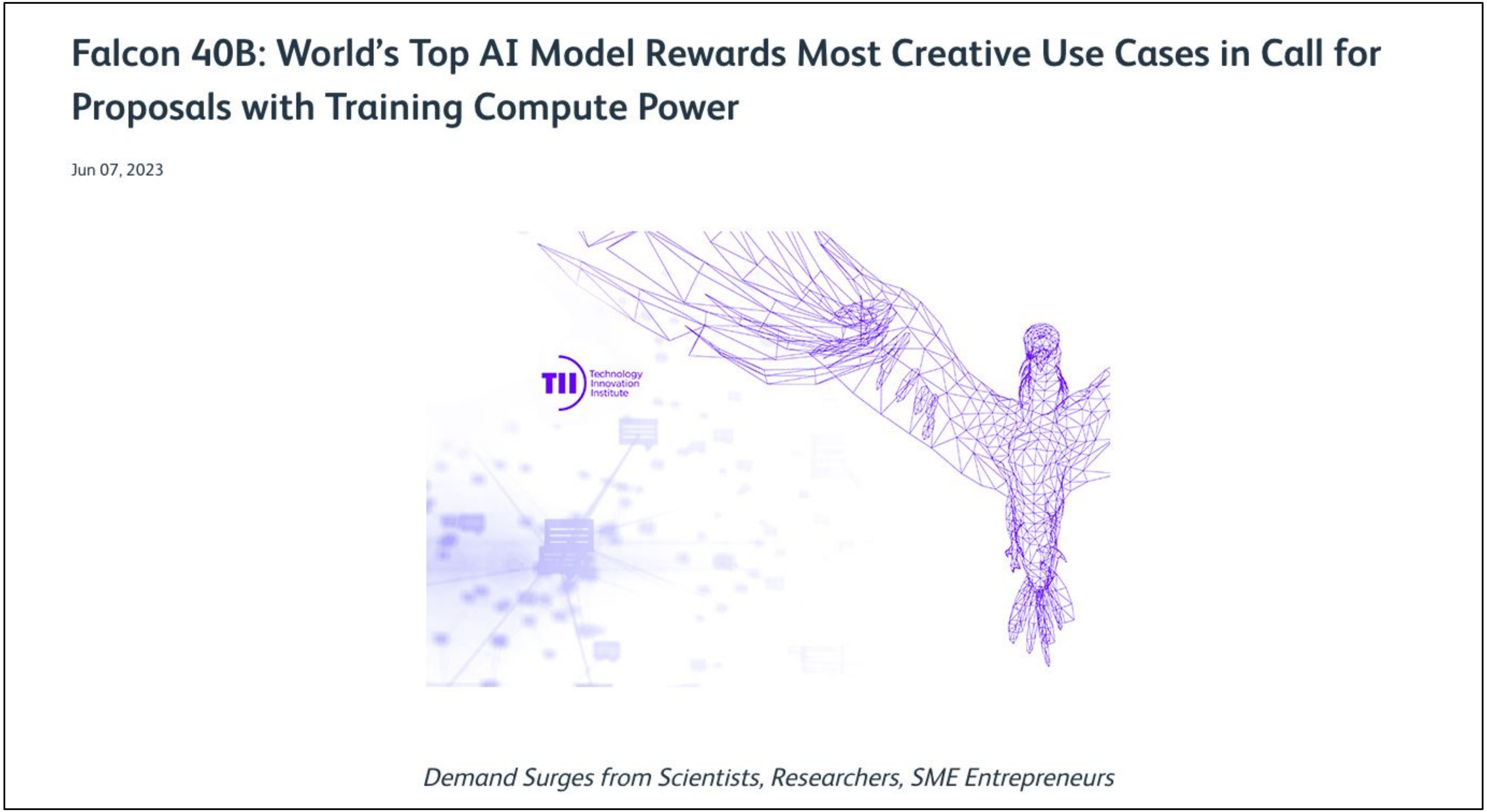
Falcon OS

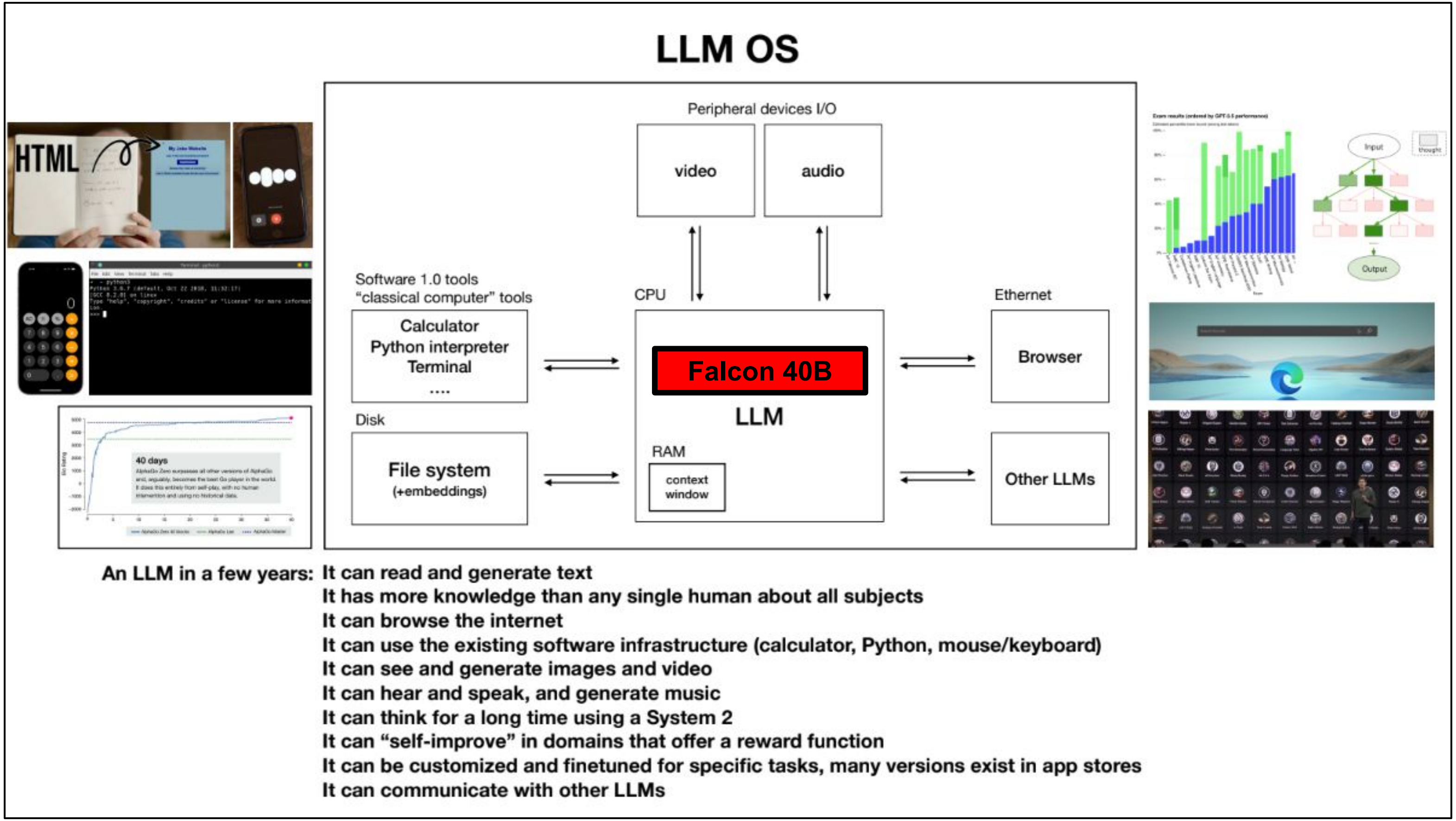
An open source LLM Operating System



Falcon 40B: World's Top AI Model Rewards Most Creative Use Cases in Call for Proposals with Training Compute Power

