Tips for building applications that use GenAI

December 10th, 2024

Juan Peredo

BOLBECK LLC

Juan Peredo - Your guide in this journey

- Architect / consultant / developer & everything in between
- Independent Cloud Application Architect / Entrepreneur
- Linkedin: linkedin.com/in/juanperedotech
- Over 15 years of IT experience in companies like:
  - Bolbeck LLC
  - AWS
  - Strategy& (PWC)
  - Booz & Co

AI can help reduce the time to code an App

GitHub Copilot

CURSOR

codeium

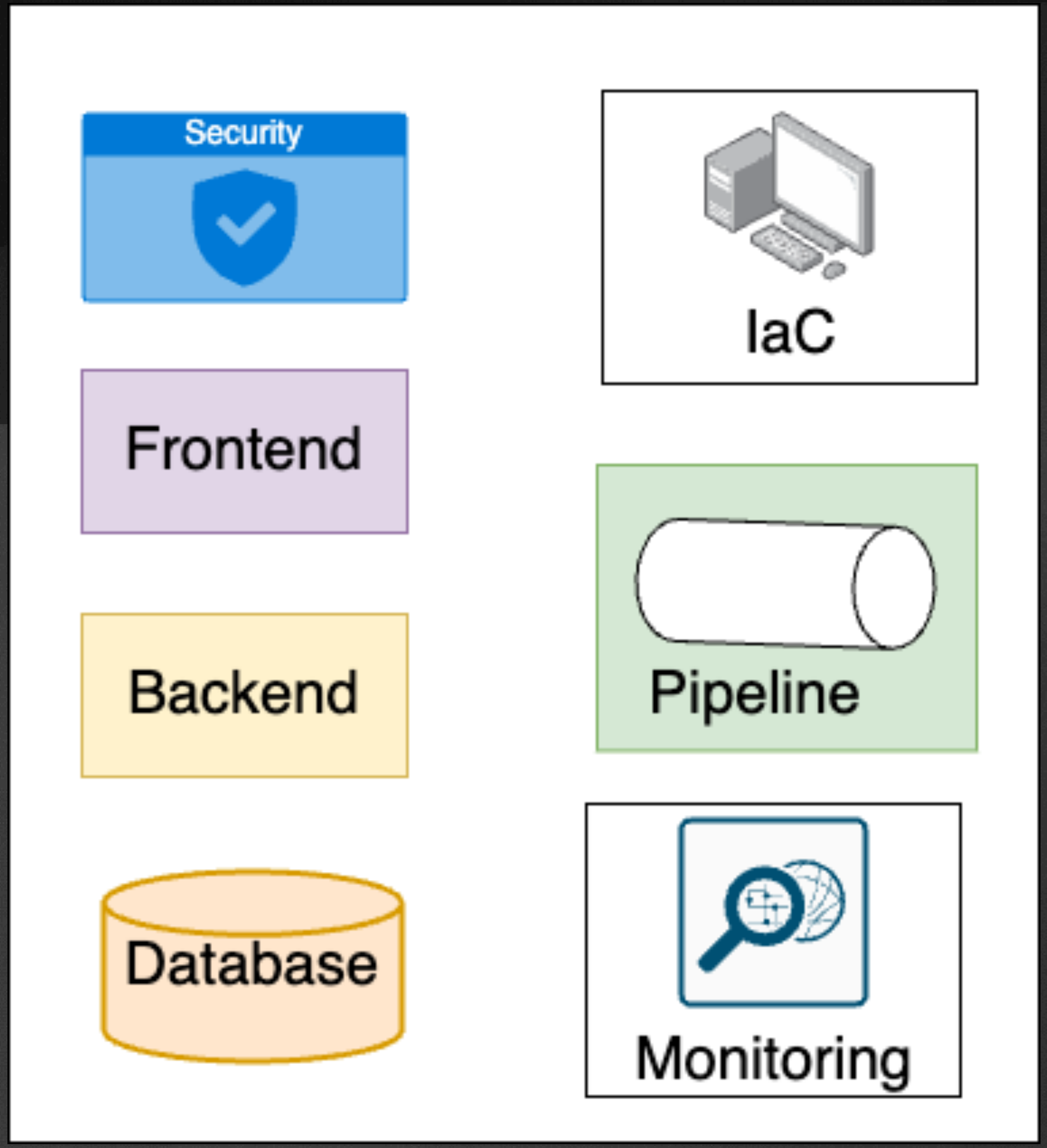
Amazon Q

Continue

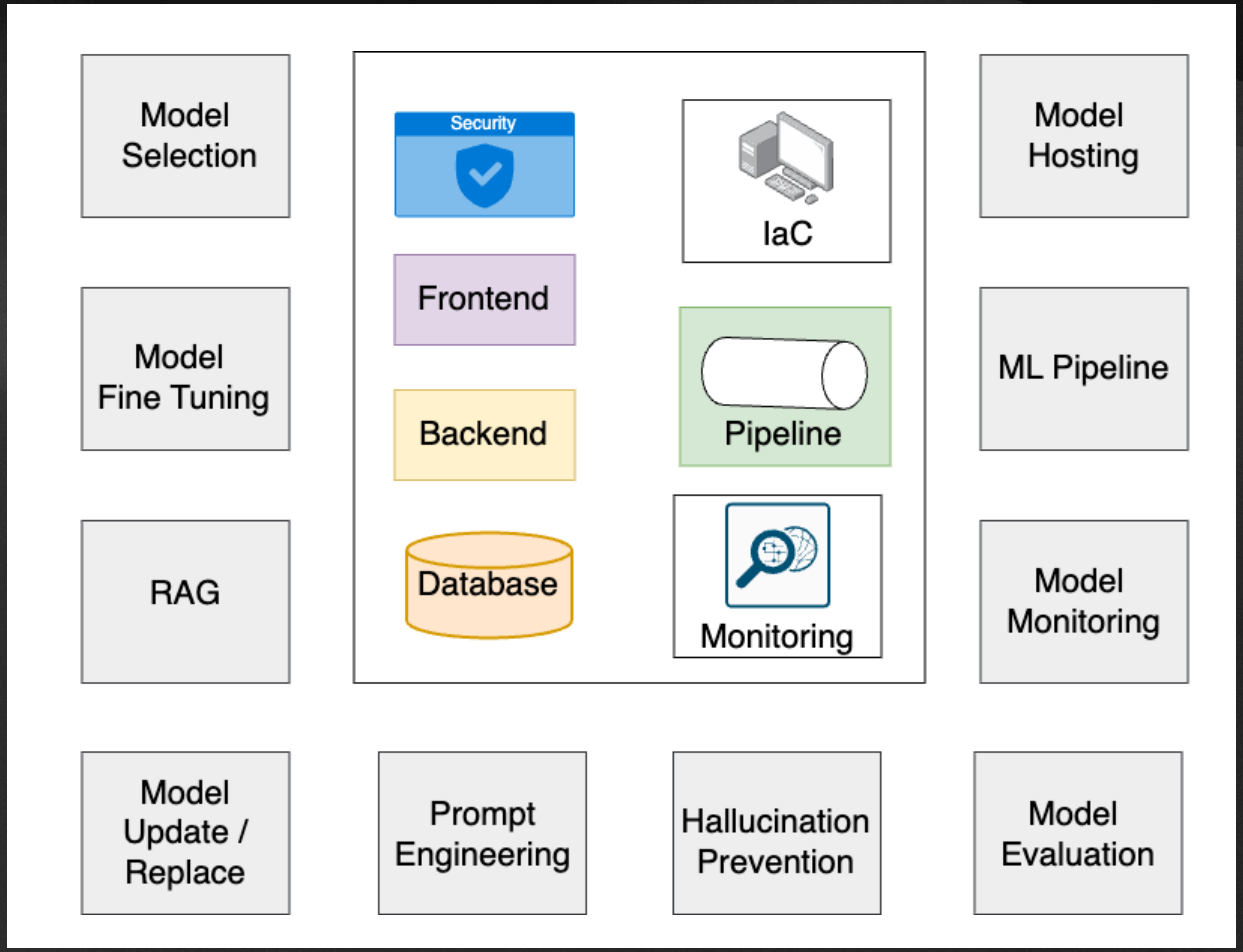
- These are plugins for your IDE except cursor that is an actual IDE
- You can also just use any LLM to help with your code via cut and paste

