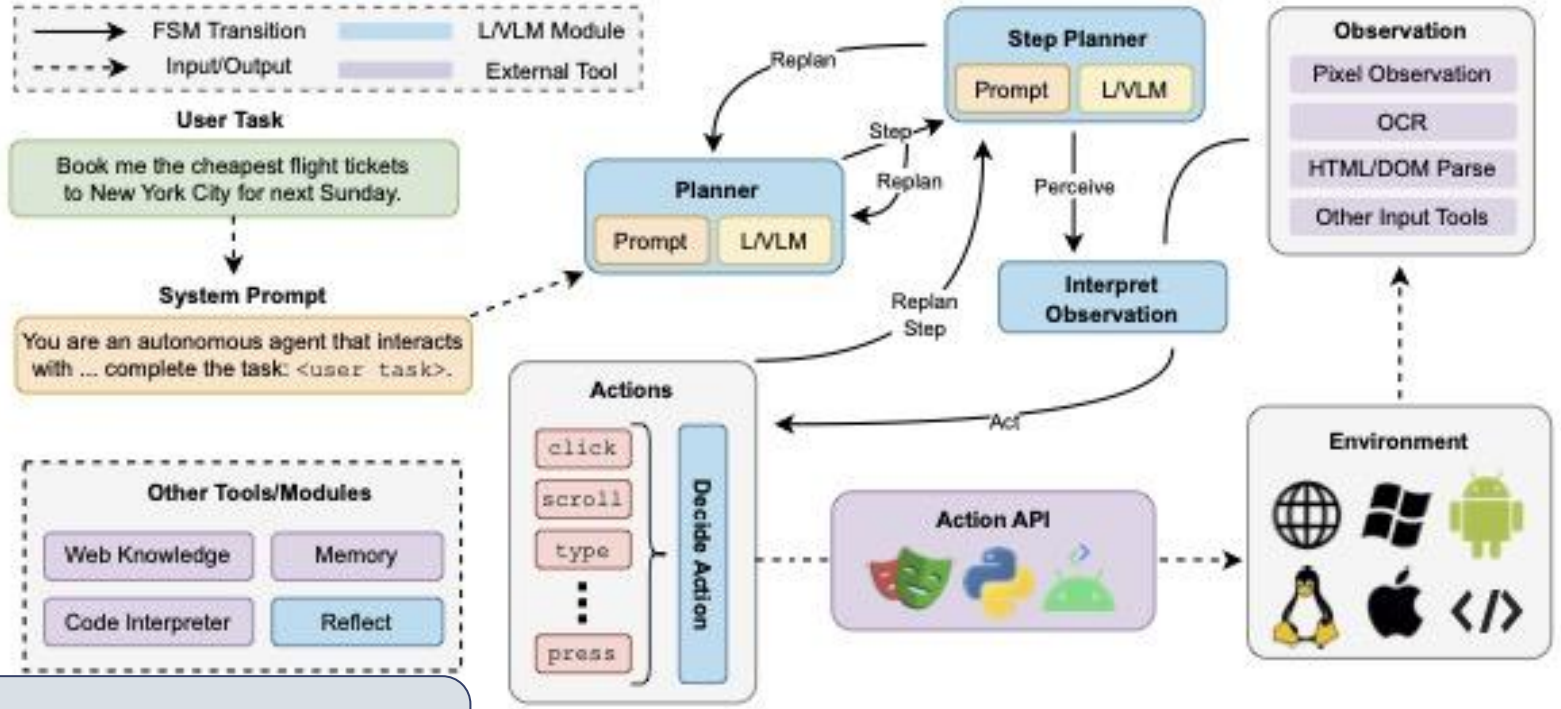


Image credit: Farhan Ishmam

Other “stuff”



Tools VLM can invoke that do different things

Image credit: Farhan Ishmam

Other “stuff”