Jun 07, 2023

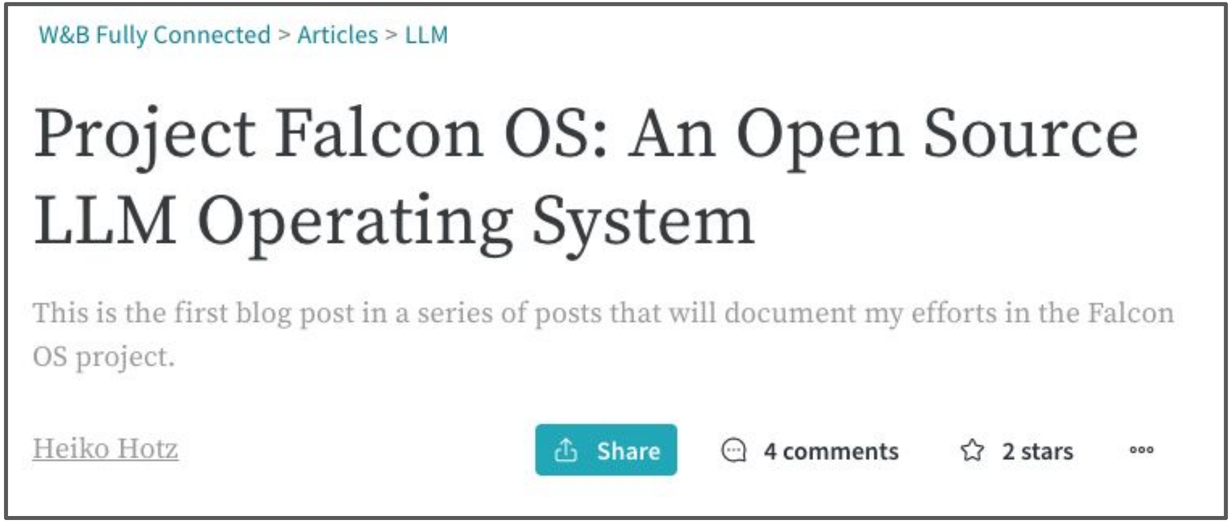
Demand Surges from Scientists, Researchers, SME Entrepreneurs



LLM OS

An LLM in a few years:
It can read and generate text
It has more knowledge than any single human about all subjects
It can browse the internet
It can use the existing software infrastructure (calculator, Python, mouse/keyboard)
It can see and generate images and video
It can hear and speak, and generate music
It can think for a long time using a System 2
It can "self-improve" in domains that offer a reward function
It can be customized and finetuned for specific tasks, many versions exist in app stores
It can communicate with other LLMs

Source: https://www.youtube.com/watch?v=zjkBMFhNj_g


W&B Fully Connected > Articles > LLM

Project Falcon OS: An Open Source LLM Operating System

This is the first blog post in a series of posts that will document my efforts in the Falcon OS project.