However, AI models also increase your app complexity

Typical app dev components



Typical AI app dev components



Where/How to host Models is important

Host locally

Ollama

vLLM

LLaMA C++

Host in the cloud

Modal

SkyPilot

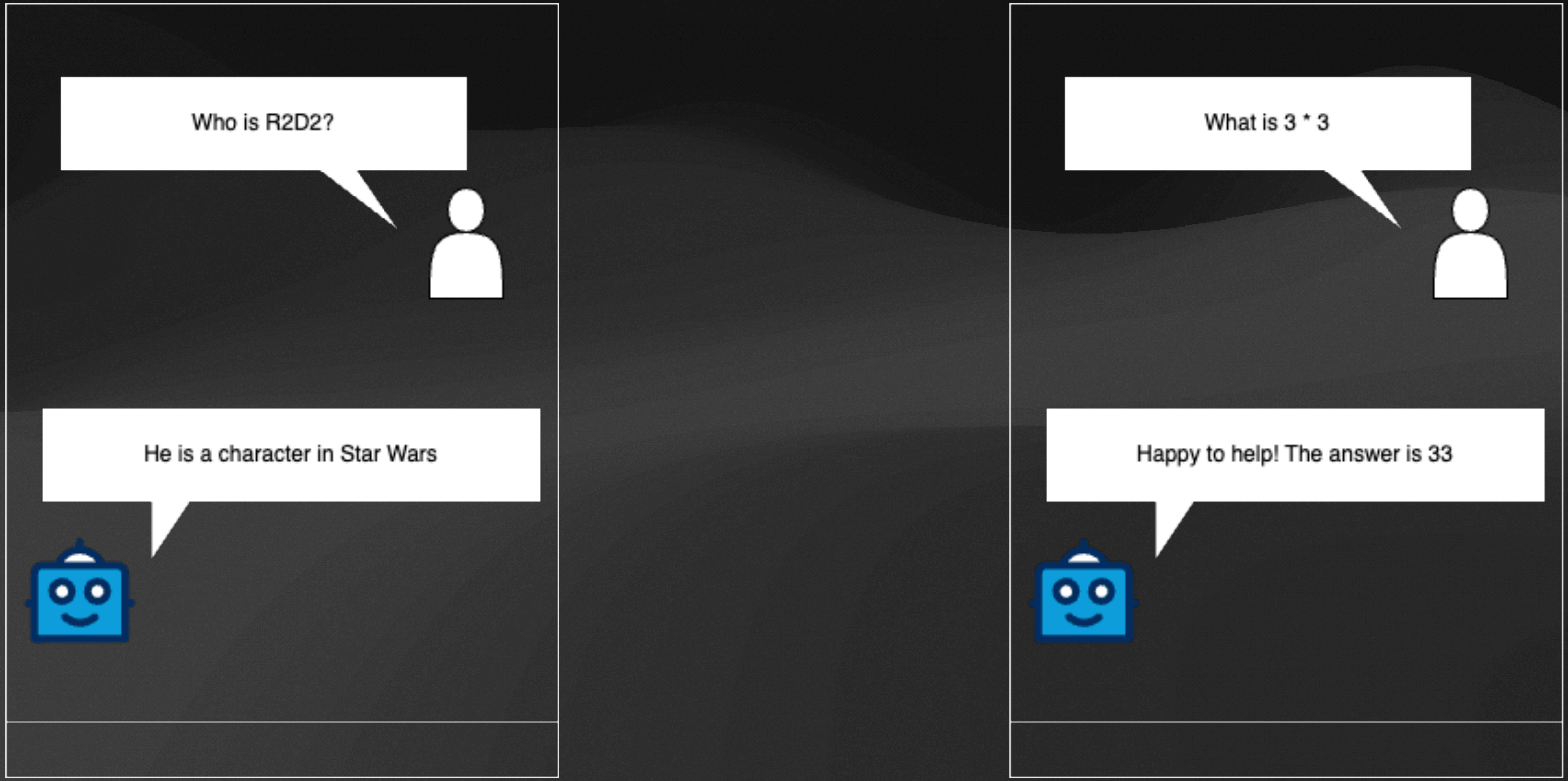
Google Cloud

aws

IBM Cloud

Alternatively, you could just use a library like HF Transformers

Building an AI chatbot is easy



Who is R2D2?

He is a character in Star Wars

What is 3 * 3

Happy to help! The answer is 33

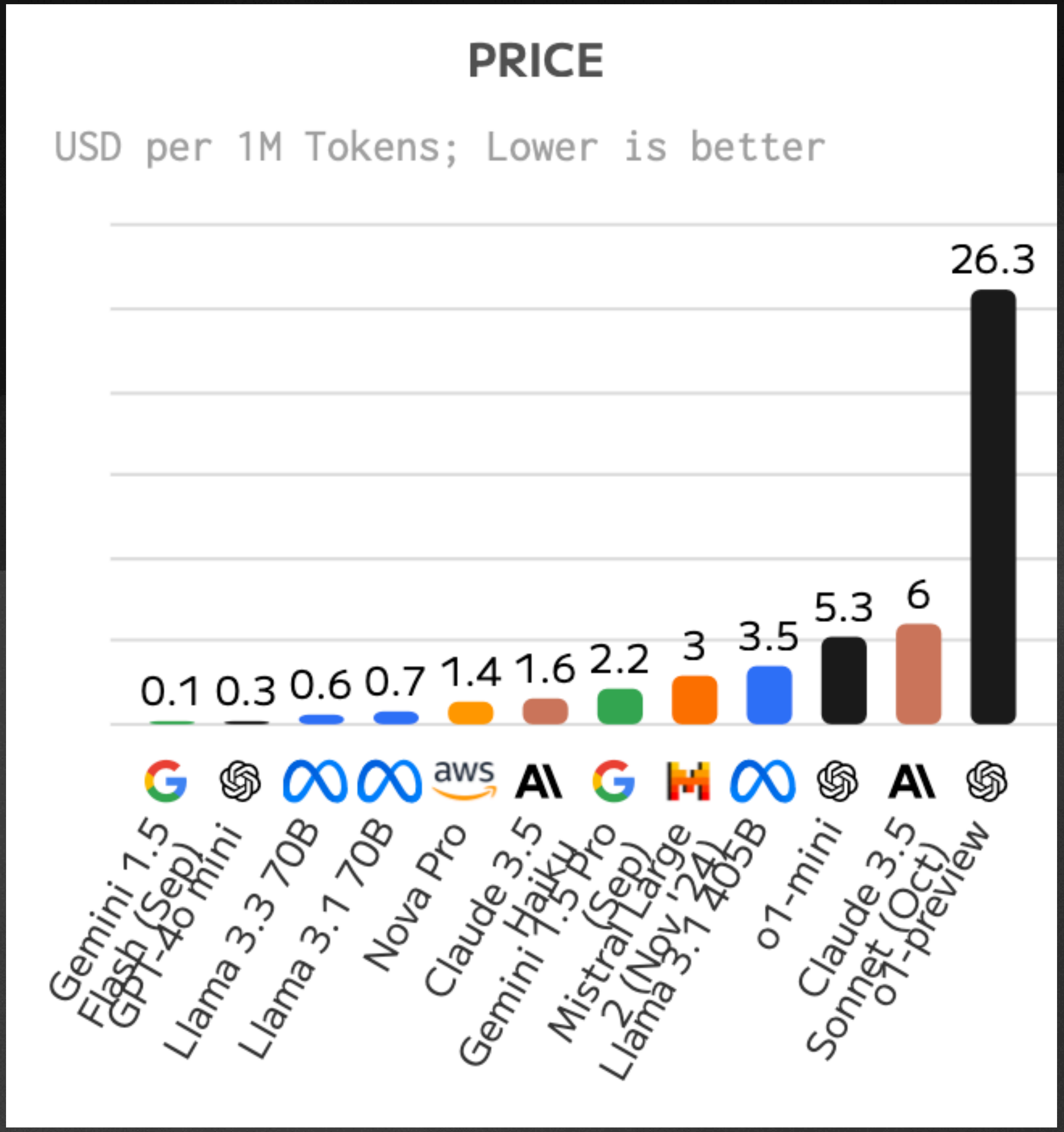
However, validating and moderating the chatbot content is hard

