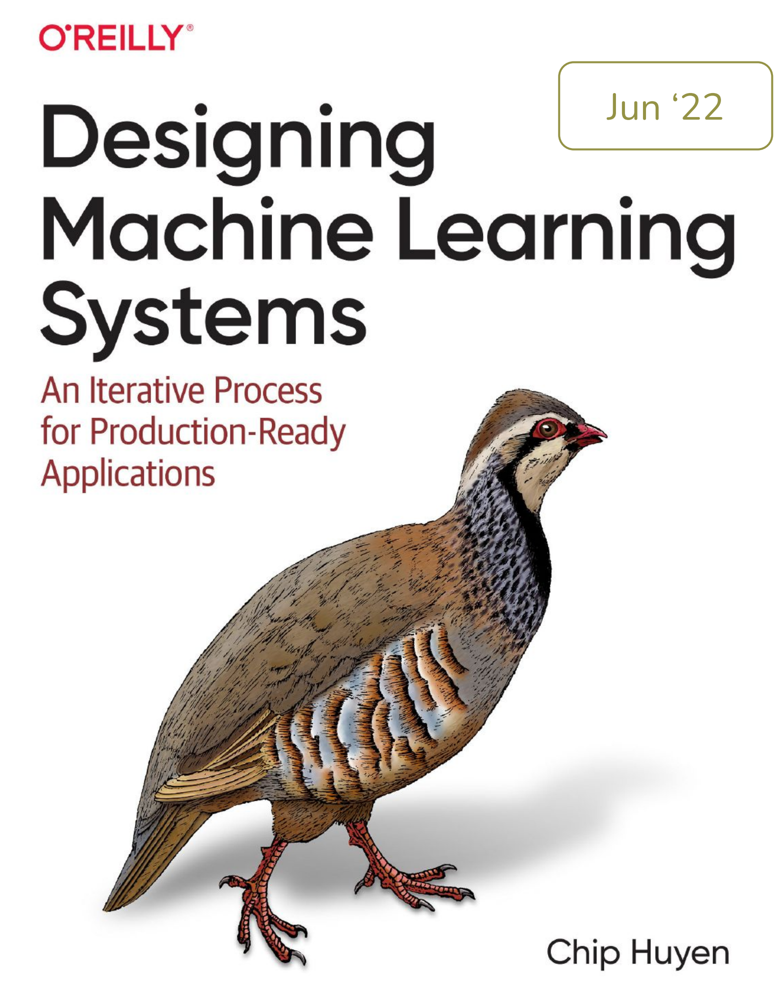
From ML Engineering to AI Engineering

Chip Huyen (@chipro | huyenchip.com)