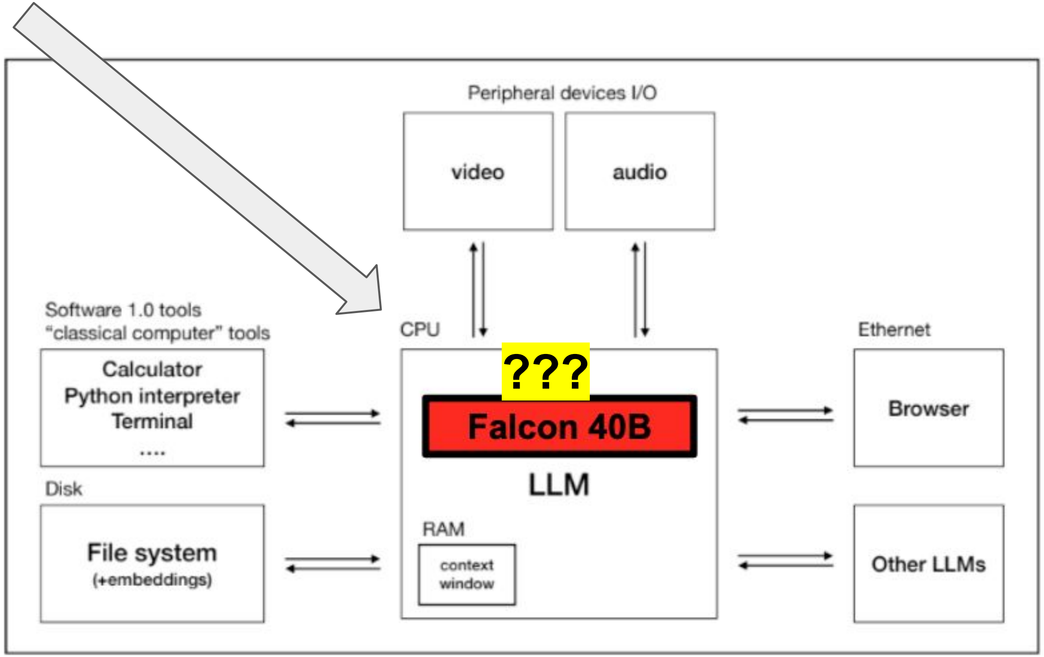
Heiko Hotz

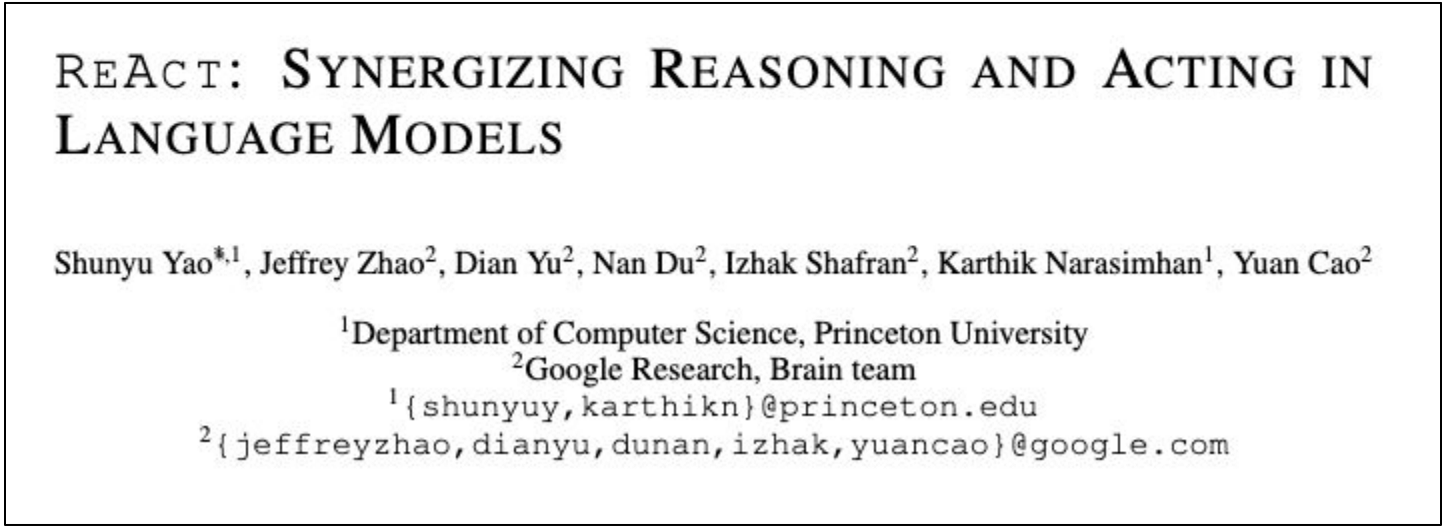
https://wandb.ai/falcon-os-wandb-team/falcon-os/reports/Project-Falcon-OS-An-Open-Source-LLM-Operating-System--Vmlldzo2MTYwMjE4

What's in this picture?

Who won the game last night?

Order a pizza!





REACT: SYNERGIZING REASONING AND ACTING IN LANGUAGE MODELS

Shunyu Yao*,[1], Jeffrey Zhao[2], Dian Yu[2], Nan Du[2], Izhak Shafran[2], Karthik Narasimhan[1], Yuan Cao[2]