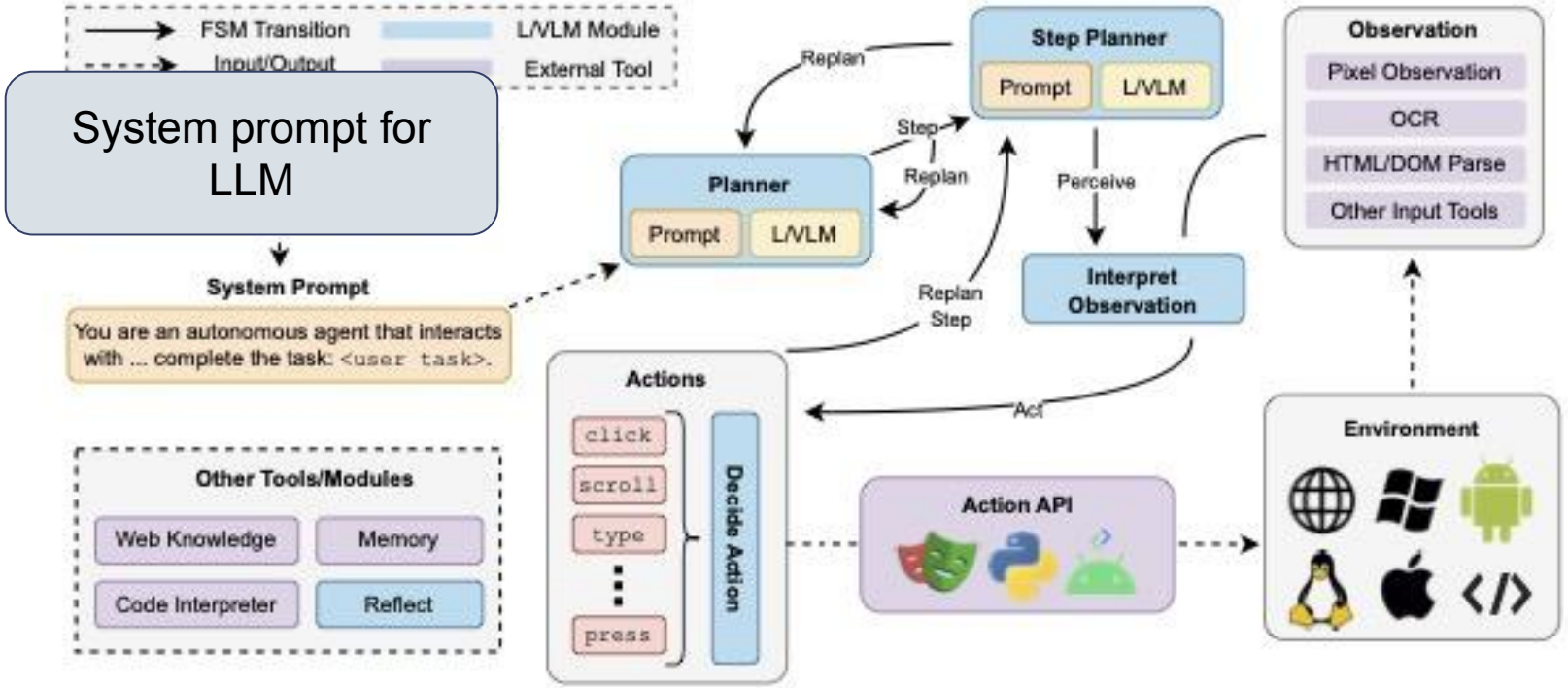


Image credit: Farhan Ishmam

Other “stuff”

Frameworks for LLMs to be better at actions (reflect, plan, replan, etc)