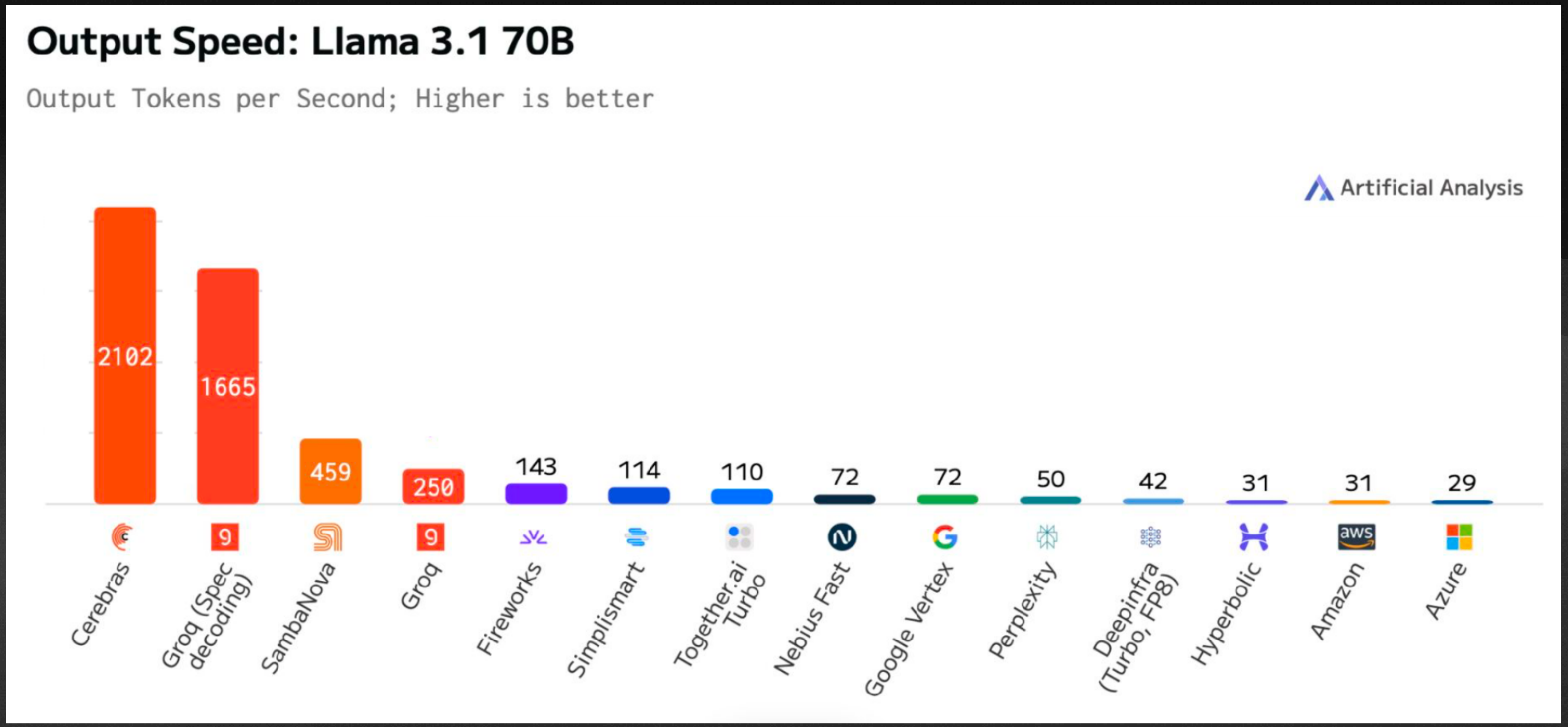
There are techniques to validate LLM output

Technique	Description	Technique Issues
Prompt Engineering	Provide guardrails to LLM in the prompt	• Need to be very precise • LLM May choose to ignore guardrails
Guardrails	Specialized LLMs that classifies answer as ‘bad’	• Adds additional latency to every LLM call • Adds cost • Is not always correct
RAG	Provide additional context to LLM as part of the prompt	• Need to add RAG store to solution • Highly dependent on the quality of the data provided • RAG may not provide enough information to get answers
Fine tuning	Provide additional training to an LLM	• Very expensive when compared to other methods above • Time consuming • No guaranties plus could degrade LLM

There are no techniques that guarantee chatbot answer is ‘good’

Choice of model and provider has a big impact

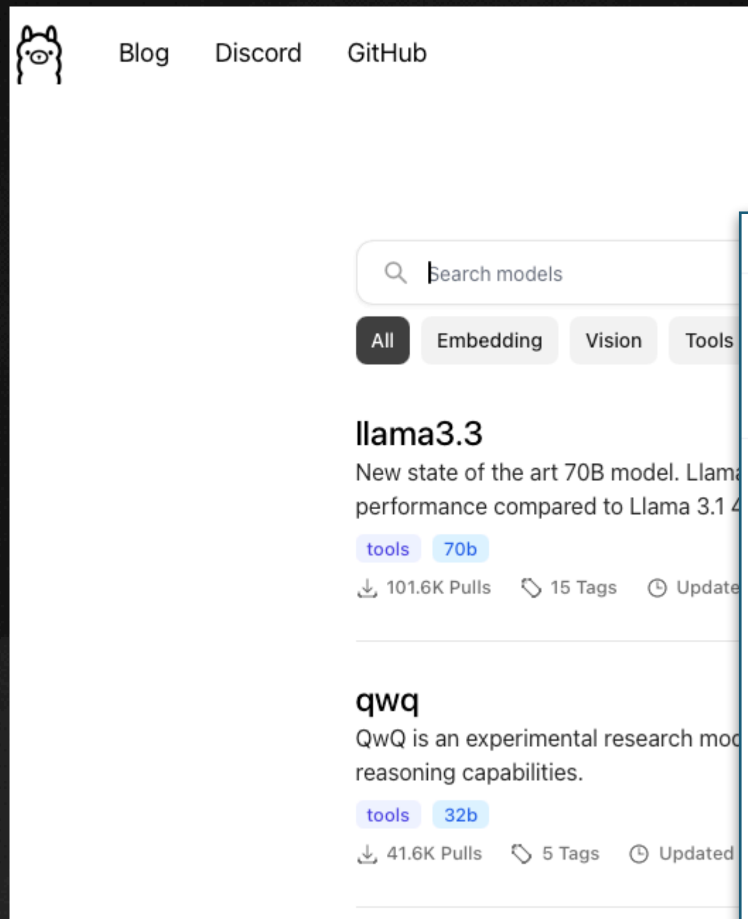




Small is beautiful and better for the environment (not to mention your wallet)