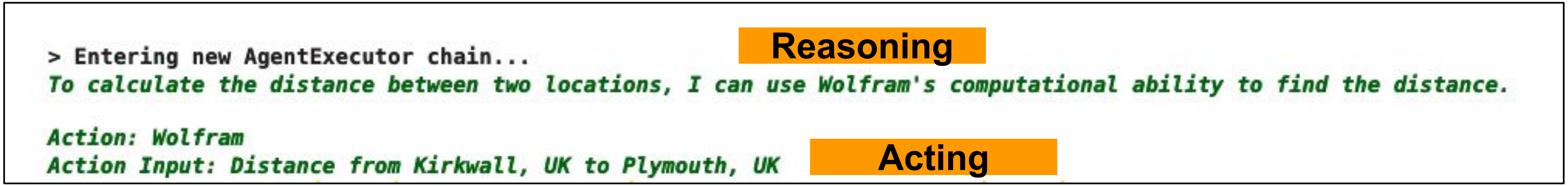
[1]Department of Computer Science, Princeton University
[2]Google Research, Brain team
[1]{shunyuy,karthikn}@princeton.edu
[2]{jeffreyzhao,dianyu,dunan,izhak,yuancao}@google.com

Source: https://arxiv.org/abs/2210.03629

DEMO TIME 🍿

User query: `How far is Kirkwall (UK) from Plymouth (UK)?`

With GPT-4:



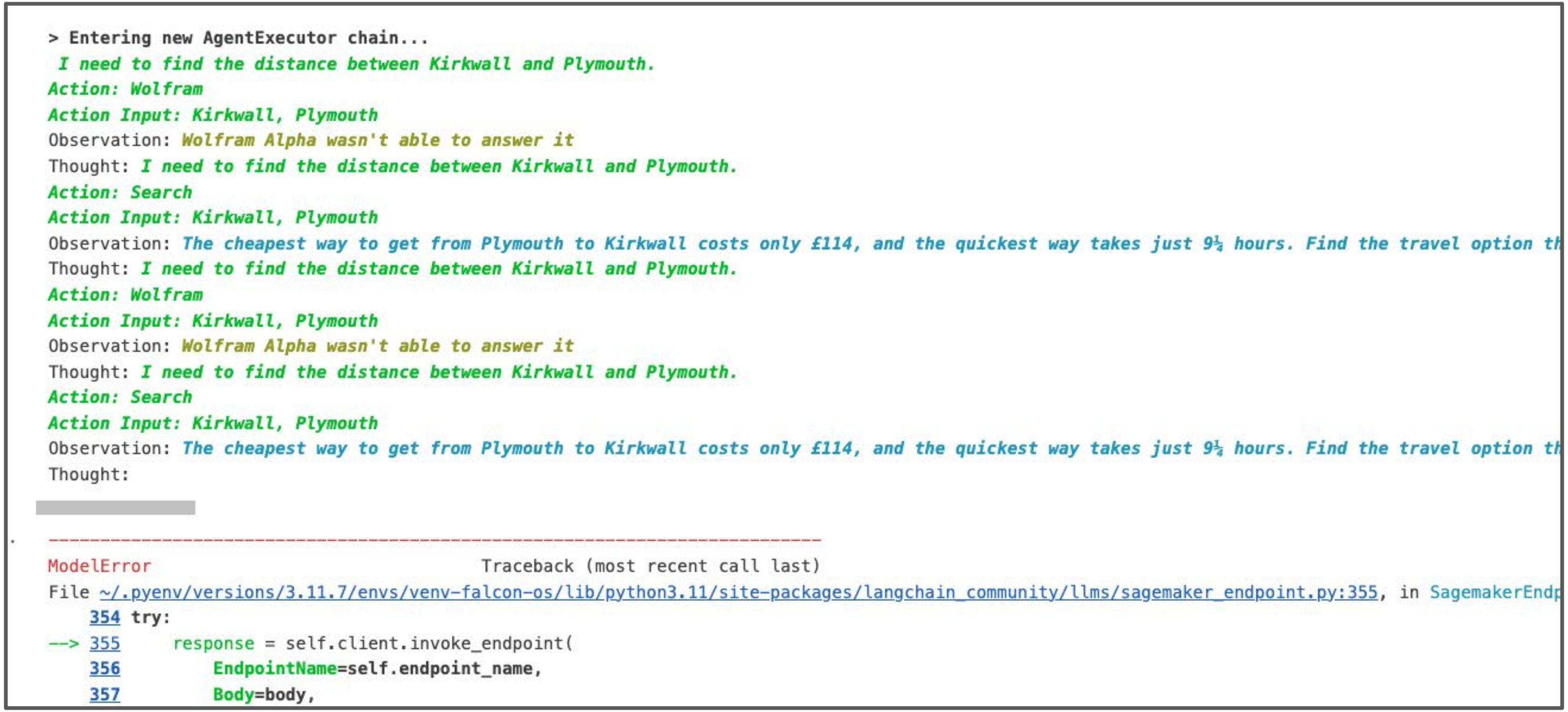
```
> Entering new AgentExecutor chain...
To calculate the distance between two locations, I can use Wolfram's computational ability to find the distance.

Action: Wolfram
Action Input: Distance from Kirkwall, UK to Plymouth, UK
```

Is (vanilla) Falcon 40B capable of this type of reasoning and acting?

