Expanding the Capabilities of LLMs Through Community Collaboration

ANTONIO VELASCO FERNÁNDEZ & JOSE PABLO CABEZA GARCÍA, ELASTACLOUD

THURSDAY, SEPTEMBER 21 • 11:55 - 12:35

PBLOCZ/LLM-COMMUNITY-TALK (GITHUB.COM)

Open Source Summit Europe 2023: Expanding the Capabilities of LLMs Throu... (sched.com)

Abstract

In this talk, we will explore how community projects are being used to enhance large language models (LLMs), such as the LLama Index, to overcome token limitations. We will discuss the development of LangChain, a community-driven ecosystem aimed at expanding LLM capabilities by providing tools like AutoGPT. By leveraging the power of community collaboration, we can overcome some of the challenges of working with LLMs, and unlock new possibilities for natural language processing. Attendees will gain a deeper understanding of the role that open source projects can play in expanding the capabilities of LLMs, and how they can contribute to these efforts.

To expand on the Subject

https://osseu2023.sched.com/overview/type/Open+AI+%26+Data+Forum?iframe=no

https://github.com/Hannibal046/Awesome-LLM

https://github.com/Mooler0410/LLMsPracticalGuide

Introduction

Jose Pablo Cabeza García — Lead Data Engineer at Elastacloud
Jose Pablo Cabeza García | LinkedIn
pblocz (Pablo Cabeza García) (github.com)

Antonio Velasco Fernández — Senior Data Scientist at Elastacloud
Antonio Velasco Fernández | LinkedIn
Antonio-Velasco (Antonio Velasco Fernández) (github.com)