Jun 2024



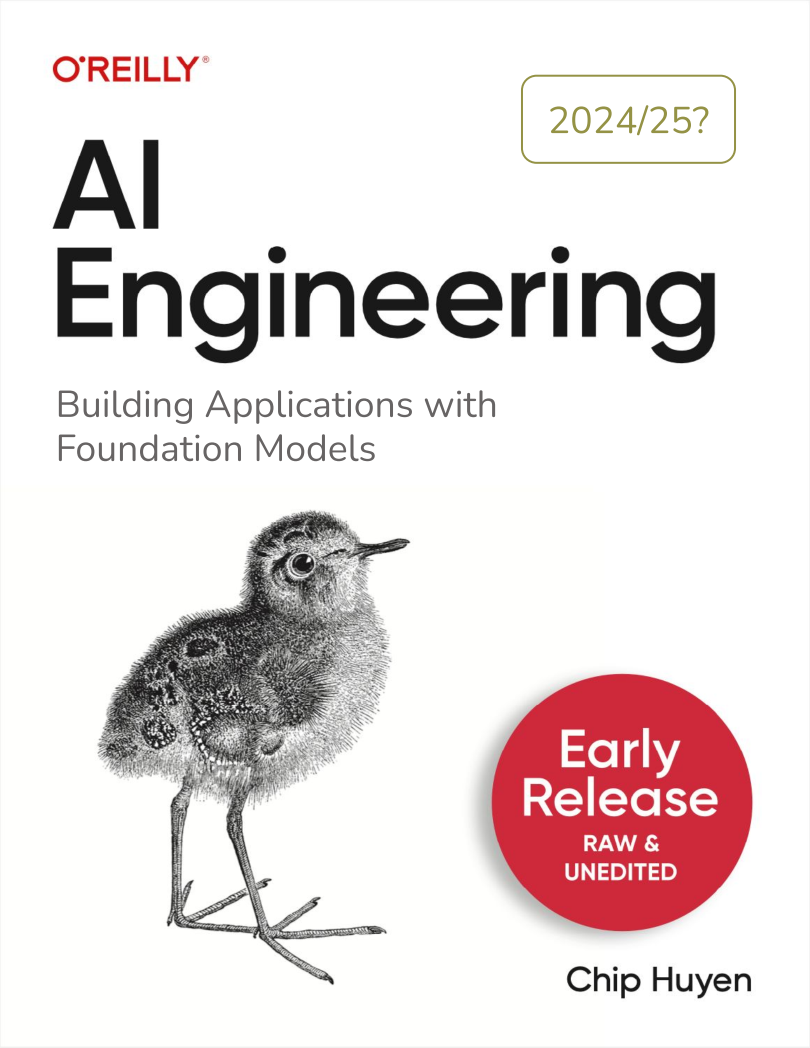
O'REILLY

Jun '22

Designing Machine Learning Systems

An Iterative Process for Production-Ready Applications

Chip Huyen



O'REILLY

2024/25?

AI Engineering

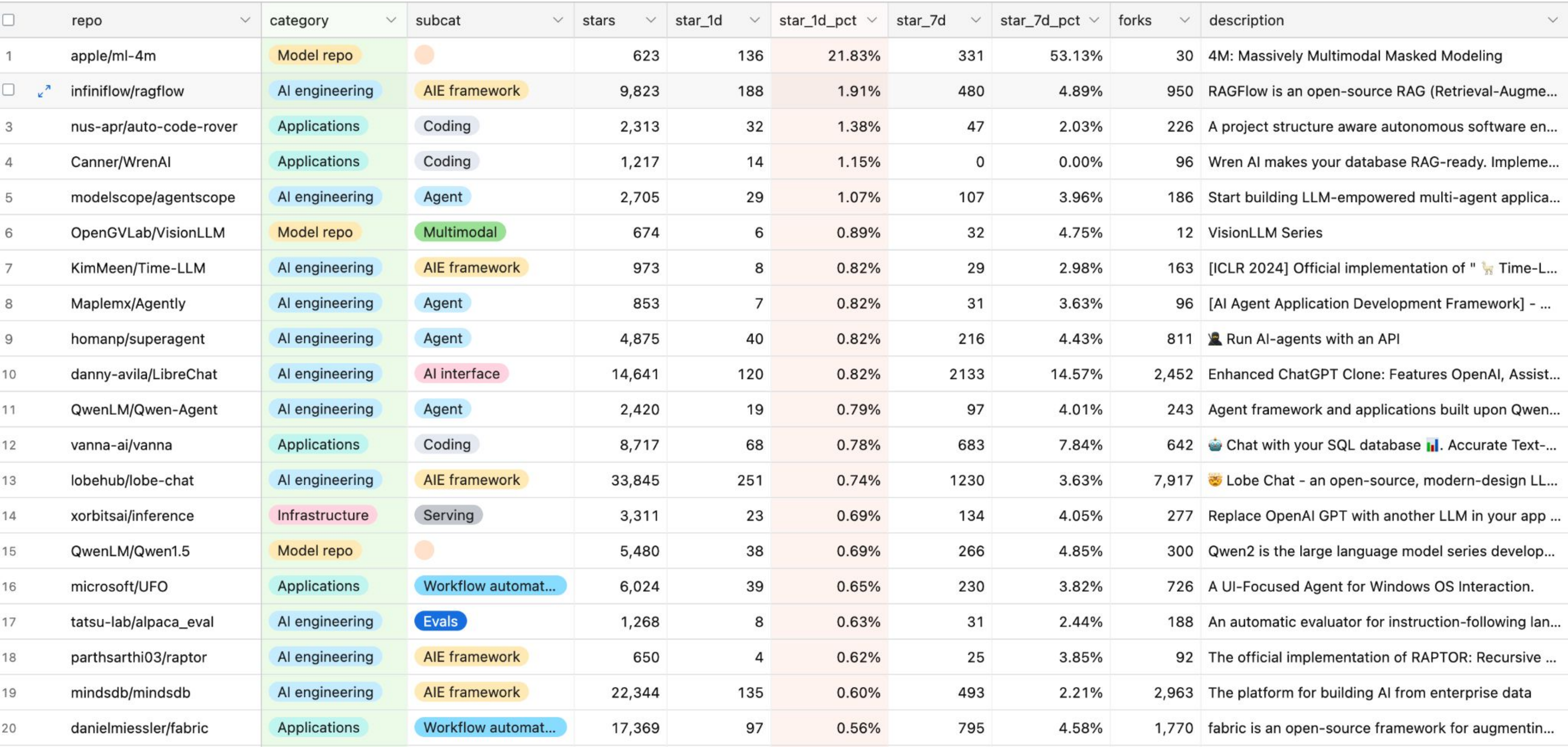
Building Applications with Foundation Models