Is it fit for ReAct???

With Falcon 40B:

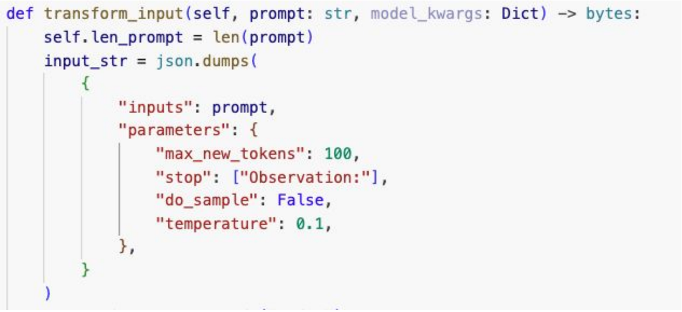


```
> Entering new AgentExecutor chain...
 I need to find the distance between Kirkwall and Plymouth.
Action: Wolfram
Action Input: Kirkwall, Plymouth
Observation: Wolfram Alpha wasn't able to answer it
Thought: I need to find the distance between Kirkwall and Plymouth.
Action: Search
Action Input: Kirkwall, Plymouth
Observation: The cheapest way to get from Plymouth to Kirkwall costs only £114, and the quickest way takes just 9¼ hours. Find the travel option th
Thought: I need to find the distance between Kirkwall and Plymouth.
Action: Wolfram
Action Input: Kirkwall, Plymouth
Observation: Wolfram Alpha wasn't able to answer it
Thought: I need to find the distance between Kirkwall and Plymouth.
Action: Search
Action Input: Kirkwall, Plymouth
Observation: The cheapest way to get from Plymouth to Kirkwall costs only £114, and the quickest way takes just 9¼ hours. Find the travel option th
Thought:

---------------------------------------------------------------------------
ModelError                                Traceback (most recent call last)
File ~/.pyenv/versions/3.11.7/envs/venv-falcon-os/lib/python3.11/site-packages/langchain_community/llms/sagemaker_endpoint.py:355, in SagemakerEndp
    354 try:
--> 355     response = self.client.invoke_endpoint(
    356         EndpointName=self.endpoint_name,
    357         Body=body,
```

Levers

Inference parameters

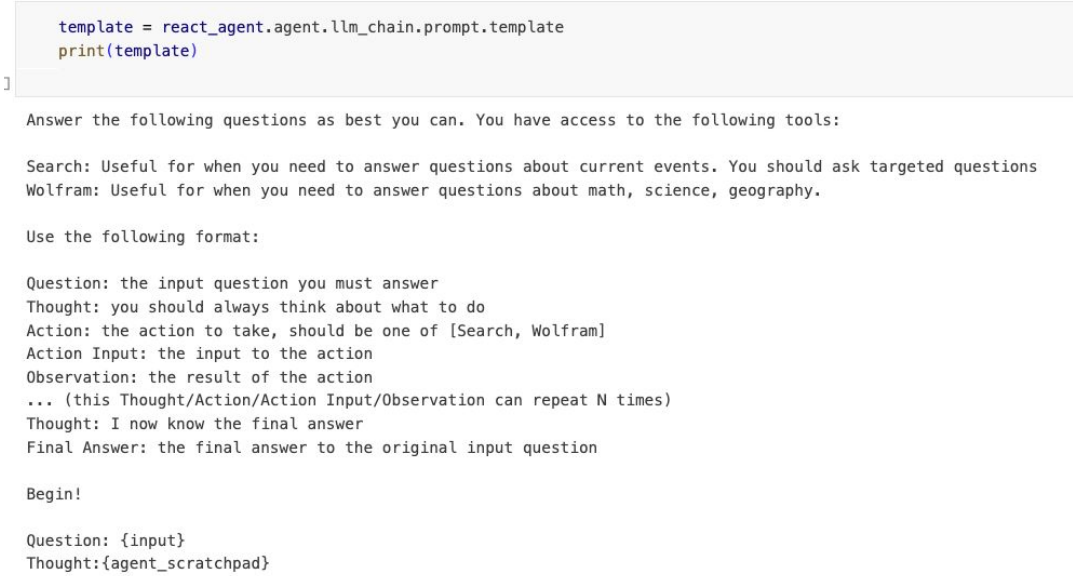


```
def transform_input(self, prompt: str, model_kwargs: Dict) -> bytes:
    self.len_prompt = len(prompt)
    input_str = json.dumps(
        {
            "inputs": prompt,
            "parameters": {
                "max_new_tokens": 100,
                "stop": ["Observation:"],
                "do_sample": False,
                "temperature": 0.1,
            },
        }
    )
```