Charts source: Artificial Analysis

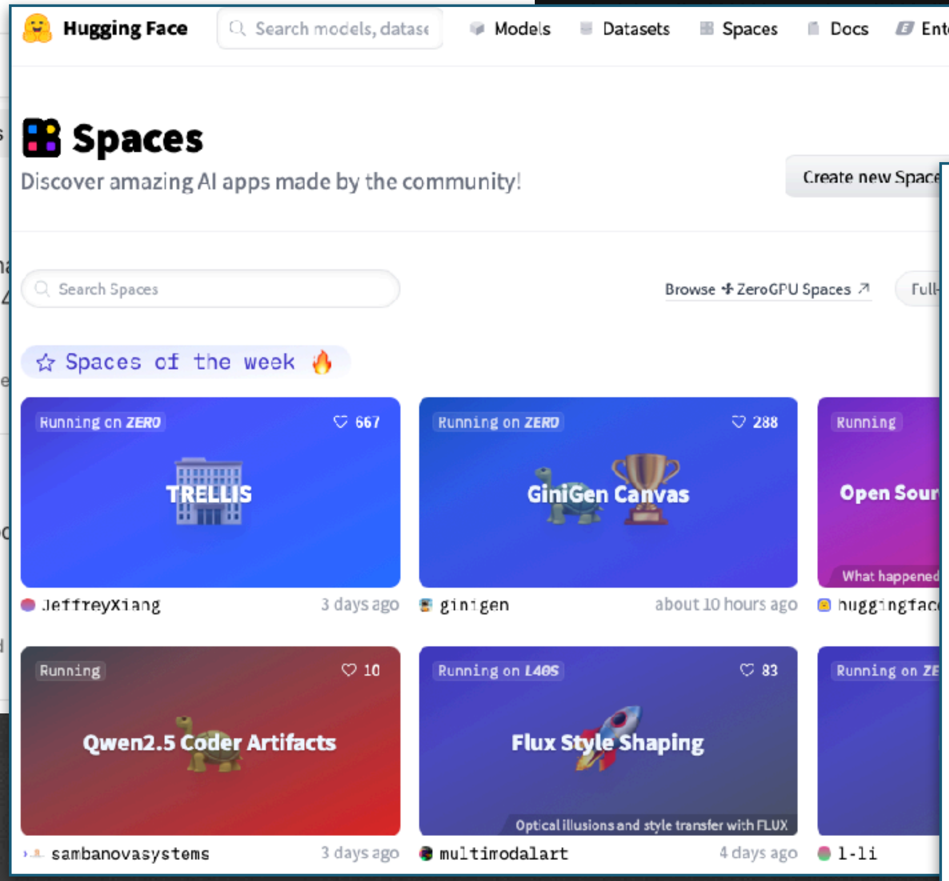
Evaluate models & choose the best for the use case



llama3.3
New state of the art 70B model. Llama
performance compared to Llama 3.1 4
tools 70b
101.6K Pulls 15 Tags Update

qwq
QwQ is an experimental research mod
reasoning capabilities.
tools 32b
41.6K Pulls 5 Tags Updated

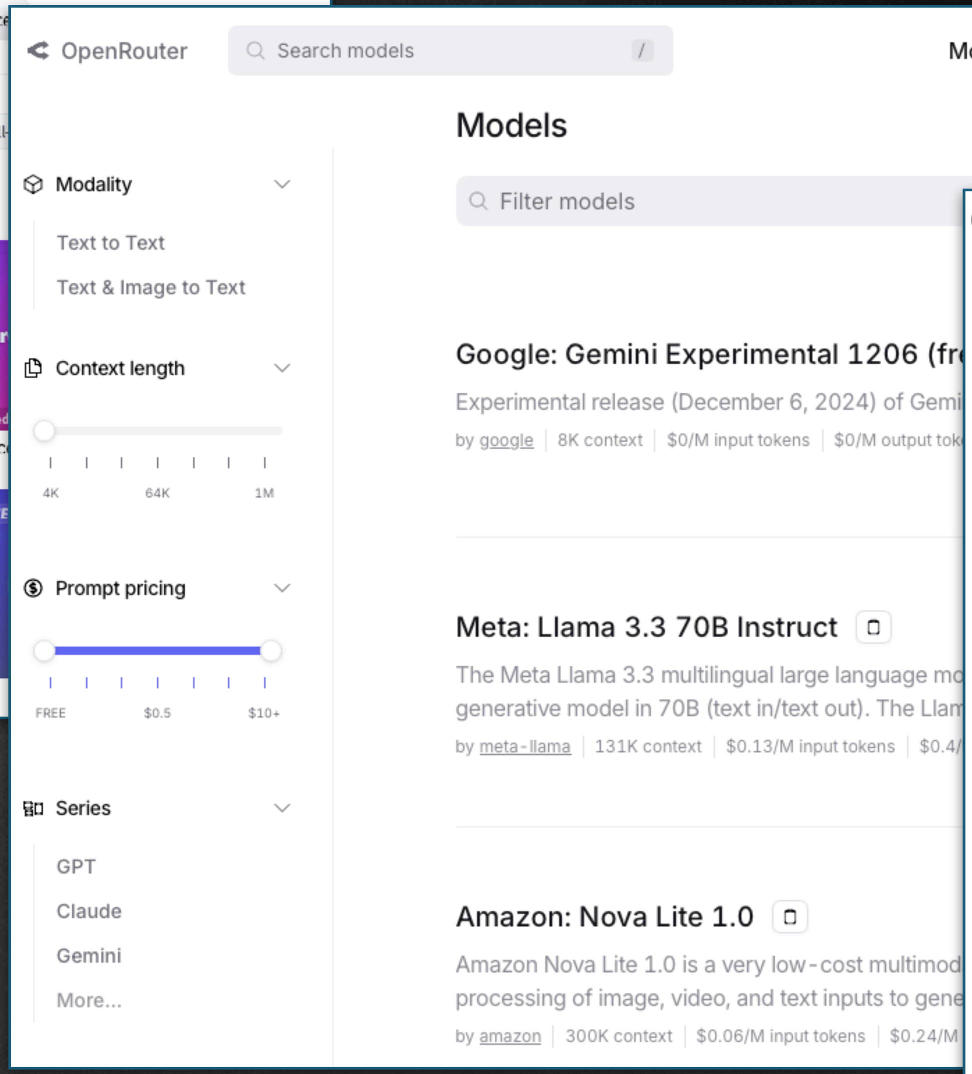
Ollama



Spaces
Discover amazing AI apps made by the community!