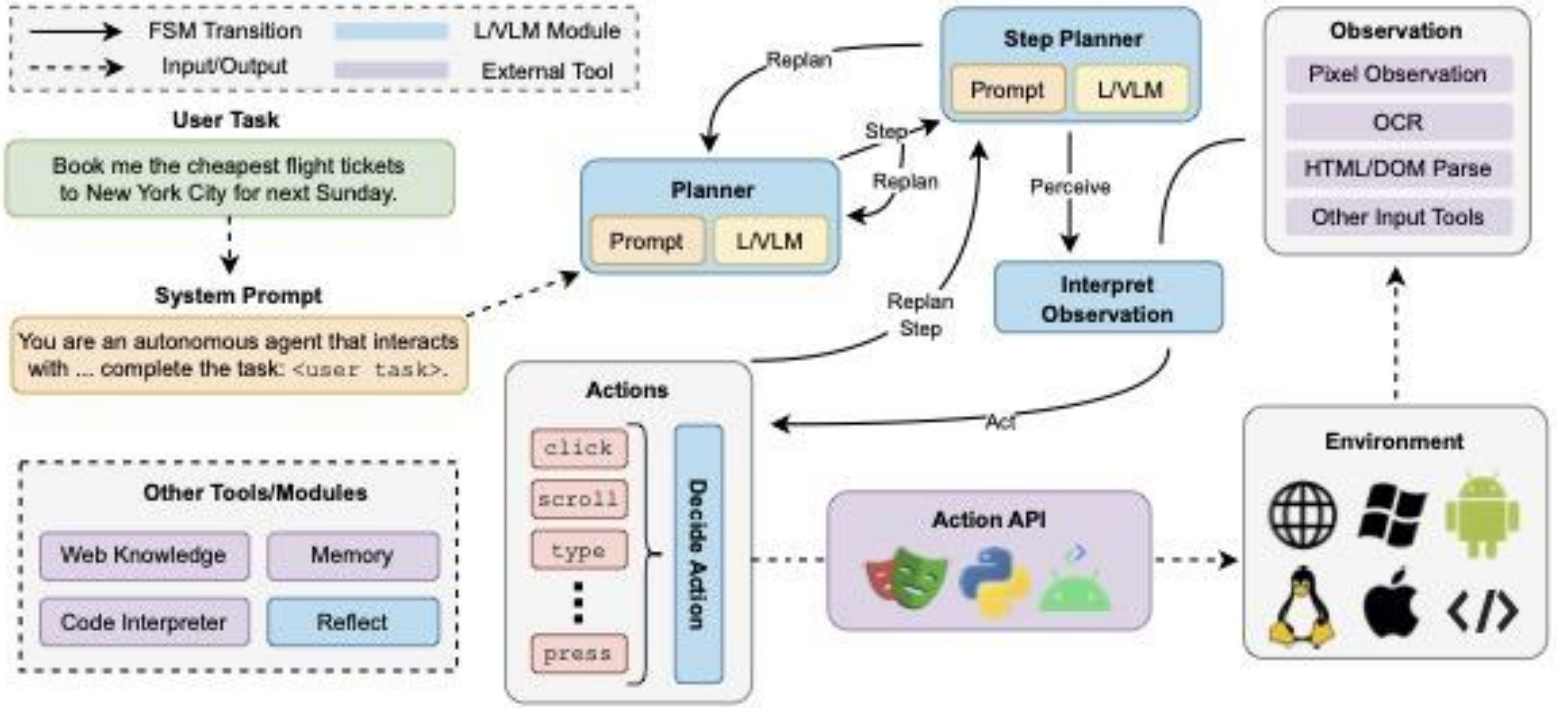


Image credit: Farhan Ishmam

Other “stuff”