Early Release
RAW & UNEDITED

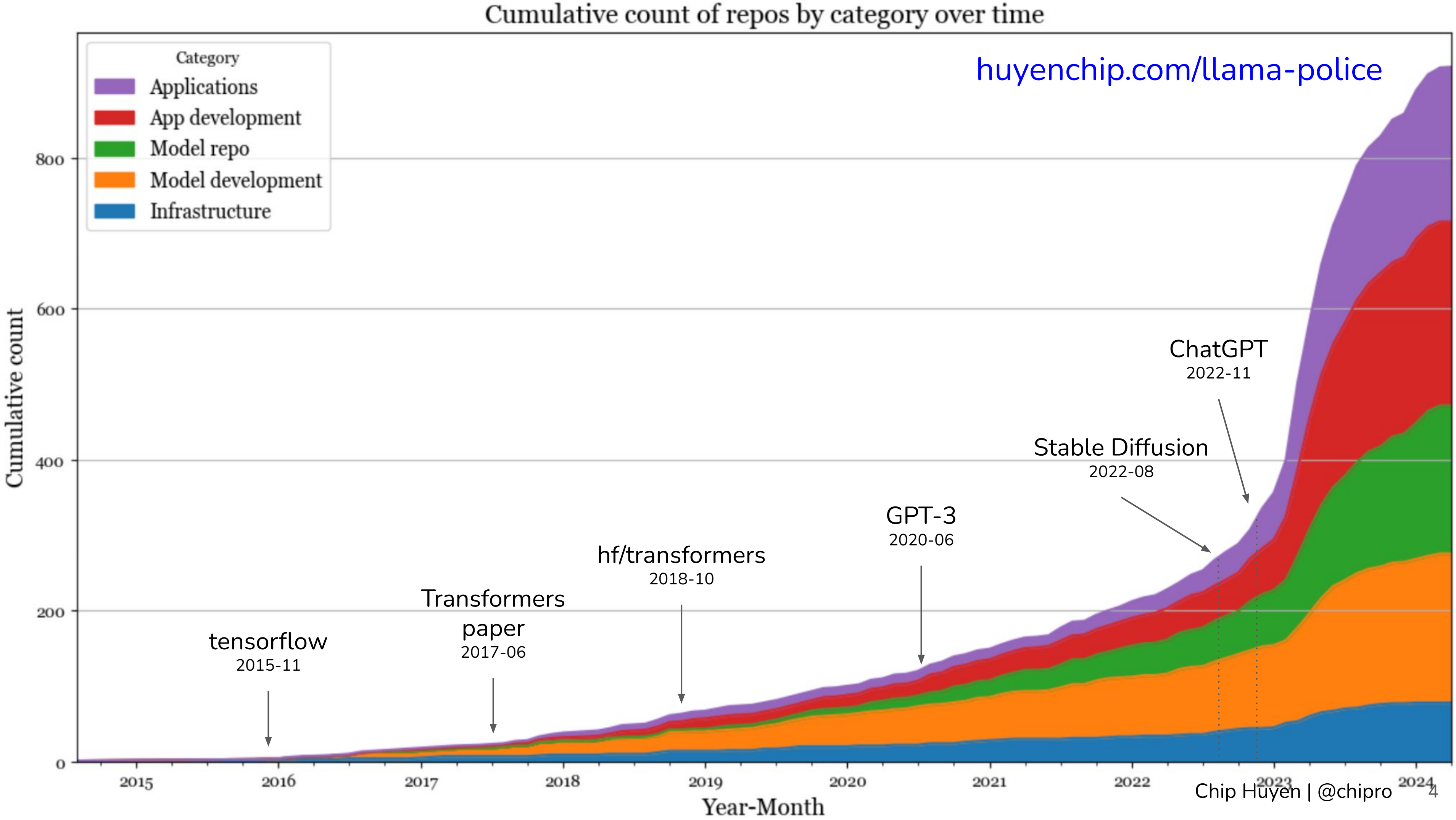
Chip Huyen

Track top 1100+ AI repos

huyenchip.com/llama-police



□	repo	category	subcat	stars	star_1d	star_1d_pct	star_7d	star_7d_pct	forks	description
1	apple/ml-4m	Model repo	●	623	136	21.83%	331	53.13%	30	4M: Massively Multimodal Masked Modeling
□ ⤢	infiniflow/ragflow	AI engineering	AIE framework	9,823	188	1.91%	480	4.89%	950	RAGFlow is an open-source RAG (Retrieval-Augme...
3	nus-apr/auto-code-rover	Applications	Coding	2,313	32	1.38%	47	2.03%	226	A project structure aware autonomous software en...
4	Canner/WrenAI	Applications	Coding	1,217	14	1.15%	0	0.00%	96	Wren AI makes your database RAG-ready. Impleme...
5	modelscope/agentscope	AI engineering	Agent	2,705	29	1.07%	107	3.96%	186	Start building LLM-empowered multi-agent applica...
6	OpenGVLab/VisionLLM	Model repo	Multimodal	674	6	0.89%	32	4.75%	12	VisionLLM Series
7	KimMeen/Time-LLM	AI engineering	AIE framework	973	8	0.82%	29	2.98%	163	[ICLR 2024] Official implementation of " 🦙 Time-L...
8	Maplemx/Agently	AI engineering	Agent	853	7	0.82%	31	3.63%	96	[AI Agent Application Development Framework] - ...
9	homanp/superagent	AI engineering	Agent	4,875	40	0.82%	216	4.43%	811	🥷 Run AI-agents with an API
10	danny-avila/LibreChat	AI engineering	AI interface	14,641	120	0.82%	2133	14.57%	2,452	Enhanced ChatGPT Clone: Features OpenAI, Assist...
11	QwenLM/Qwen-Agent	AI engineering	Agent	2,420	19	0.79%	97	4.01%	243	Agent framework and applications built upon Qwen...
12	vanna-ai/vanna	Applications	Coding	8,717	68	0.78%	683	7.84%	642	🤖 Chat with your SQL database 📊. Accurate Text-...
13	lobehub/lobe-chat	AI engineering	AIE framework	33,845	251	0.74%	1230	3.63%	7,917	🤯 Lobe Chat - an open-source, modern-design LL...
14	xorbitsai/inference	Infrastructure	Serving	3,311	23	0.69%	134	4.05%	277	Replace OpenAI GPT with another LLM in your app ...