Agenda

Introduction

LLMs development and overview

- Open Source LLMs
- Private LLMs

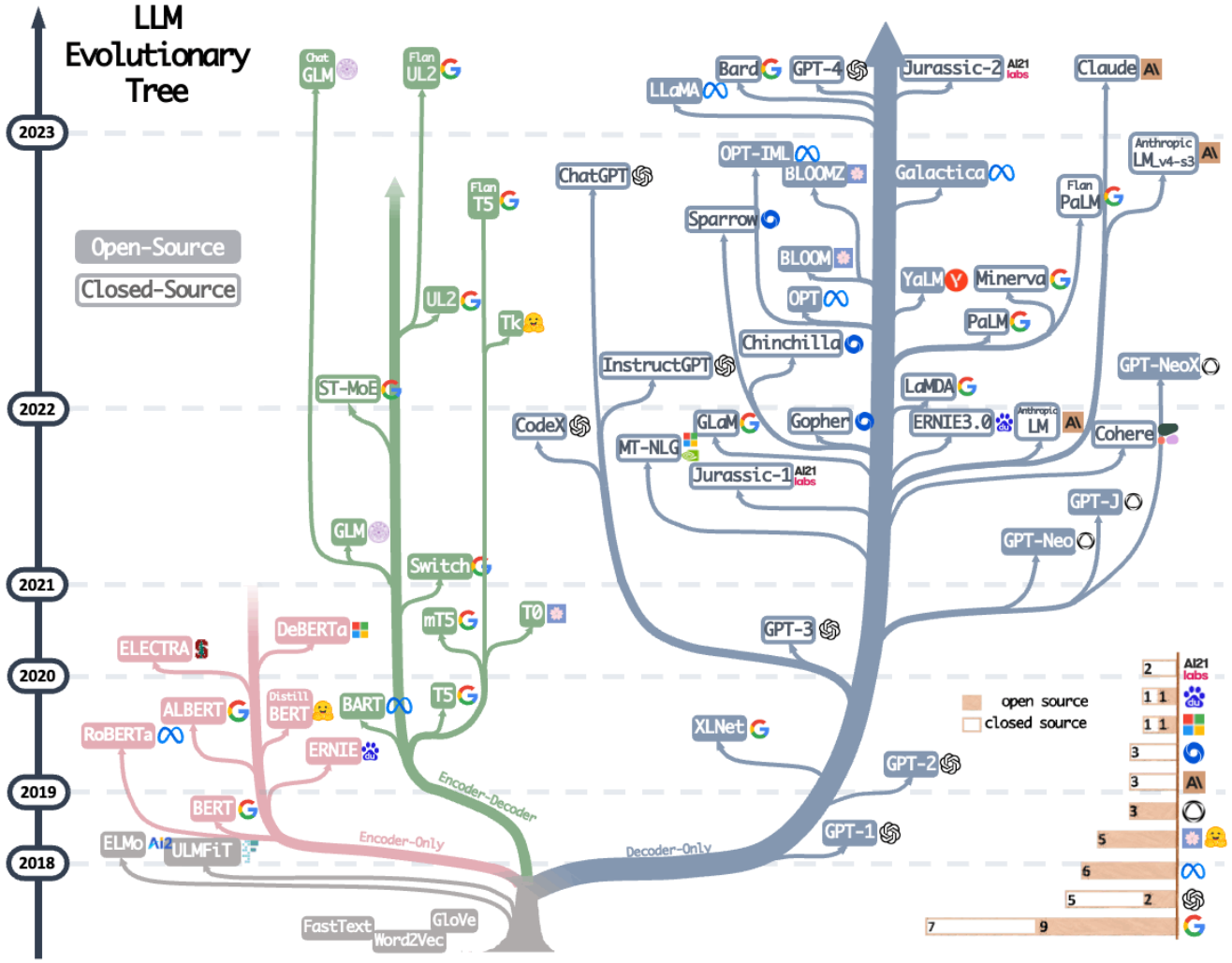
Private & Public development

Community efforts

Practical Examples

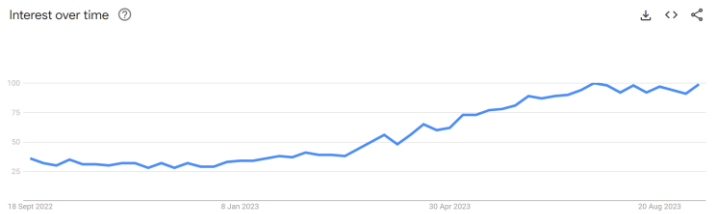
Evolution of LLMs

- Encoder only
- Encoder-Decoder
- Decoder only

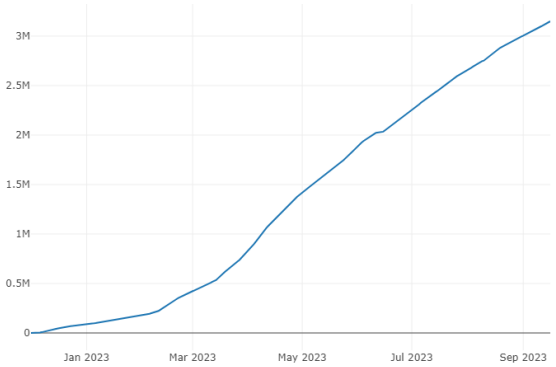


https://github.com/Mooler0410/LLMsPracticalGuide/

Google trends llama llm



Google trends LLM



r/chatgpt

The LLM hype

Foundational papers

Attention Is All You Need
2017
https://arxiv.org/abs/1706.03762

BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
2019
https://arxiv.org/abs/1810.04805

LLaMA: Open and Efficient Foundation Language Models
2023
https://arxiv.org/abs/2302.13971

Open Source LLMs Overview

Llama 2 by Meta: Groundbreaking LLM for text and code generation, supported on Azure and Windows, trained to minimize harmful outputs.

UL2 and Flan-UL2 by Google: Open-source LLMs based on T5 model, learn from unlabeled text, perform tasks like summarization and sentiment analysis.

BLOOM by BigScience: Massive LLM with 176 billion parameters, handles multiple languages and domains, based on GPT-3 architecture.