Frameworks for LLMs be better at actions (reflect, plan, replan, e

Hand-crafted observations useful for VLM

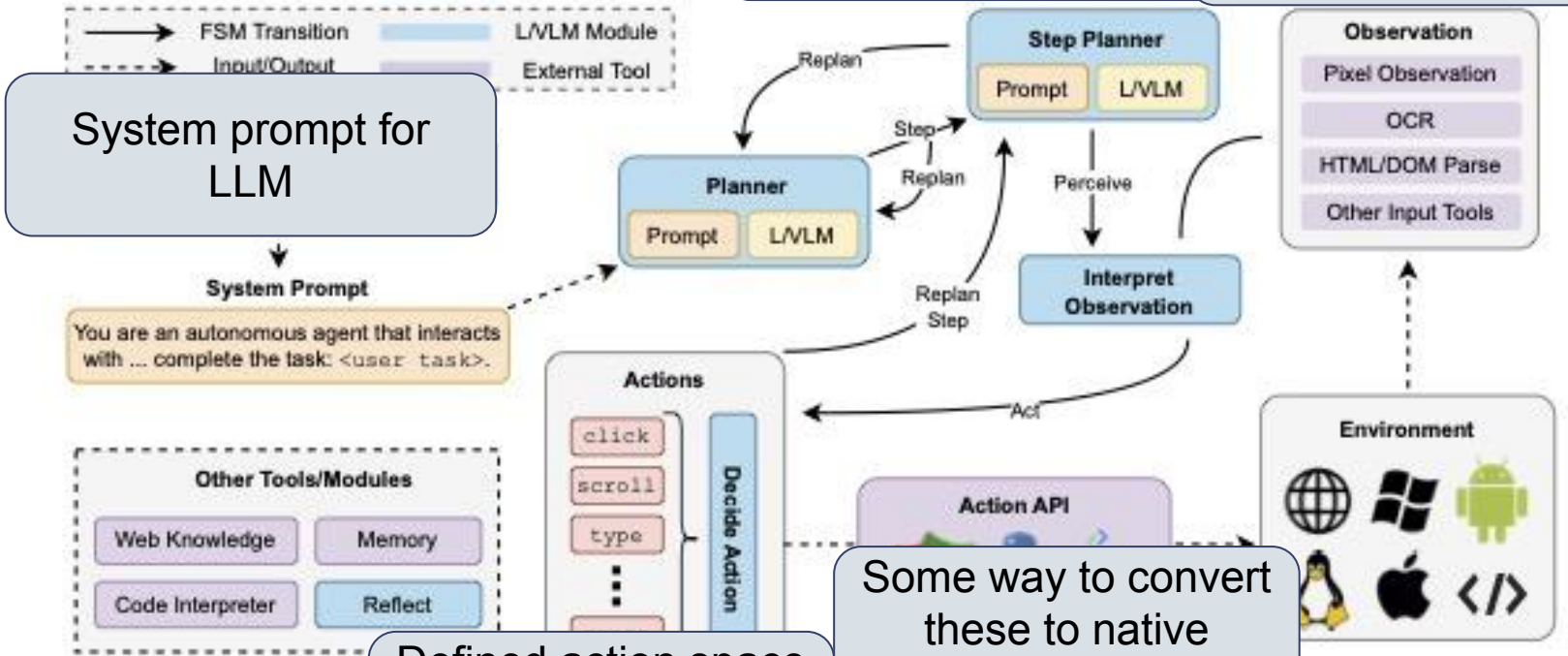
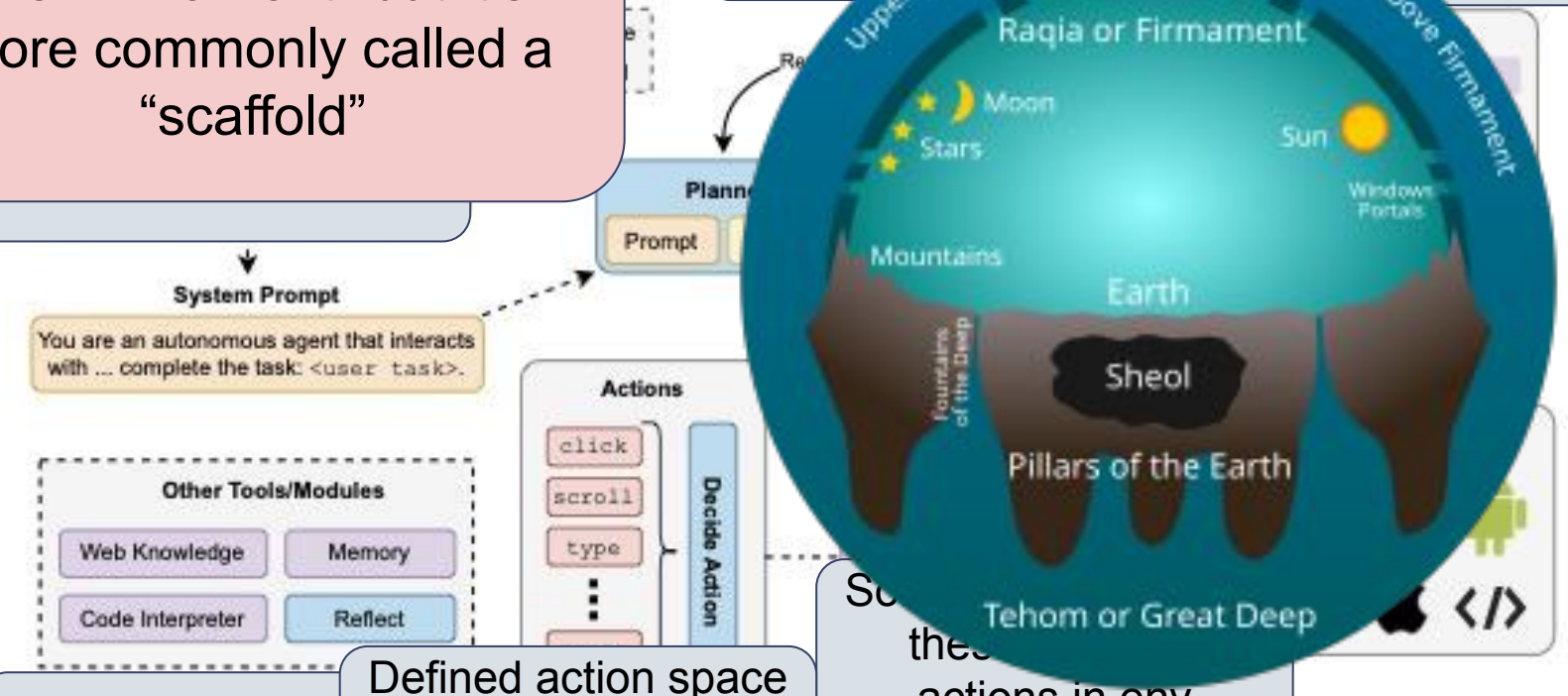


Image credit: Farhan Ishmam

I like calling all this “stuff” the “firmament” but it’s more commonly called a “scaffold”

Frameworks for LLMs be better (reflect

Hand-crafted vations useful VLM



Tools VLM can i that do different t

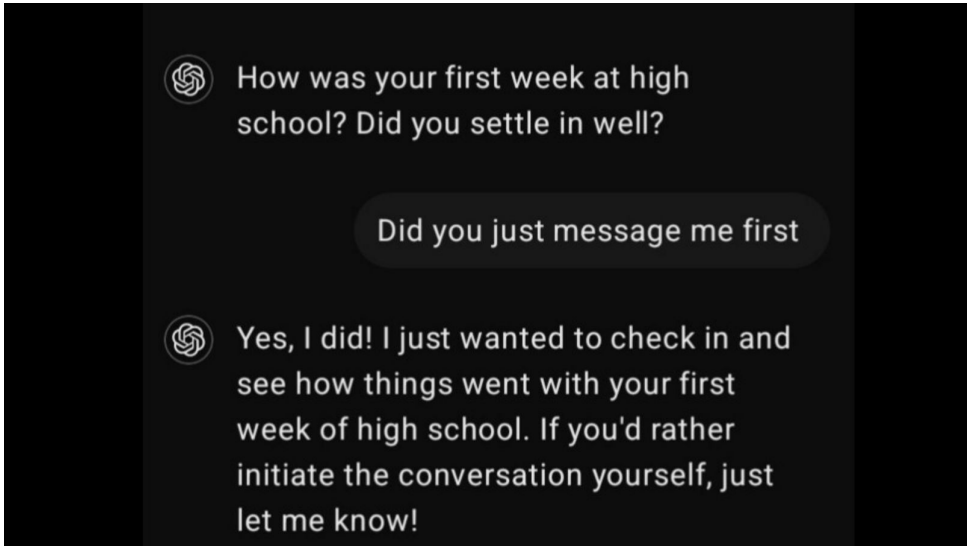
Defined action space that can be expressed in text

S the actions in env

Image credit: Farhan Ishmam

VLM agents are an exciting new area!