15	QwenLM/Qwen1.5	Model repo	●	5,480	38	0.69%	266	4.85%	300	Qwen2 is the large language model series develop...
16	microsoft/UFO	Applications	Workflow automat...	6,024	39	0.65%	230	3.82%	726	A UI-Focused Agent for Windows OS Interaction.
17	tatsu-lab/alpaca_eval	AI engineering	Evals	1,268	8	0.63%	31	2.44%	188	An automatic evaluator for instruction-following lan...
18	parthsarthi03/raptor	AI engineering	AIE framework	650	4	0.62%	25	3.85%	92	The official implementation of RAPTOR: Recursive ...
19	mindsdb/mindsdb	AI engineering	AIE framework	22,344	135	0.60%	493	2.21%	2,963	The platform for building AI from enterprise data
20	danielmiessler/fabric	Applications	Workflow automat...	17,369	97	0.56%	795	4.58%	1,770	fabric is an open-source framework for augmentin...



huyenchip.com/llama-police

tensorflow
2015-11

Transformers paper
2017-06

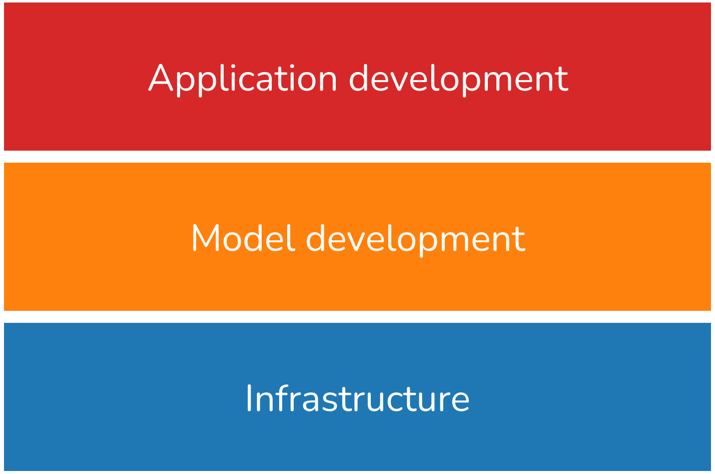
hf/transformers
2018-10

GPT-3
2020-06

Stable Diffusion
2022-08

ChatGPT
2022-11

The New AI Stack



Application development

Prompt engineering
AI interface
RAG
Evaluation

Model development

Dataset engineering
Inference optimization
Modeling & training

Infrastructure

Compute management
Serving
Gateway
Monitoring

The Rise of AI Engineering

AI engineering: the process of building applications with foundation models

1. Foundation models ⊃ LLMs
2. Foundation models often needed to adapted to specific needs

AI Engineering vs. ML Engineering

How ML production has changed with foundation models.

1. Model-as-a-service

   (+) Anyone, even those with no/minimal AI background, can now leverage AI to build applications

   (+) More consumer applications!!

   (+/-) Need UX to serve those applications

AI Engineering vs. ML Engineering

How ML production has changed with foundation models.

1. Model-as-a-service

   (+) Anyone, even those with no/minimal AI background, can now leverage AI to build applications

   (+) More consumer applications!!

