

Assistant Agents

CS6960 MultiModal LLM Agents

Kenneth Marino

Announcements

- Projects
 - Working on grading / feedback for proposals
 - Begin working on project if you haven't already
 - Get set up on CHPC or other compute
 - ...

Any Questions

Assistant Agents

- Kind of a bucket term
 - Agents that can “assist people”

Assistant Agents

- Things we can call assistant agents (in presentations)
 - LLM with search/retrieval tools



A scholarly research assistant that combines literature understanding and data-driven discovery. Asta uses 108M+ abstracts and 12M+ full-text papers to find, summarize, and analyze scientific evidence. A project from [AI2](#).

Find papers Generate a report Analyze data

Describe the papers you're looking for



Show example queries

Explore research prototypes in [AstaLabs](#)

Level 1

Question: What was the actual enrollment count of the clinical trial on H. pylori in acne vulgaris patients from Jan-May 2018 as listed on the NIH website?

Ground truth: 90

Level 2



Question: If this whole pint is made up of ice cream, how many percent above or below the US federal standards for butterfat content is it when using the standards as reported by Wikipedia in 2020? Answer as + or - a number rounded to one decimal place.

Ground truth: +4.6

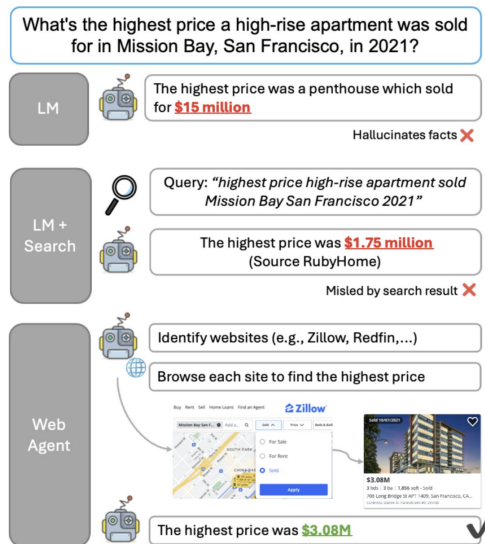
Level 3

Question: In NASA's Astronomy Picture of the Day on 2006 January 21, two astronauts are visible, with one appearing much smaller than the other. As of August 2023, out of the astronauts in the NASA Astronaut Group that the smaller astronaut was a member of, which one spent the least time in space, and how many minutes did he spend in space, rounded to the nearest minute? Exclude any astronauts who did not spend any time in space. Give the last name of the astronaut, separated from the number of minutes by a semicolon. Use commas as thousands separators in the number of minutes.

Ground truth: White; 5876

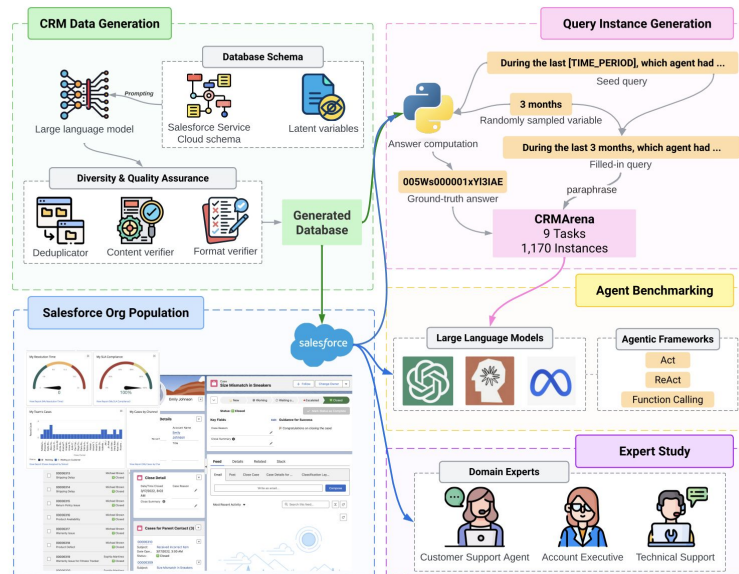
Assistant Agents

- Things we can call assistant agents (in presentations)
 - LLM with search/retrieval tools
 - Web agents