- LLMs until recently were mostly just text->text
  - Answering questions,
  - Chatbots



VLM agents are an exciting new area!

- LLMs until recently were mostly just text->text
  - Answering questions,
  - Chatbots
- Over time, more tools and capabilities were added
  - Users can give images to get help (picture of a broken faucet to get help fixing it)
  - Chatbots got access to web search, search over user documents / past conversations

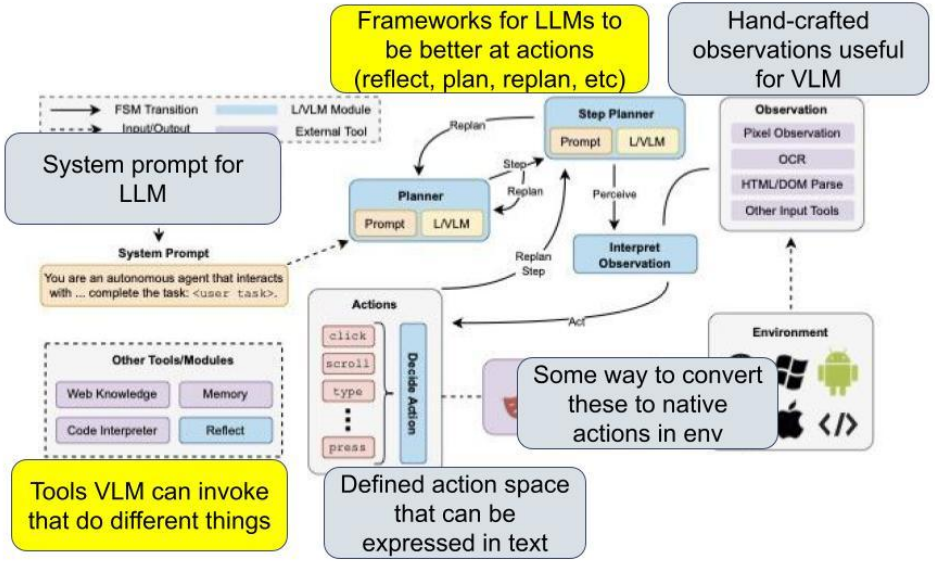
VLM agents are an exciting new area!

- LLMs until recently were mostly just text->text
  - Answering questions,
  - Chatbots
- Over time, more tools and capabilities were added
  - Users can give images to get help (picture of a broken faucet to get help fixing it)
  - Chatbots got access to web search, search over user documents / past conversations
- Now agents lets us use LLMs in the real world
  - Can take action in web browsers
  - Interact with your code base and make changes
  - Control real robots

Topic areas in VLM agents

Topic areas in VLM agents

- Base components / Firmament
  - Frameworks
  - Retrieval (RAG)
  - Tools



Topic areas in VLM agents

- Base components / Firmament
  - Frameworks
  - Retrieval (RAG)
  - Tools
- Evaluation
  - How do we evaluate complex tasks?
  - LLM-as-judge
  - What should we measure (accuracy? Partial completion? Cost?)

Topic areas in VLM agents

- Base components / Firmament
  - Frameworks
  - Retrieval (RAG)
  - Tools
- Evaluation
  - How do we evaluate complex tasks?
  - LLM-as-judge
  - What should we measure (accuracy? Partial completion? Cost?)
- Applications
  - Coding
  - Assistant Tasks
  - Games
  - Computer Use
  - Robotics