GPT-J-6B by EleutherAI: Open-source LLM similar to GPT-3, uses JAX, generates coherent text for various domains.

MPT-7B by MosaicML: New standard for open-source LLMs, performs well on multiple benchmarks and tasks, based on Megatron-LM architecture.

Private & Public development

BBC · 4mon

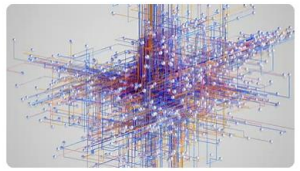
Sam Altman: CEO of OpenAI calls for US to regulate artificial intelligence

Sam Altman, the CEO of OpenAI, the company behind ChatGPT, testified before a US Senate committee on Tuesday about the possibilities - and pitfalls - of the new technology. In a matter of months ...

CNET on MSN · 10h

AI and You: Big Tech Says AI Regulation Needed, Microsoft Takes On Copyright Risks

"Regulate AI risk, not AI algorithms," IBM CEO Arvind Krishna said in a statement. "Not all uses of AI carry the same level ...



The Guardian · 3d

Tech leaders agree on AI regulation but divided on how in Washington forum

Bill Gates, Sundar Pichai, Sam Altman and others gathered for 'one of the most important conversations of the year' ...

Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs

Introducing MPT-7B, the first entry in our MosaicML Foundation Series. MPT-7B is a transformer trained from scratch on 1T tokens of text and code. It is open source, available for commercial use, and matches the quality of LLaMA-7B. MPT-7B was trained on the MosaicML platform in 9.5 days with zero human intervention at a cost of ~$200k.

alleywatch.com · 10d

Hugging Face Raises $235M for its Open Platform for AI and Machine Learning

Hugging Face is the most-used open platform for AI builders. More than 50,000 organizations are using it to build, train, and ...

Business Wire

LlamaIndex Raises $8.5M to Unlock Large Language Models Capabilities with Personal Data

SAN FRANCISCO--(BUSINESS WIRE)--LlamaIndex, the data framework for Large Language Models (LLMs), announced it raised $8.5M in seed funding...

Jun 6, 2023

LLM Ecosystem

Prompt Engineering

- Zero shot
- Few shot
- Chain of Thought
- Self-Consistency
- Tree of Thought

Vector Databases

elasticsearch

Pinecone

Weaviate

Chroma

Orchestration

Llama Index

Lang Chain

OpenAI

Bard

cohere

Hugging Face

LLM APIs

Community effort

LlamaIndex

- Offers data connectors to ingest your existing data sources and data formats (APIs, PDFs, docs, SQL, etc.)
- Provides ways to structure your data (indices, graphs) so that this data can be easily used with LLMs.
- Provides an advanced retrieval/query interface over your data: Feed in any LLM input prompt, get back retrieved context and knowledge-augmented output.
- Allows easy integrations with your outer application framework (e.g. with LangChain, Flask, Docker, ChatGPT, anything else).

Embeddings Vector Databases

Natural Language		**Embeddings**
Machine learning		(0.234, 0.655, 2.405)
Random forest		(1.112, 1.754, 3.193)
Mathematics	LLM	(0.987, 0.734, 0.112)
Neural Network		(3.401, 0.599, 2.345)
Back Propagation		(0.452, 2.465, 1.478)