Assistant Agents

- Things we can call assistant agents (in presentations)
 - LLM with search/retrieval tools
 - Web agents
 - Customer Relationship Management



Assistant Agents

- Things we can call assistant agents (in presentations)
 - LLM with search/retrieval tools
 - Web agents
 - Customer Relationship Management
 - Many other things can fall into this bucket

Assistant Agents

- Things we can call assistant agents (in presentations)
 - LLM with search/retrieval tools
 - Web agents
 - Customer Relationship Management
 - Many other things can fall into this bucket

How is Prof. Marino going to make an interesting lecture out of this?

Assistant Agents

- Things we can call assistant agents (in presentations)
 - LLM with search/retrieval tools
 - Web agents
 - Customer Relationship Management
 - Many other things can fall into this bucket

How is Prof. Marino going to make an interesting lecture out of this?



Recall: About Me

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



Recall: About Me

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



Didn't this used to be called
DeepMind?

A non-NDA-violating history of Google DeepMind

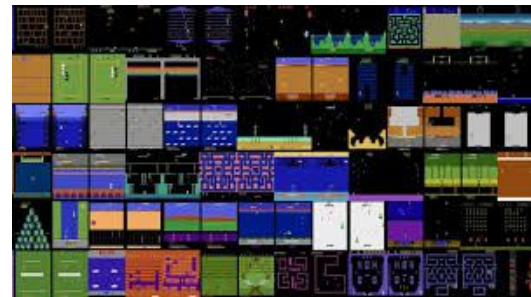
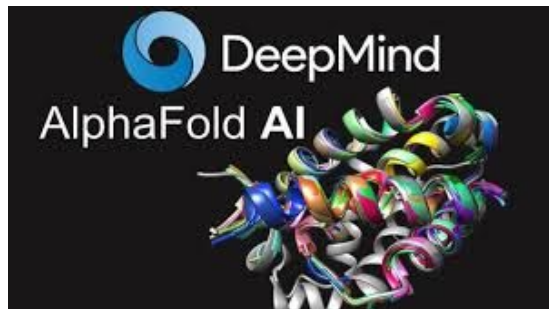
This relates to Assistant Agents, I promise



2021: The old DeepMind



AlphaGo



Atari



StarCraft

2021: The old DeepMind



Games and
RL is the way
to AGI

Then everything changed with ChatGPT



Situation at DeepMind and Google



Situation at DeepMind and Google

- Suddenly very behind
- We need an answer to ChatGPT
- Investors worried about search
- Execs wondering why we have at least 3 AI groups

Refocus to LLMs



DeepMind and Google Smosh logos



Pictured: Font crimes

Google DeepMind launches Bard

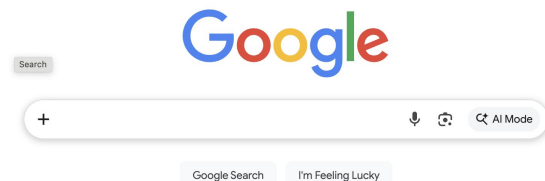
Can we reconcile LLMs with the old ways?

Recall: An earlier definition of agents

- 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

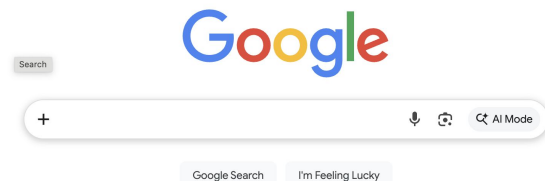
Solutions: Start making LLMs more “agentic”

Solutions: Start making LLMs more “agentic”



Integrated search was the initial differentiator with ChatGPT

Solutions: Start making LLMs more “agentic”



Keep adding tools?