   (+/-) Need UX to serve those applications

AI has become a common component in SWE, the way JavaScript and databases are

→ AI engineering is coming closer to full-stack

AI Engineering vs. ML Engineering

How ML production has changed with foundation models.

1. Model-as-a-service
2. Open-ended evaluation

   Open-ended responses

   (+) can be used for many applications

   (-) harder to evaluate

AI Engineering vs. ML Engineering

How ML production has changed with foundation models.

1. Model-as-a-service
2. Open-ended evaluation

   Open-ended responses

   (+) can be used for many applications

   (-) harder to evaluate

New eval approaches:

- Comparative evals (e.g. Chatbot Arena)
- AI-as-a-judge

AI Engineering vs. ML Engineering

How ML production has changed with foundation models.

1. Model-as-a-service
2. Open-ended evaluation

   Open-ended responses

   (+) can be used for many applications

   (-) harder to evaluate

⚠️ Evaluation is the biggest challenge for GenAI apps ⚠️

AI Engineering vs. ML Engineering

How ML production has changed with foundation models.

1. Model-as-a-service
2. Open-ended evaluation
3. Feature engineering -> context construction

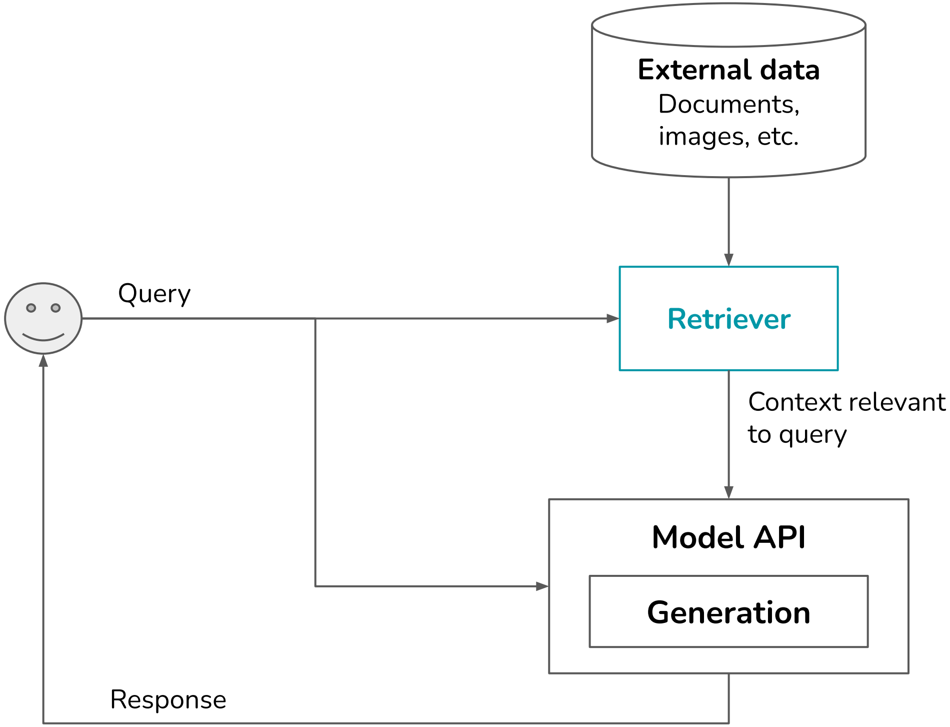
Model is more likely to hallucinate
when it doesn’t have the right info

AI Engineering vs. ML Engineering

How ML production has changed with foundation models.

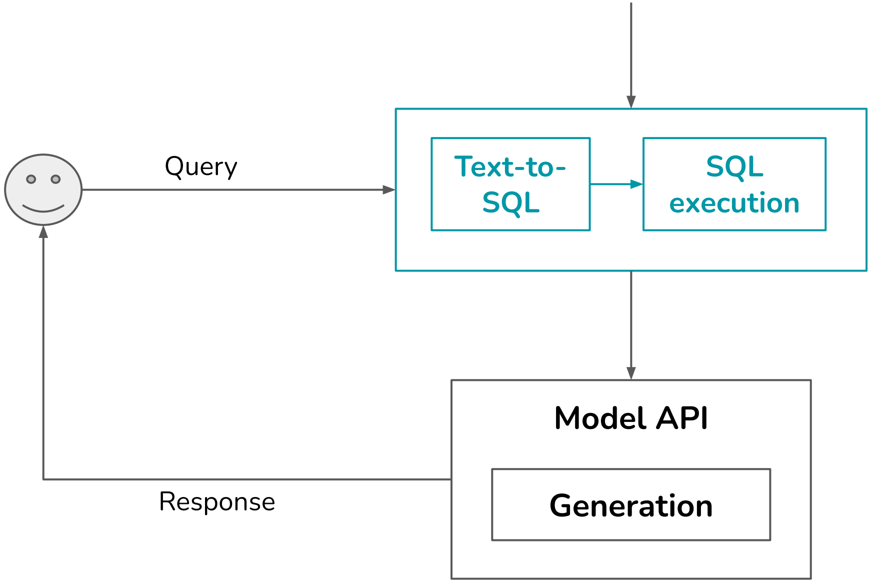
1. Model-as-a-service
2. Open-ended evaluation
3. Feature engineering -> context construction
   - Retrieval (RAG)
   - Tools that model can use to gather info (agentic)