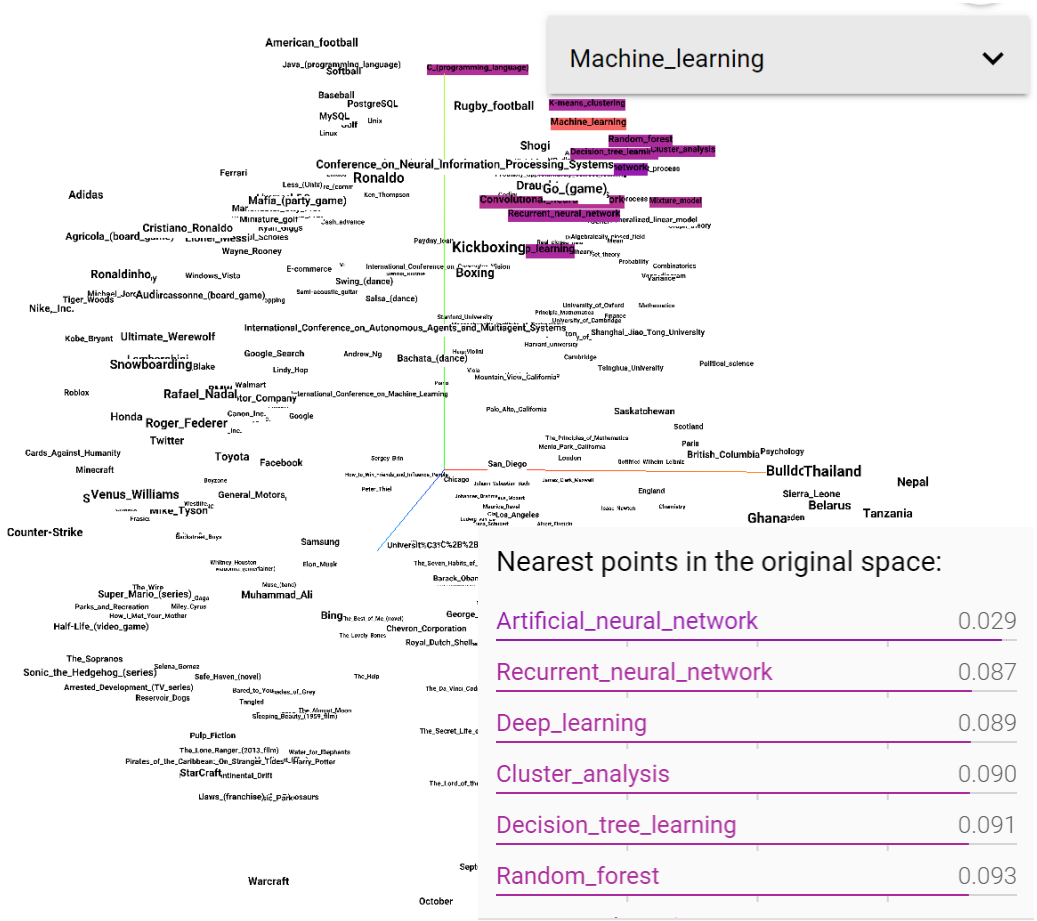
Example:

Corpus: [The, Black, Orange, Small, Cat]

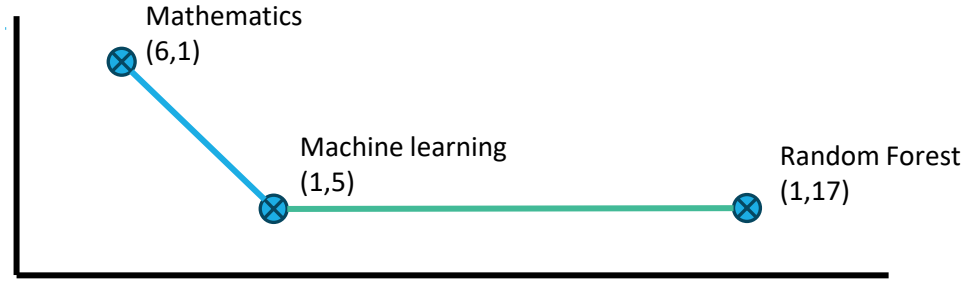
Sentence: The Black Cat
Embedding: [1,1,0,0,1]

Sentence: The Small Orange
Embedding: [1,0,1,1,0]