Business idea

- Merge the business of Google
 - Organize the world information
- Agentic intelligence
 - Take actions over time
- Natural convergence
 - Use LLMs to make an assistant

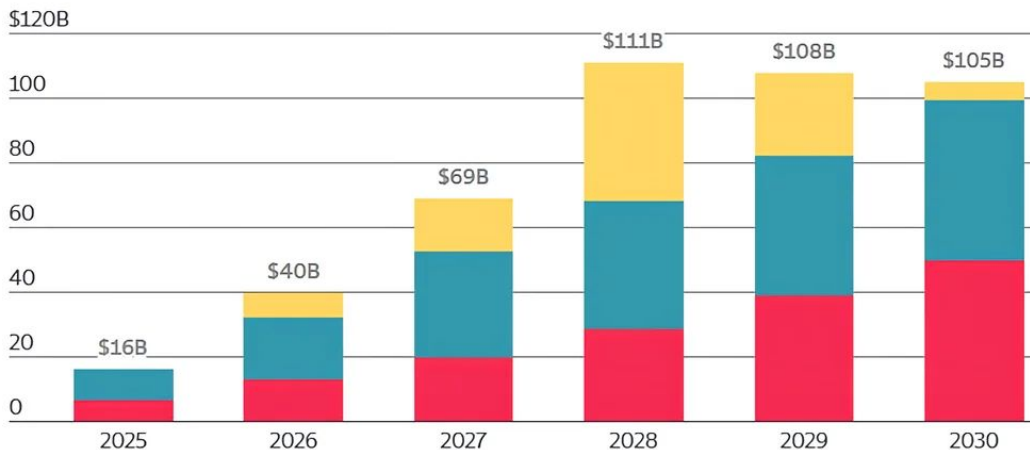
How do we cover costs

OpenAI's Cost to Compute



OpenAI plans to spend about \$450 billion renting servers through 2030, including \$100 billion in backup servers it says are 'monetizable.'

● Inference compute ● R&D compute (excl. non cash) ● Monetizable compute



Note: Numbers are projections and rounded

Source: The Information reporting

How do we cover costs

This is a billion
with a B

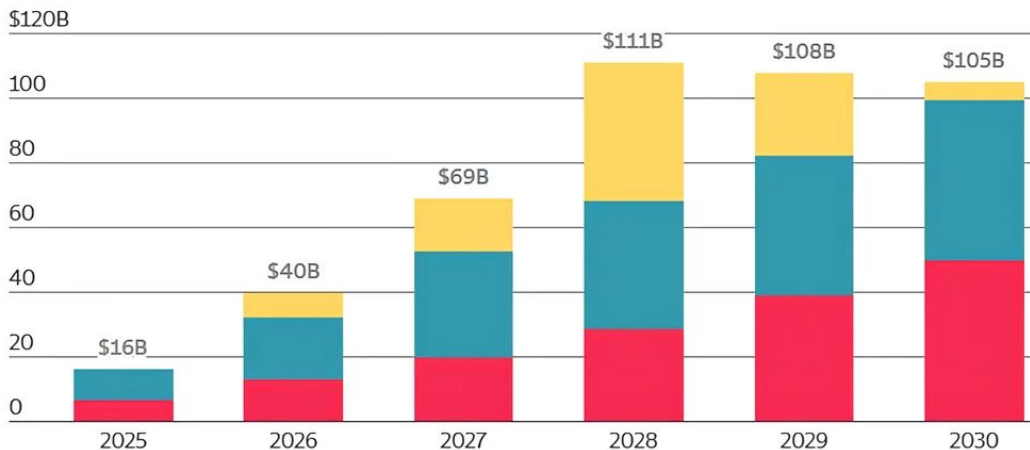


OpenAI's Cost to Compute



OpenAI plans to spend about \$450 billion renting servers through 2030, including \$100 billion in backup servers it says are 'monetizable.'