Enhance context with retrieval



Retrieval with tabular data

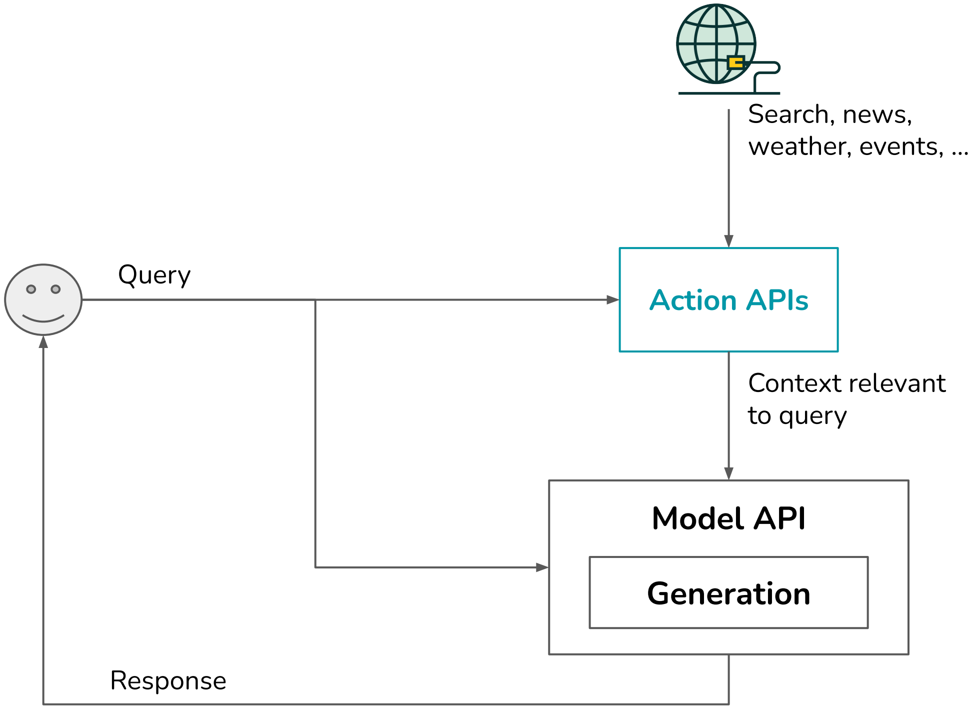
Product ID	Product name	Price/unit ($)	Units	Total
2044	Meow Mix Seasoning	10.99	1	10.99
3492	Purr & Shake	25	2	50
2045	Fruity Fedora	18	1	18
…	…	…	…	…