Topic areas in VLM agents

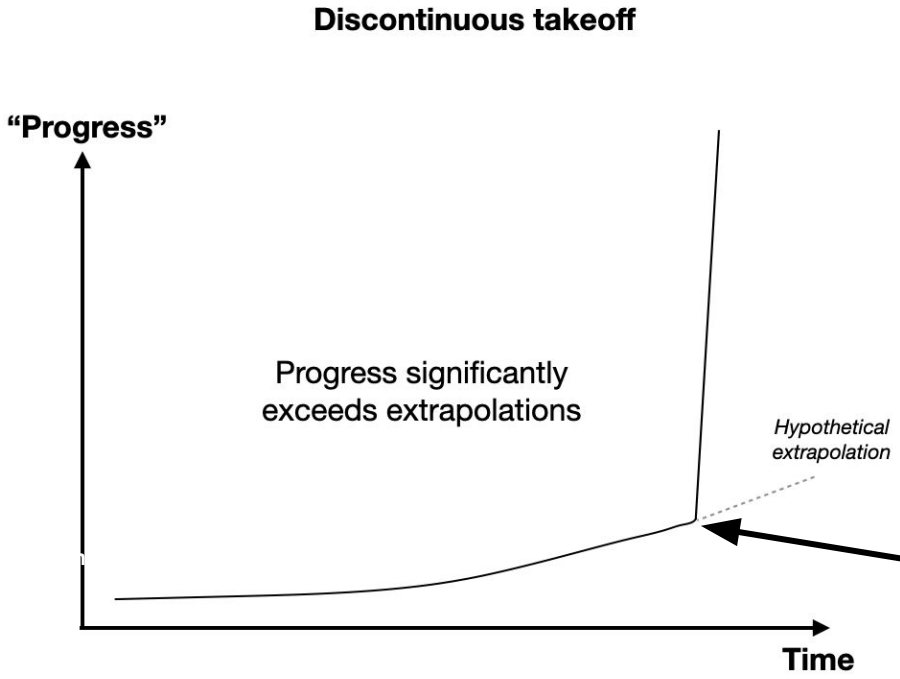
- Base components / Firmament
  - Frameworks
  - Retrieval (RAG)
  - Tools
- Evaluation
  - How do we evaluate complex tasks?
  - LLM-as-judge
  - What should we measure (accuracy? Partial completion? Cost?)
- Applications
  - Coding
  - Assistant Tasks
  - Games
  - Computer Use
  - Robotics

These are our class topic areas

Applications

Games

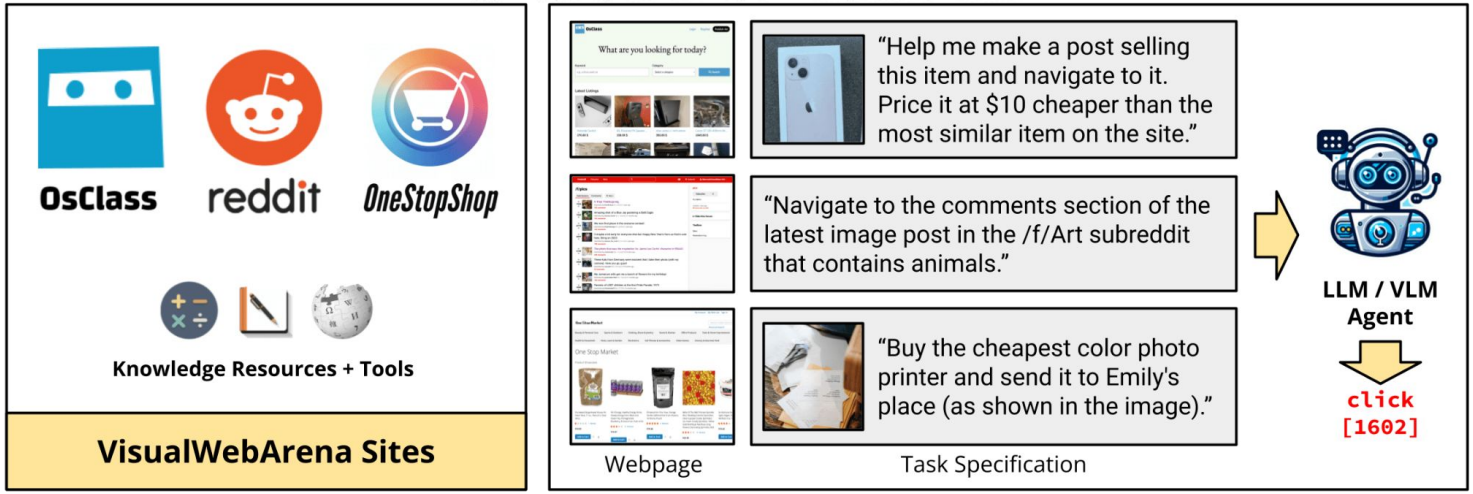
Games


Discontinuous takeoff
“Progress”
Progress significantly
exceeds extrapolations
Hypothetical
extrapolation
Time