● Inference compute ● R&D compute (excl. non cash) ● Monetizable compute



Note: Numbers are projections and rounded

Source: The Information reporting

Solutions: Start making LLMs more “agentic”

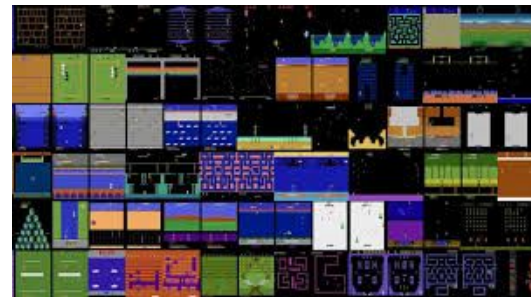
Chatbots -> Tools -> Assistants -> ??? -> AGI & Money



Can we still go back?



AlphaGo

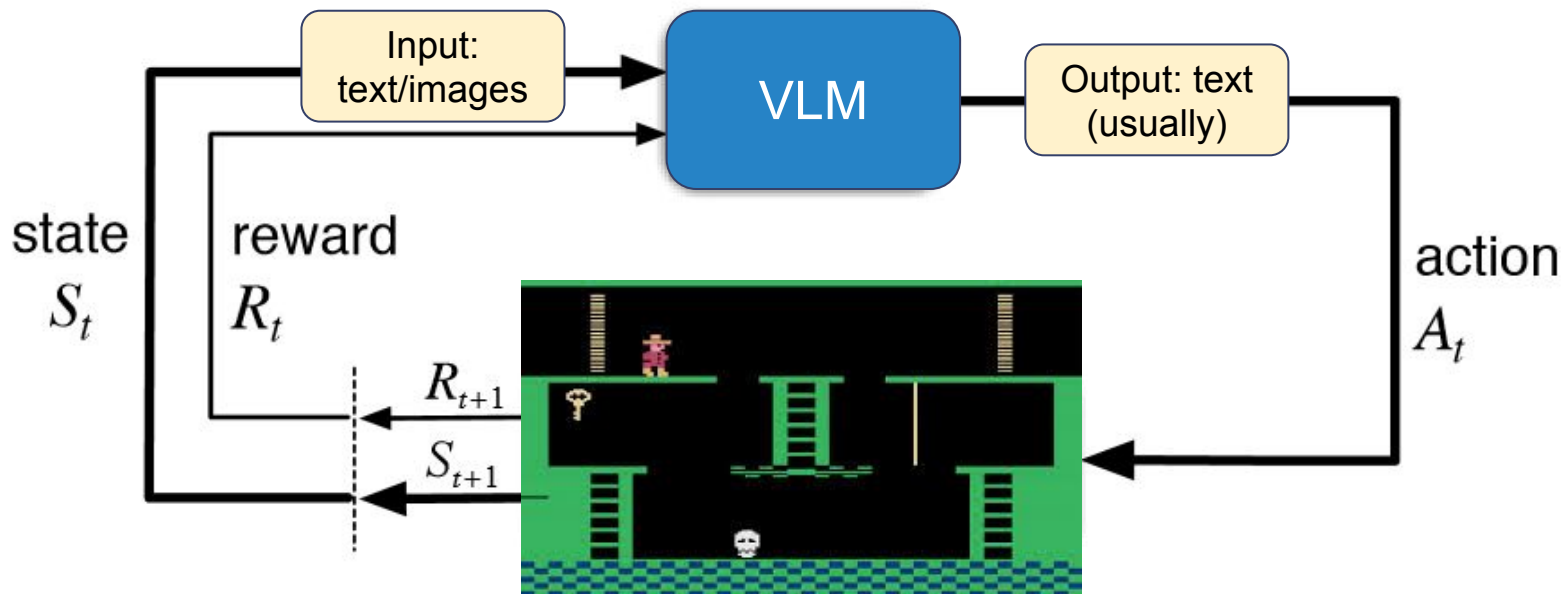


Atari



StarCraft

VLM Agents - Old DeepMind meets LLMs



Any Questions



Questions

Now for the presentations!