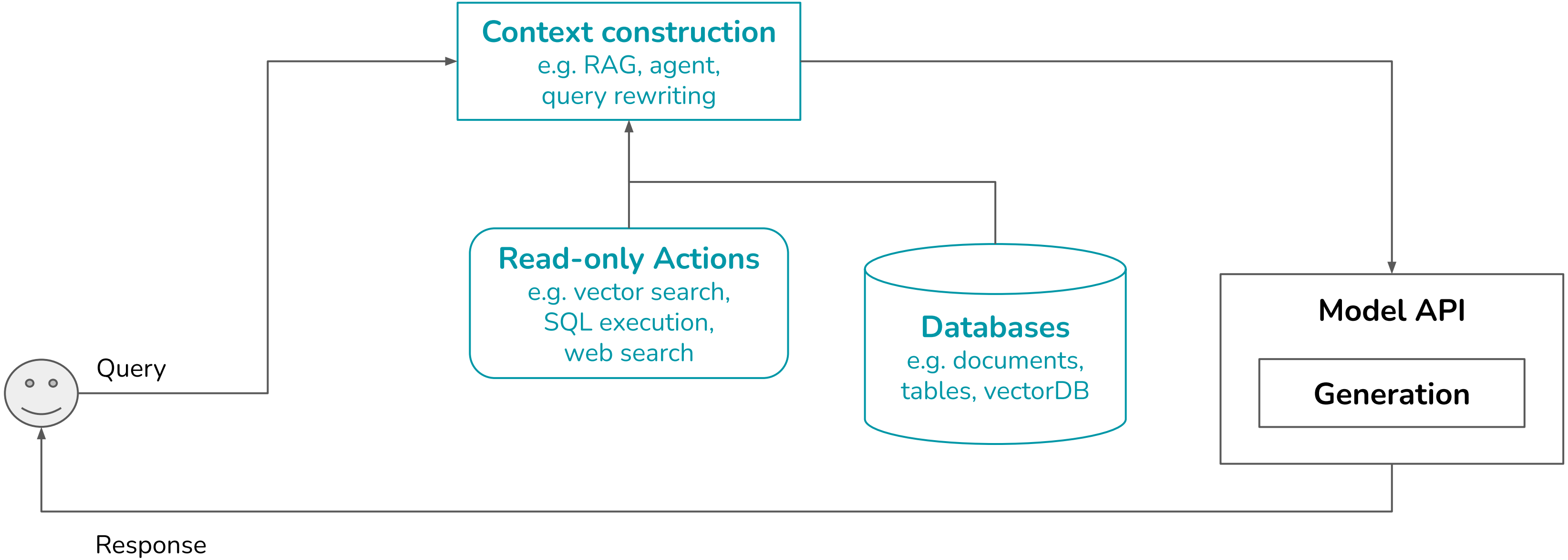


How many units of Fruity Fedora were sold in the last 7 days?

```
SELECT SUM(units) AS total_units_sold
FROM Sales
WHERE product_name = 'Fruity Fedora'
AND timestamp >= DATE_SUB(CURDATE(), INTERVAL 7 DAY);
```

Retrieval with tools




Context construction
e.g. RAG, agent,
query rewriting
Read-only Actions
e.g. vector search,
SQL execution,
web search
Databases
e.g. documents,
tables, vectorDB
Model API
Generation
Query
Response