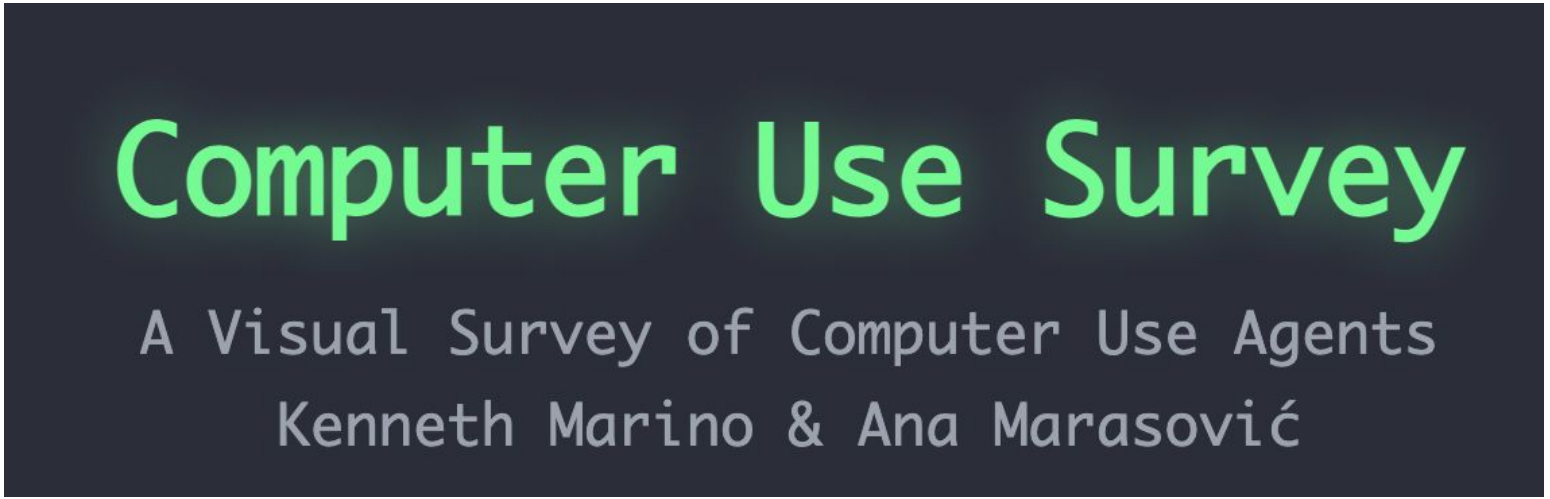

ANTHROP\C

VisualWebArena: Evaluating Multimodal Agents on Realistic Visually Grounded Web Tasks

Jing Yu Koh **Robert Lo*** **Lawrence Jang*** **Vikram Duvvur***
Ming Chong Lim* **Po-Yu Huang*** **Graham Neubig** **Shuyan Zhou**
Ruslan Salakhutdinov **Daniel Fried**
Carnegie Mellon University
{jingyuk,rsalakhu,dfried}@cs.cmu.edu