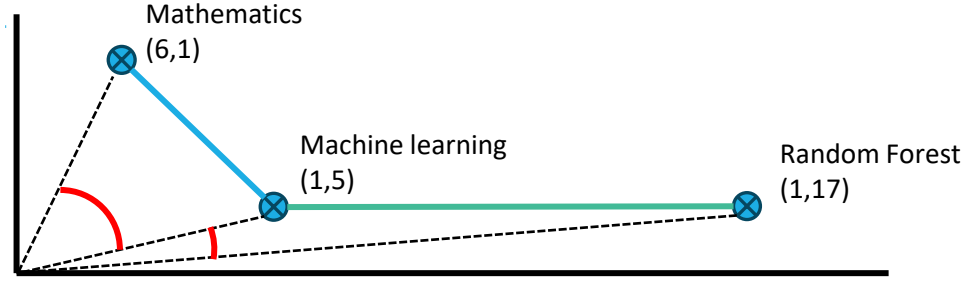
Embeddings Vector Databases: Searches



Euclidean Distance



Cosine Similarity



Community effort

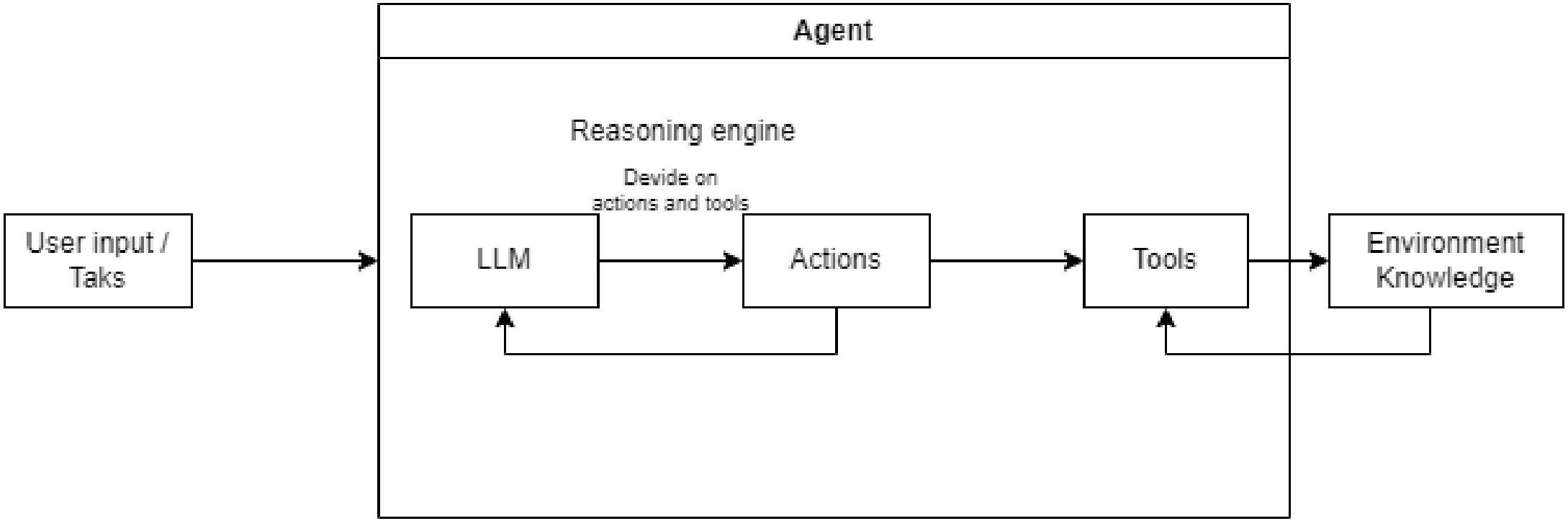
LangChain
Langstream
Haystack

Large language models (LLMs) are a powerful tool. However, they are limited in isolation. The real value comes when you combine them with other sources of computation or knowledge.