Temperature, Top K,
Top P, …

Different wording,
few-shot examples,
…

Prompt template



```
    template = react_agent.agent.llm_chain.prompt.template
    print(template)
]

Answer the following questions as best you can. You have access to the following tools:

Search: Useful for when you need to answer questions about current events. You should ask targeted questions
Wolfram: Useful for when you need to answer questions about math, science, geography.

Use the following format:

Question: the input question you must answer
Thought: you should always think about what to do
Action: the action to take, should be one of [Search, Wolfram]
Action Input: the input to the action
Observation: the result of the action
... (this Thought/Action/Action Input/Observation can repeat N times)
Thought: I now know the final answer
Final Answer: the final answer to the original input question

Begin!

Question: {input}
Thought:{agent_scratchpad}
```

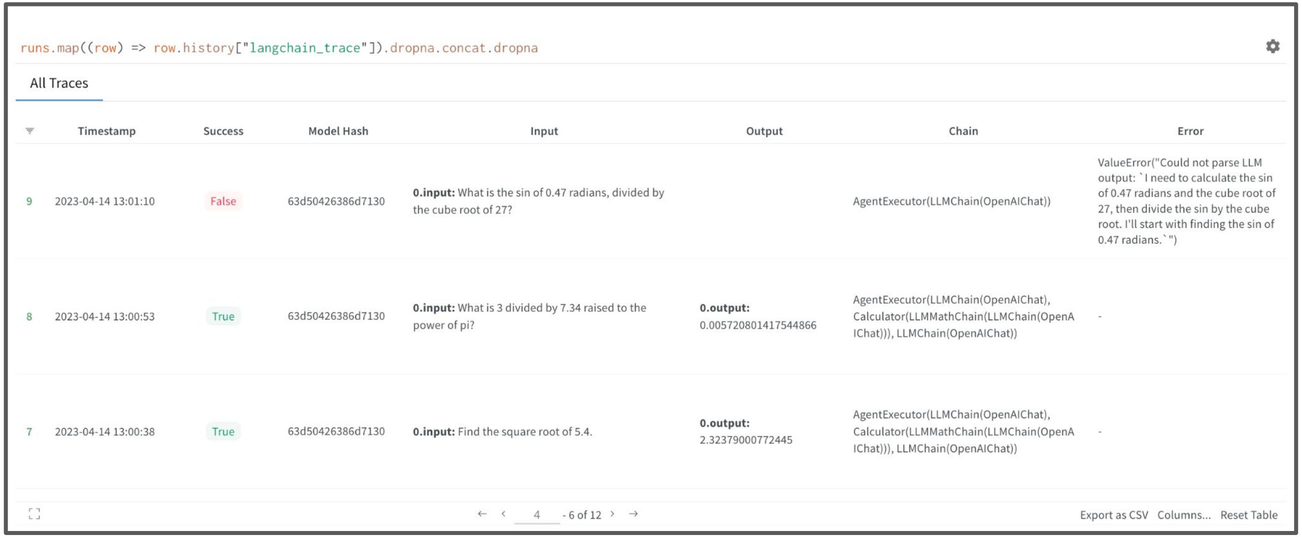
Hypothesis:
Training the model (aka fine-tuning) to "ReAct" will yield the best results