JeffreyXiang 3 days ago
ginigen about 10 hours ago
huggingfac
sambanovasystems 3 days ago
multimodalart 4 days ago
l-li

Hugging Face



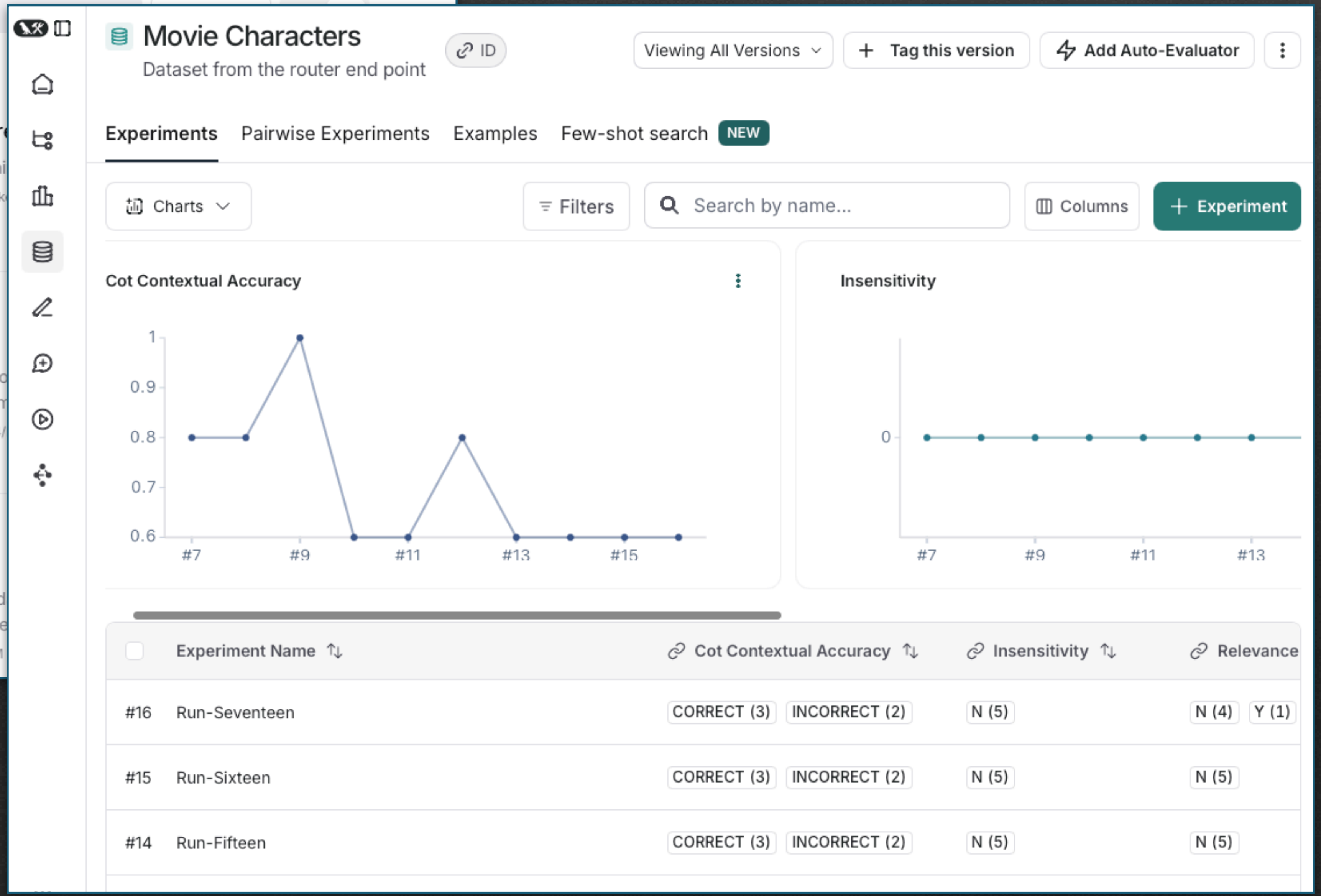
Models

Google: Gemini Experimental 1206 (fr
Experimental release (December 6, 2024) of Gemi
by google | 8K context | $0/M input tokens | $0/M output tok

Meta: Llama 3.3 70B Instruct
The Meta Llama 3.3 multilingual large language mo
generative model in 70B (text in/text out). The Llam
by meta-llama | 131K context | $0.13/M input tokens | $0.4/

Amazon: Nova Lite 1.0
Amazon Nova Lite 1.0 is a very low-cost multimod
processing of image, video, and text inputs to gene
by amazon | 300K context | $0.06/M input tokens | $0.24/M

OpenRouter



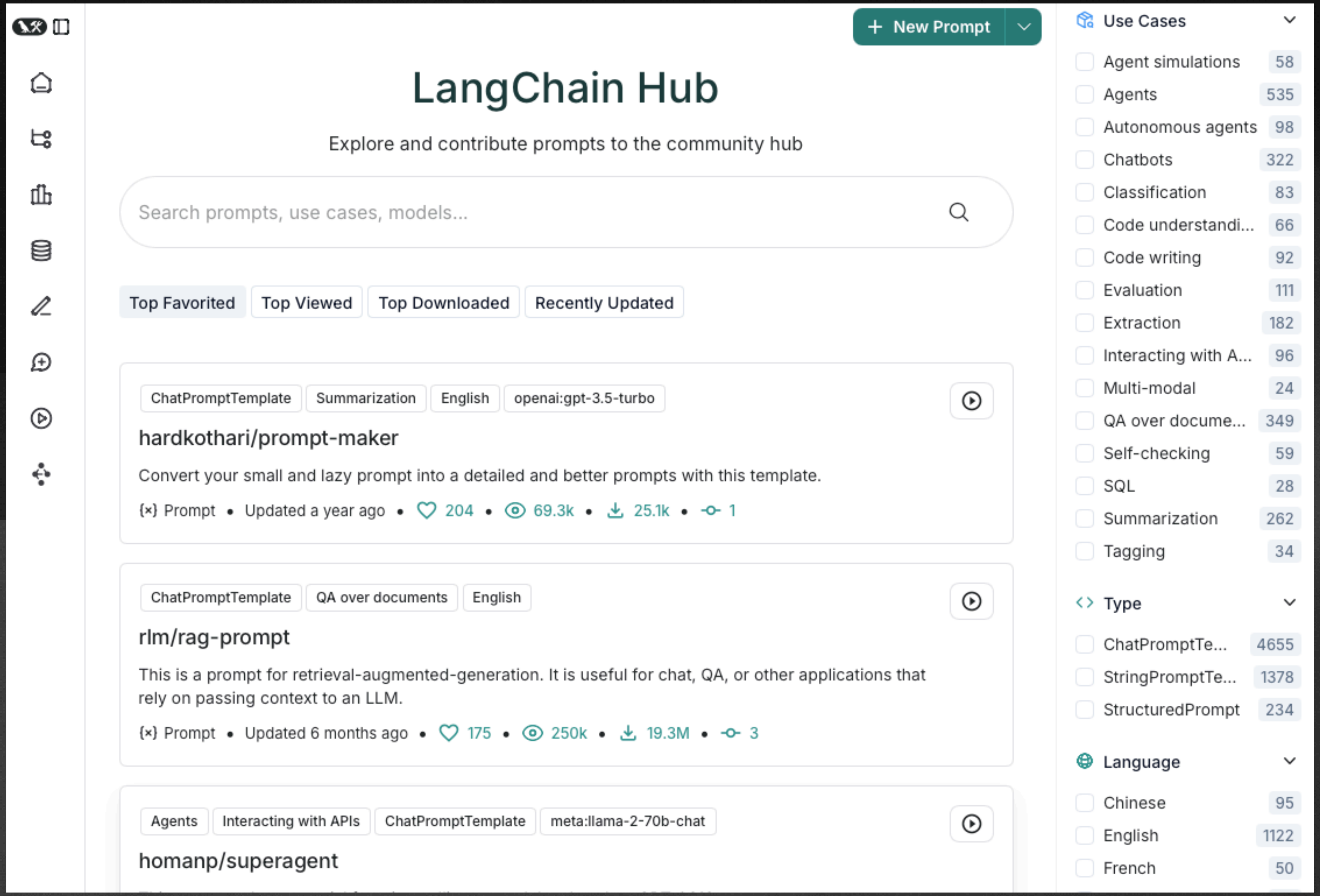
Movie Characters
Dataset from the router end point

	Experiment Name	Cot Contextual Accuracy	Insensitivity	Relevance
#16	Run-Seventeen	CORRECT (3) INCORRECT (2)	N (5)	N (4) Y (1)
#15	Run-Sixteen	CORRECT (3) INCORRECT (2)	N (5)	N (5)
#14	Run-Fifteen	CORRECT (3) INCORRECT (2)	N (5)	N (5)

LangSmith

Evaluation must take place at each step of the way

Externalize your prompts



LangChain Hub

Explore and contribute prompts to the community hub

Search prompts, use cases, models...