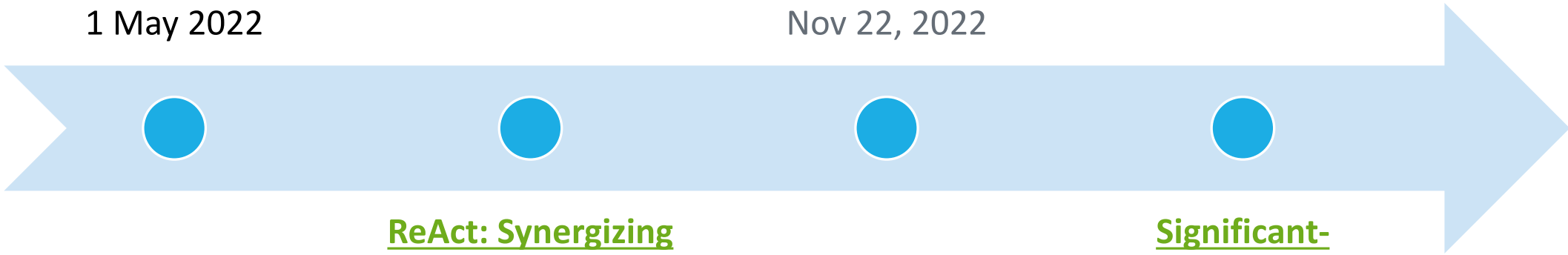
Virtually all industry and personal applications may benefit from the solutions shared by the community.

- ? Question Answering over specific documents
- Chatbots
- Agents
- Pipelines

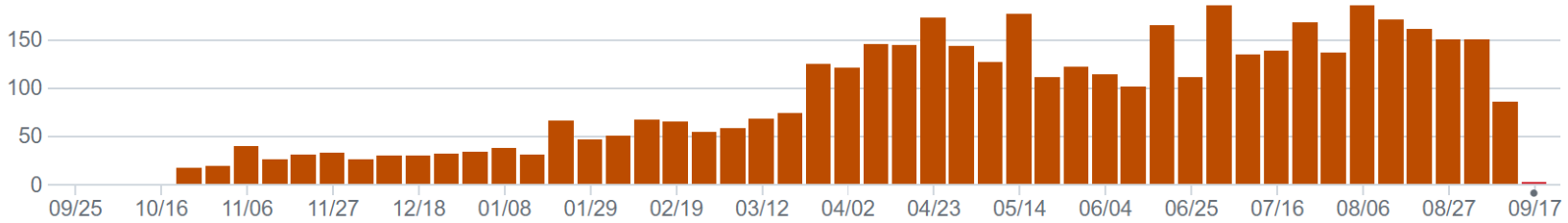
How do LLM agents work?