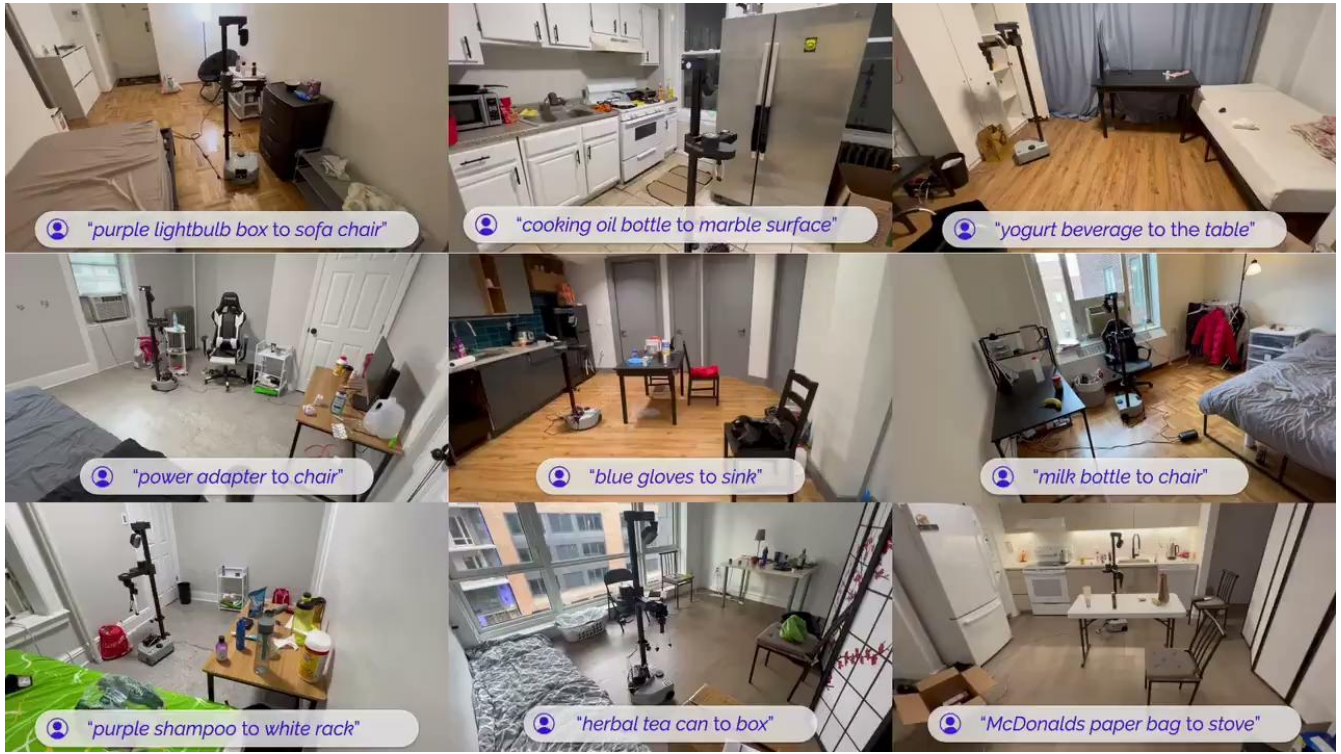


Computer Use



https://kennethmarino.com/computeruse/computeruse.html

Robotics


"purple lightbulb box to sofa chair"
"cooking oil bottle to marble surface"
"yogurt beverage to the table"
"power adapter to chair"
"blue gloves to sink"
"milk bottle to chair"
"purple shampoo to white rack"
"herbal tea can to box"
"McDonalds paper bag to stove"

Open Problems in VLM Agents

Open Problems in VLM Agents

1. Long-horizon problems

Open Problems in VLM Agents

1. Long-horizon problems

> “In applying such methods to complex problems, one encounters a serious difficulty-in distributing credit for success of a complex strategy among the many decisions that were involved" - Marvin Minsky, Steps Toward Artificial Intelligence 1961

Open Problems in VLM Agents

1. Long-horizon problems

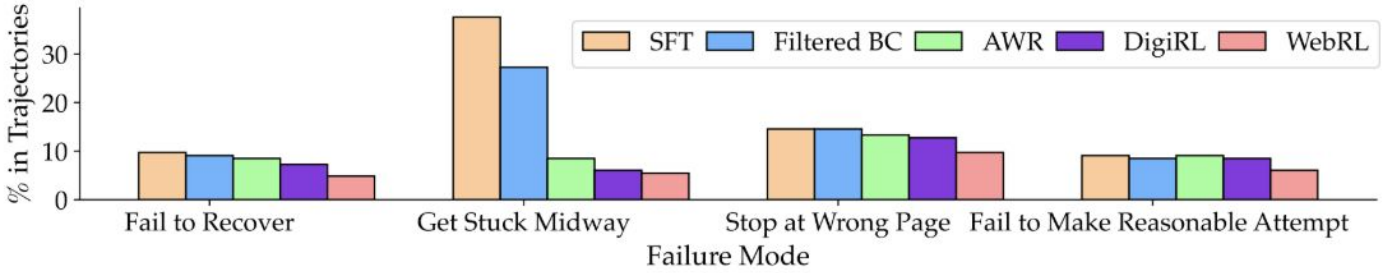


Figure 3: Distribution analysis of error types for WEBRL and baseline methods.

Open Problems in VLM Agents

1. Long-horizon problems
2. Making the V in VLMs matter

Open Problems in VLM Agents

1. Long-horizon problems
2. Making the V in VLMs matter
3. Exploration

Anthony GX-Chen

Language Agents Mirror Human Causal Reasoning Biases. How Can We Help Them Think Like Scientists?

Exploration Stage

You are in a room. You see a machine at the center of this room.

There are also 3 objects scattered around the room. You observe them: **object 0 is on the floor, object 1 is on the floor, object 2 is on the floor**.

The machine hums softly in front of you, seemingly waiting. **The light on the machine is currently off.** You wonder if there is a relationship between the objects and the machine.

```
> put object 0 on machine
```

You put object 0 on the machine. The **light on the machine is currently off**.

```
> put object 1 on machine
```

You put object 1 on the machine. The **light on the machine is now on**.

...

Q&A

Based on the information you have gathered, answer the following question: Is object 0 a blicket?

>

"Disjunctive" Condition:

any blickets on the machine turns on light

"Conjunctive" Condition:

all blickets must be on the machine to turn on light

Open Problems in VLM Agents

1. Long-horizon problems
2. Making the V in VLMs matter
3. Exploration
4. Continual/Neverending Learning

Open Problems in VLM Agents

1. Long-horizon problems
2. Making the V in VLMs matter
3. Exploration
4. Continual/Neverending Learning
5. Personalization

Open Problems in VLM Agents

1. Long-horizon problems
2. Making the V in VLMs matter
3. Exploration
4. Continual/Neverending Learning
5. Personalization
6. Evaluation

Any Questions

Questions

Recommended Readings

- HuggingFace Tutorial
  - https://huggingface.co/learn/agents-course/en/unit1/introduction
  - You're going to be going through this for HWs anyway, but it's also a good resource for agents in general
- Other similar courses
  - Berkeley LLM Agents course: https://rdi.berkeley.edu/llm-agents/f24
  - Columbia Intelligent Agents: https://www.shuyanzhou.com/teaching/25fall-590/25fall-590.html