Challenges / open questions:

- **How to create the training dataset?**
  - **Crowd-sourcing?**
  - **Another LLM?**
- **What's the best format of the training dataset?**
  - **Single or multi turn examples? Or both?**
  - **Different examples for different tools? How many?**
- **Fine-tuning**
  - **With (Q)LoRA?**
  - **How many adapters? Where to store them? Do we need an adapter registry?**
  - **Can we combine adapters? Can we swap them out quickly?**
- **…**

Next steps

- Determining a first set of tools to work with (besides Web-Search and Wolfram Alpha)
- Identifying and quantifying the gaps when using ReAct with Falcon 40B (potentially with W&B’s prompt traces)



```
runs.map((row) => row.history["langchain_trace"]).dropna.concat.dropna
```

All Traces

	Timestamp	Success	Model Hash	Input	Output	Chain	Error
9	2023-04-14 13:01:10	False	63d50426386d7130	**0.input:** What is the sin of 0.47 radians, divided by the cube root of 27?		AgentExecutor(LLMChain(OpenAIChat))	ValueError("Could not parse LLM output: `I need to calculate the sin of 0.47 radians and the cube root of 27, then divide the sin by the cube root. I'll start with finding the sin of 0.47 radians.`")
8	2023-04-14 13:00:53	True	63d50426386d7130	**0.input:** What is 3 divided by 7.34 raised to the power of pi?	**0.output:** 0.005720801417544866	AgentExecutor(LLMChain(OpenAIChat)), Calculator(LLMMathChain(LLMChain(OpenAIChat))), LLMChain(OpenAIChat))	-
7	2023-04-14 13:00:38	True	63d50426386d7130	**0.input:** Find the square root of 5.4.	**0.output:** 2.32379000772445	AgentExecutor(LLMChain(OpenAIChat)), Calculator(LLMMathChain(LLMChain(OpenAIChat))), LLMChain(OpenAIChat))	-

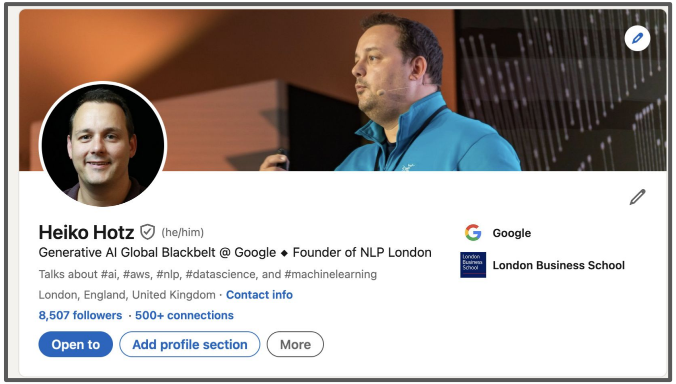
https://docs.wandb.ai/guides/prompts

Summary

- Falcon LLM as the core of an LLM OS
- -> Reasoning and Acting - ReAct
- Falcon is not there yet
- But maybe we can get it there through
    - Prompt Engineering
    - Experimentation
    - Fine-tuning

Thank you!

LinkedIn



Heiko Hotz (he/him)

Generative AI Global Blackbelt @ Google ◆ Founder of NLP London

Talks about #ai, #aws, #nlp, #datascience, and #machinelearning

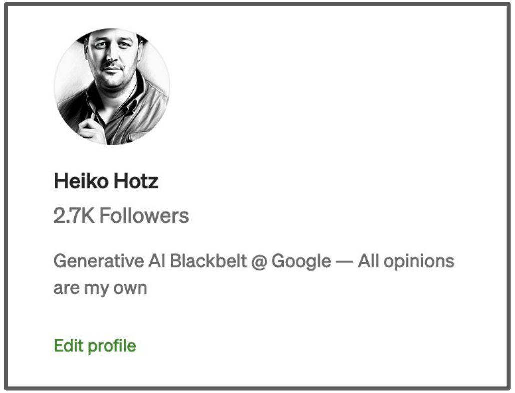
London, England, United Kingdom · Contact info

8,507 followers · 500+ connections

Google

London Business School

Medium



Heiko Hotz

2.7K Followers

Generative AI Blackbelt @ Google — All opinions are my own