Agents timeline

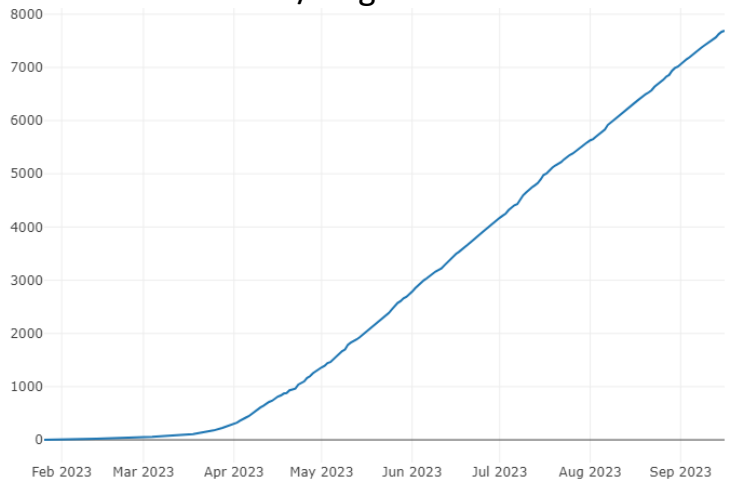


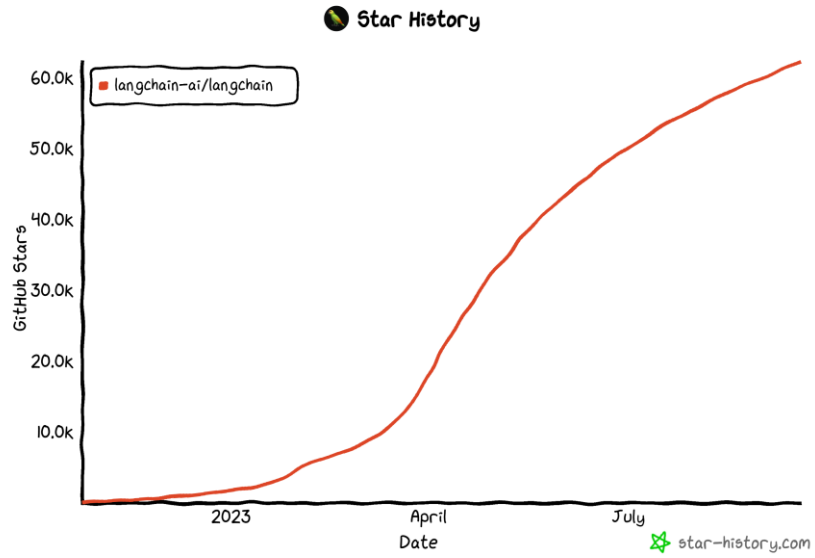
Evolution of Langchain



commits

r/langchain





Community effort

Skypilot

Framework for running LLMs and other AI or batch jobs. Allows management of different cloud providers, maximizing resource availability, cost savings and avoiding provider lock-in.

Supported providers:

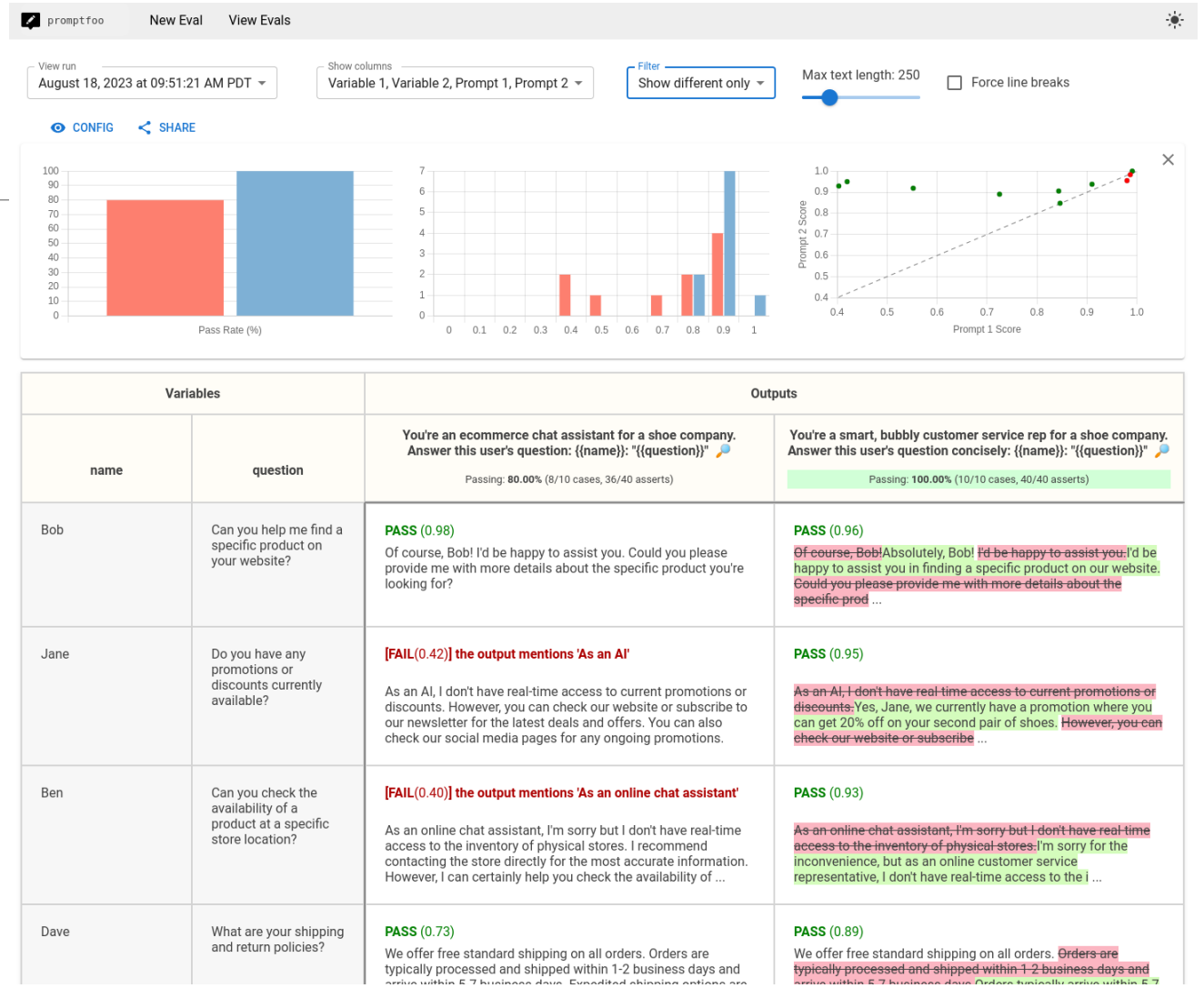
Community effort

Psychic	Data integration platform that allows the retrieval of context documents by providing standardized connection to most usual data sources. Notion, Slack, Zendesk, Confluence or Google Drive.
FastChat	Platform to train and evaluate most common large language model chatbots. It also provides with a web UI to consume them.
vLLM	Allows building an API server for a large language model. Which doubles as a toolset to build an offline batched inference on a dataset.

Community effort

Promptfoo

Promptify



Examples

Python Agent

Dataframe Agent

Github Agent

DnD context assistant

Llama Index Vector database

pblocz/llm-community-talk (github.com)

Referenced Repositories

1. lm-sys/FastChat: An open platform for training, serving, and evaluating large language models. Release repo for Vicuna and Chatbot Arena. (github.com)

2. skypilot-org/skypilot: SkyPilot: Run LLMs, AI, and Batch jobs on any cloud. Get maximum savings, highest GPU availability, and managed execution—all with a simple interface. (github.com)

3. vllm-project/vllm: A high-throughput and memory-efficient inference and serving engine for LLMs (github.com)

4. deepset-ai/haystack: :mag: LLM orchestration framework to build customizable, production-ready LLM applications. Connect components (models, vector DBs, file converters) to pipelines or agents that can interact with your data. With advanced retrieval methods, it's best suited for building RAG, question answering, semantic search or conversational agent chatbots. (github.com)

5. psychic-api/psychic: Data integration platform for LLMs. Connect to SaaS tools with turnkey auth and sync documents from N data sources with only one integration (github.com)

6. jerryjliu/llama_index: LlamaIndex (GPT Index) is a data framework for your LLM applications (github.com)

7. langchain-ai/langchain: ⚡ Building applications with LLMs through composability ⚡ (github.com)

8. rogeriochaves/langstream: Build robust LLM applications with true composability 🔗 (github.com)

9. promptfoo/promptfoo: Test your prompts. Evaluate and compare LLM outputs, catch regressions, and improve prompt quality. (github.com)

10. promptslab/Promptify: Prompt Engineering | Prompt Versioning | Use GPT or other prompt based models to get structured output. Join our discord for Prompt-Engineering, LLMs and other latest research (github.com)

Useful links

https://github.com/Hannibal046/Awesome-LLM

https://github.com/Mooler0410/LLMsPracticalGuide