Top Favorited | Top Viewed | Top Downloaded | Recently Updated

ChatPromptTemplate · Summarization · English · openai:gpt-3.5-turbo

hardkothari/prompt-maker

Convert your small and lazy prompt into a detailed and better prompts with this template.

{x} Prompt • Updated a year ago • ♡ 204 • 69.3k • 25.1k • 1

ChatPromptTemplate · QA over documents · English

rlm/rag-prompt

This is a prompt for retrieval-augmented-generation. It is useful for chat, QA, or other applications that rely on passing context to an LLM.

{x} Prompt • Updated 6 months ago • ♡ 175 • 250k • 19.3M • 3

Agents · Interacting with APIs · ChatPromptTemplate · meta:llama-2-70b-chat

homanp/superagent

Use Cases	
□ Agent simulations	58
□ Agents	535
□ Autonomous agents	98
□ Chatbots	322
□ Classification	83
□ Code understandi...	66
□ Code writing	92
□ Evaluation	111
□ Extraction	182
□ Interacting with A...	96
□ Multi-modal	24
□ QA over docume...	349
□ Self-checking	59
□ SQL	28
□ Summarization	262
□ Tagging	34

Type	
□ ChatPromptTe...	4655
□ StringPromptTe...	1378
□ StructuredPrompt	234

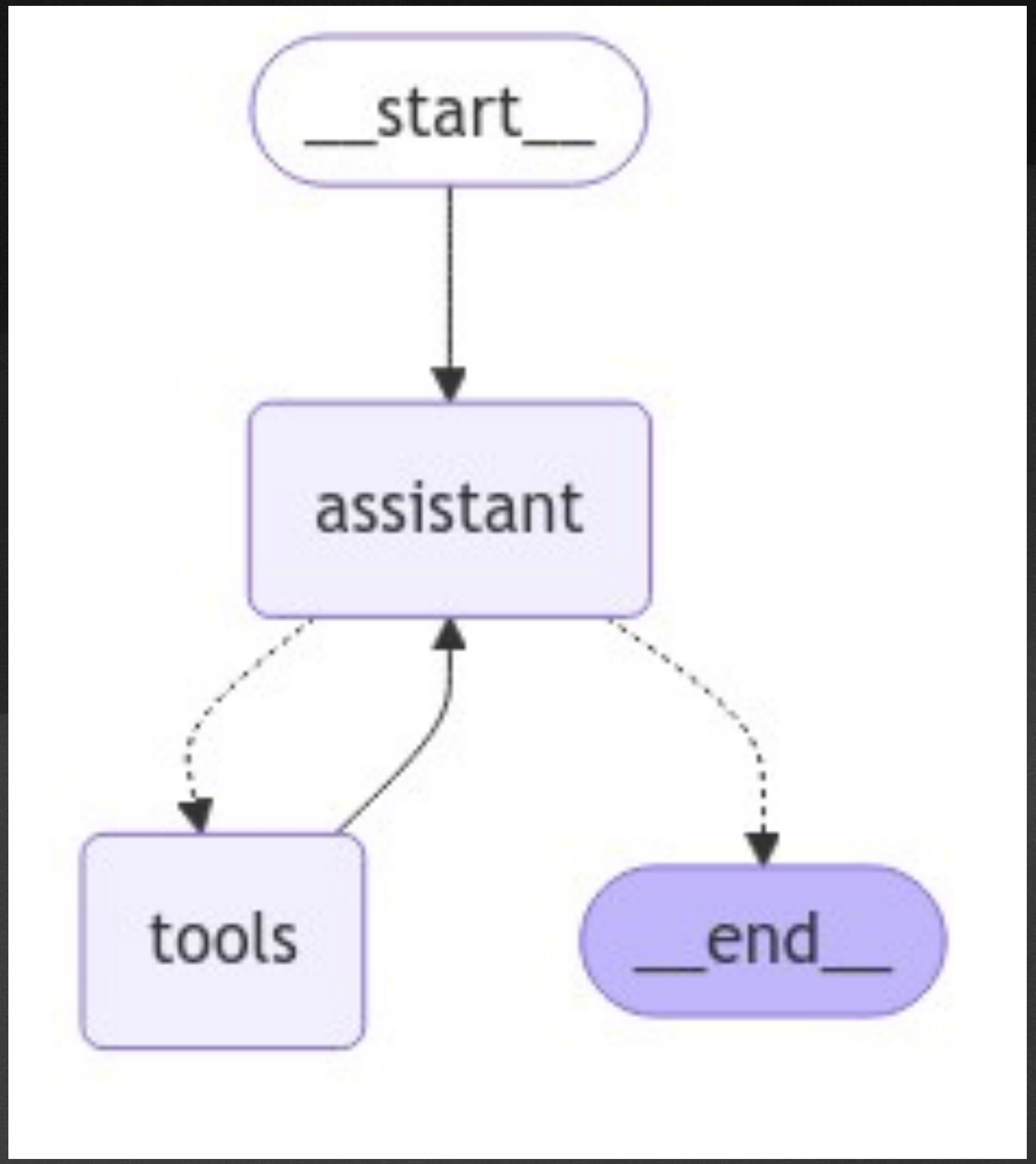
Language	
□ Chinese	95
□ English	1122
□ French	50

Externalizing prompts facilitates:

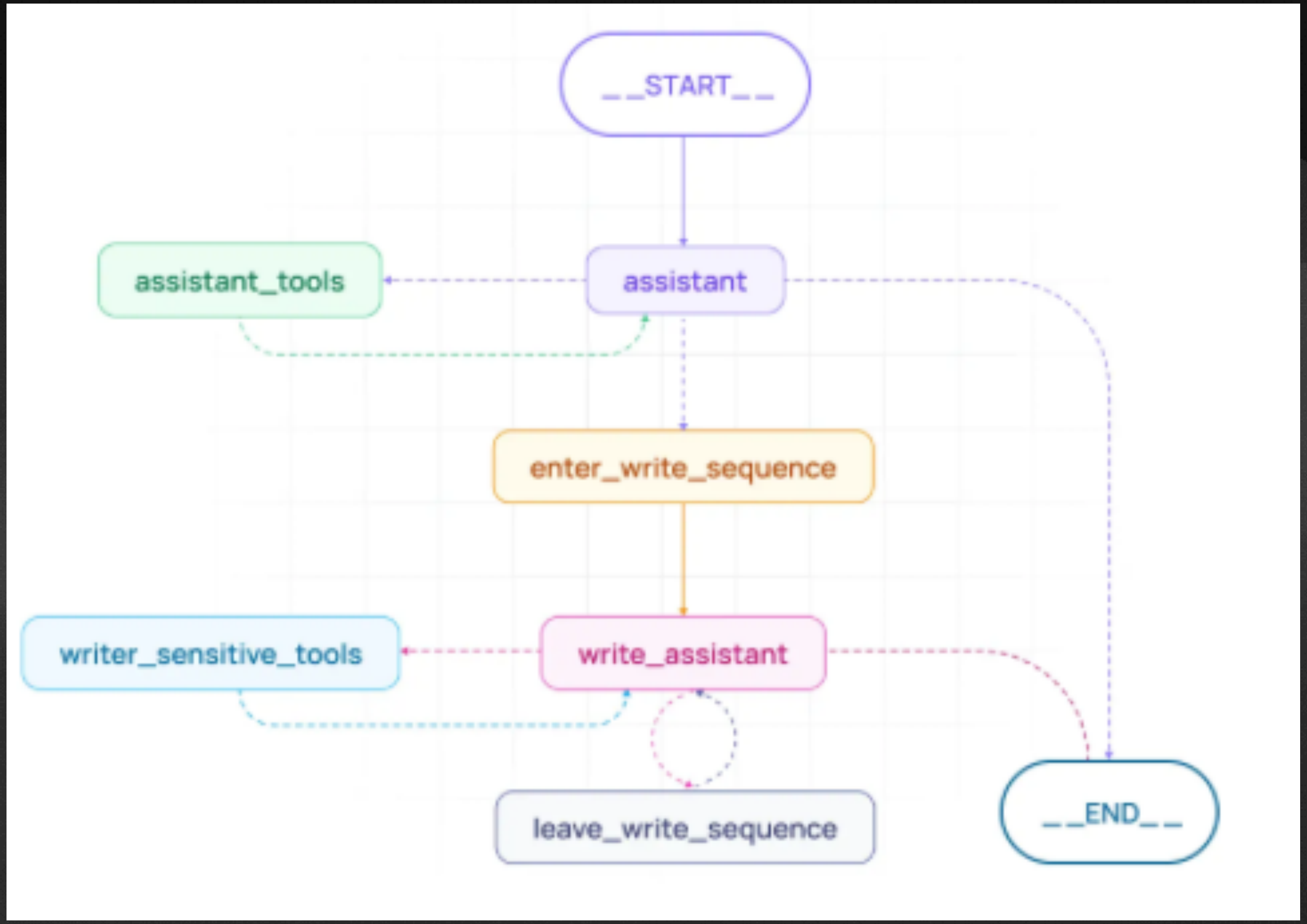
- Expert input
- Future proofing
- Faster development

Prompts should not be embedded in the codebase

Use agents to go beyond the chatbot



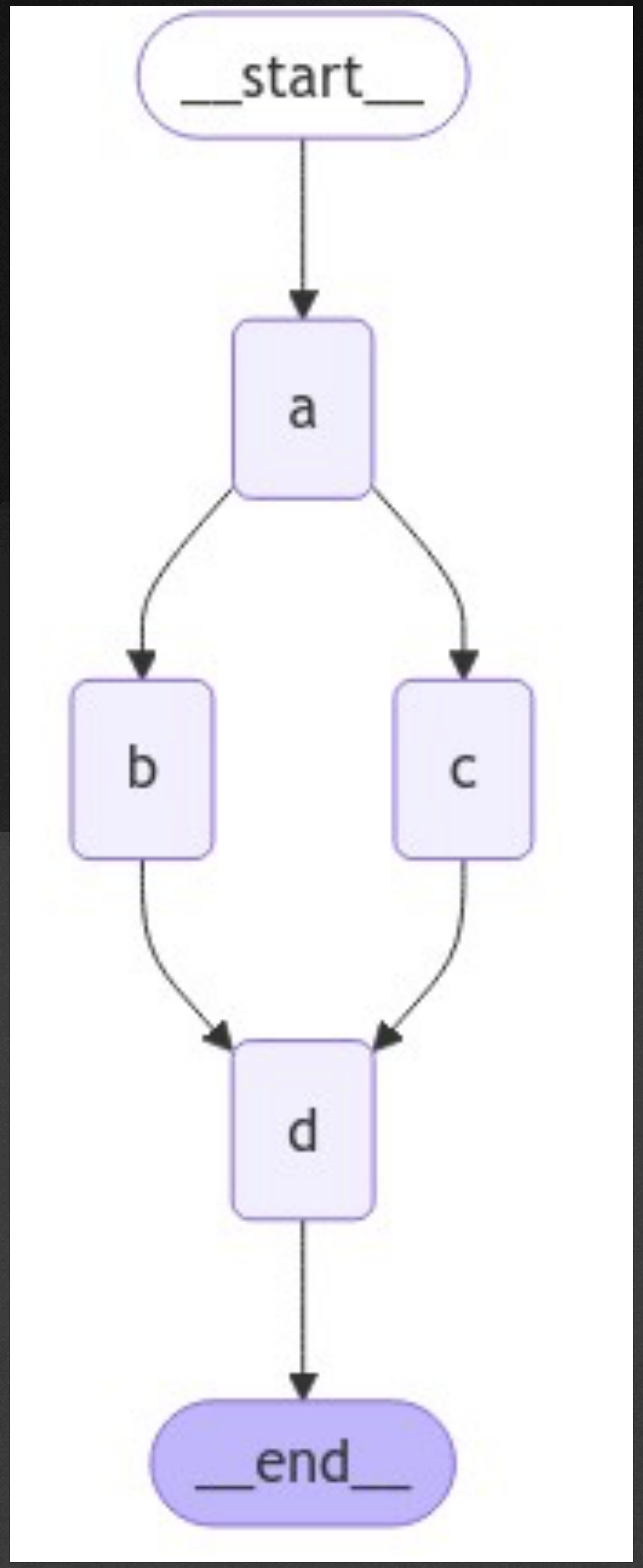
Simple tool calling agent



Simple assistant agent

Agents allow LLMs to interact with the real world

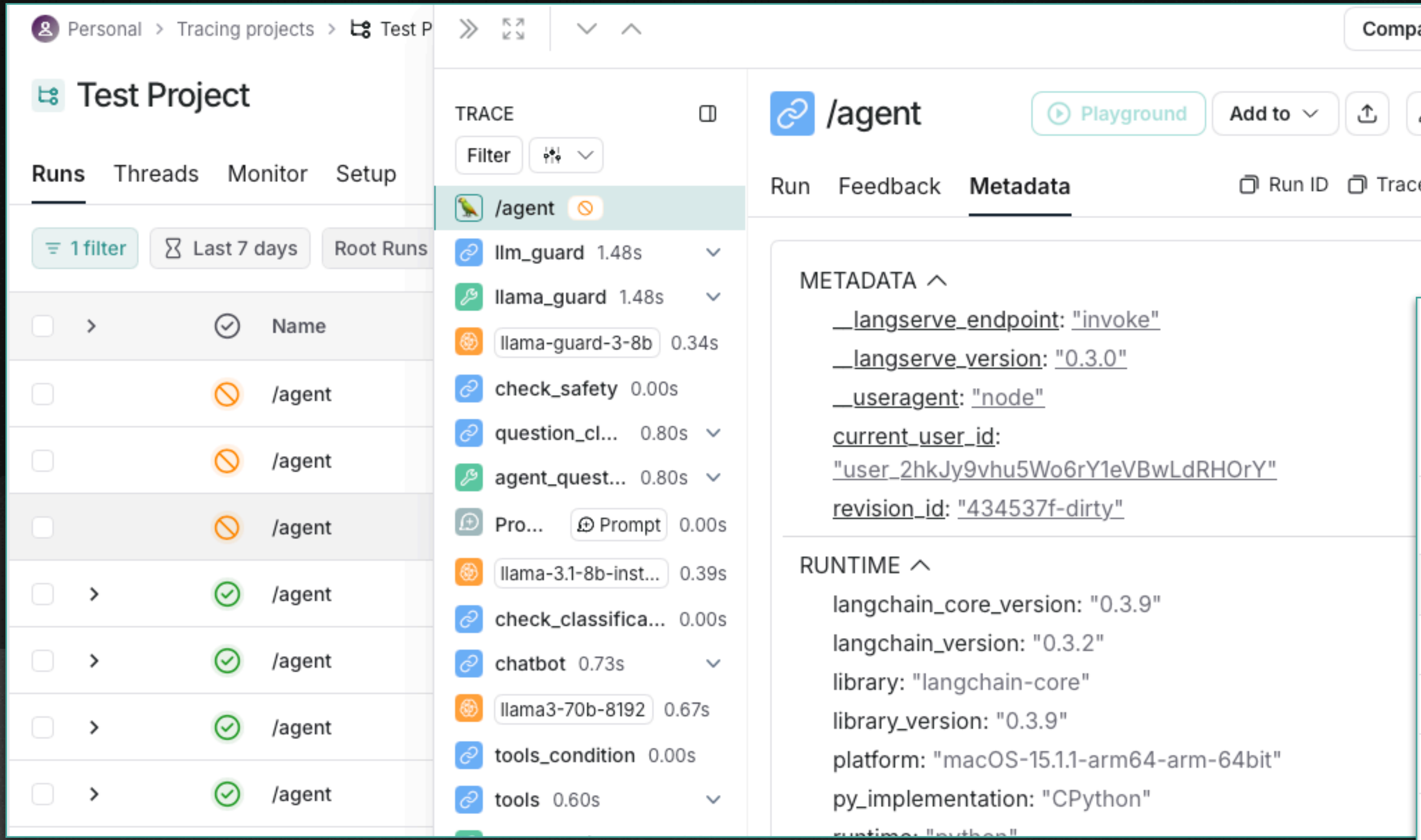
Take advantage of platforms to speed up agents



- Models are the biggest source of latency
  - Take seconds to run
- Platforms like LangGraph allow complex chaining of models and code
  - Concurrent calls
  - Branching

Agents are the future of LLM development

Get proper observability for your models



/agent
llm_guard 1.48s
llama_guard 1.48s
llama-guard-3-8b 0.34s
check_safety 0.00s
question_cl... 0.80s
agent_quest... 0.80s
Pro... Prompt 0.00s
llama-3.1-8b-inst... 0.39s
check_classifica... 0.00s
chatbot 0.73s
llama3-70b-8192 0.67s
tools_condition 0.00s
tools 0.60s

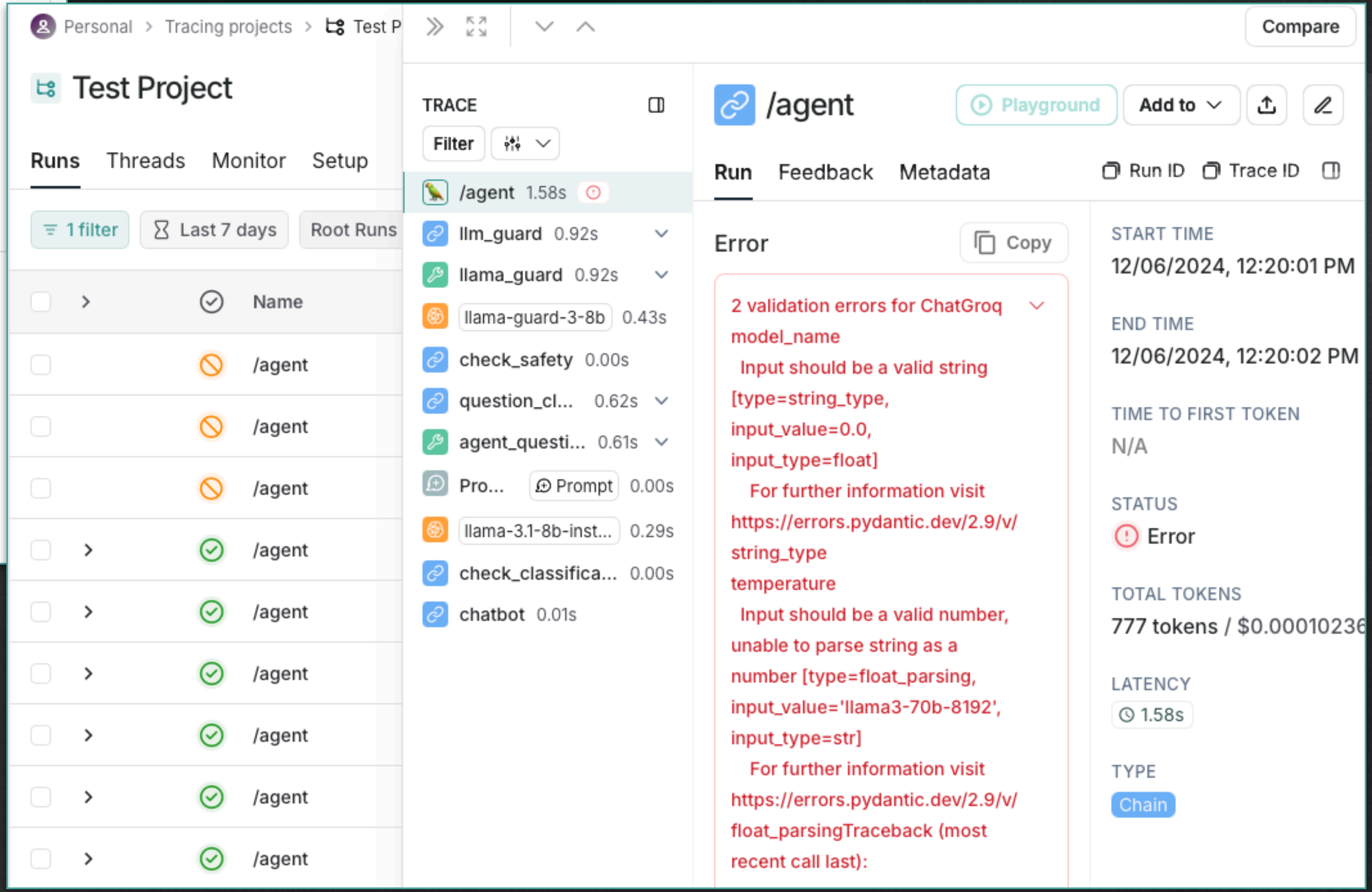
/agent

METADATA

```
__langserve_endpoint: "invoke"
__langserve_version: "0.3.0"
__useragent: "node"
current_user_id:
"user_2hkJy9vhu5Wo6rY1eVBwLdRHOrY"
revision_id: "434537f-dirty"
```

RUNTIME

```
langchain_core_version: "0.3.9"
langchain_version: "0.3.2"
library: "langchain-core"
library_version: "0.3.9"
platform: "macOS-15.1.1-arm64-arm-64bit"
py_implementation: "CPython"
```



/agent 1.58s
llm_guard 0.92s
llama_guard 0.92s
llama-guard-3-8b 0.43s
check_safety 0.00s
question_cl... 0.62s
agent_questi... 0.61s
Pro... Prompt 0.00s
llama-3.1-8b-inst... 0.29s
check_classifica... 0.00s
chatbot 0.01s

/agent

Error

```
2 validation errors for ChatGroq
model_name
  Input should be a valid string
[type=string_type,
input_value=0.0,
input_type=float]
    For further information visit
https://errors.pydantic.dev/2.9/v/
string_type
temperature
  Input should be a valid number,
unable to parse string as a
number [type=float_parsing,
input_value='llama3-70b-8192',
input_type=str]
    For further information visit
https://errors.pydantic.dev/2.9/v/
float_parsingTraceback (most
recent call last):
```

START TIME
12/06/2024, 12:20:01 PM

END TIME
12/06/2024, 12:20:02 PM

TIME TO FIRST TOKEN
N/A

STATUS
Error

TOTAL TOKENS
777 tokens / $0.00010236

LATENCY
1.58s

TYPE
Chain

Agent run traces in LangSmith

Tracing model output, specially with agents, is hard without the necessary tooling

Questions ?

Juan Peredo

linkedin.com/in/juanperedotech

Thanks to AI Camp for organizing this meetup and to IBM for hosting us