AI Engineering vs. ML Engineering

1. Model-as-a-service
2. Open-ended evaluation
3. Feature engineering -> context construction
4. Bigger size

   (-) Higher latency

   (-) More expensive

   (-) Requiring more expertise to host

AI Engineering vs. ML Engineering

1. Model-as-a-service
2. Open-ended evaluation
3. Feature engineering -> context construction
4. Bigger size

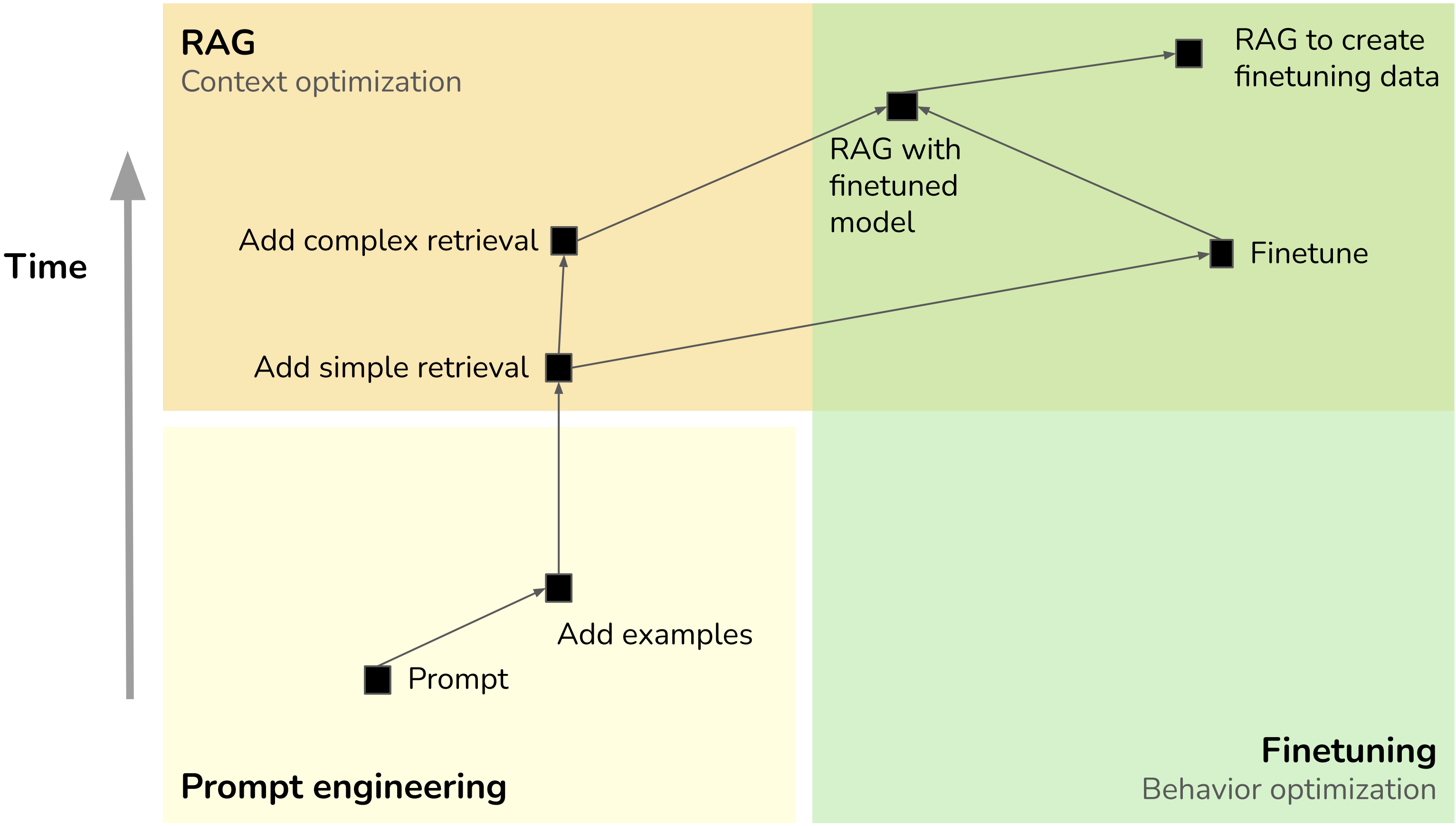
   (-) Higher latency

   (-) More expensive

   (-) Requiring more expertise to host

- Inference optimization+++

  Hardware, algorithm, model architecture
- Cache
- Parameter-efficient finetuning


RAG
Context optimization
RAG to create finetuning data
RAG with finetuned model
Time
Add complex retrieval
Finetune
Add simple retrieval
Add examples
Prompt
Prompt engineering
Finetuning
Behavior optimization

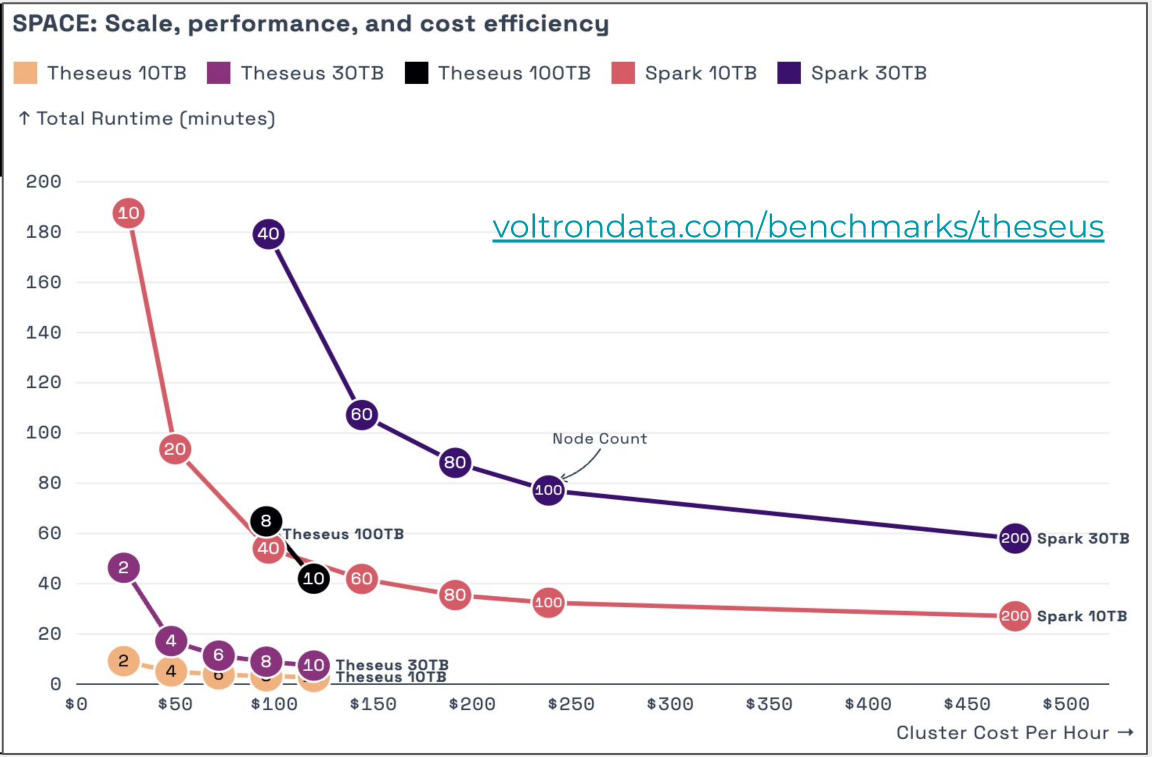
THESEUS: GPU-native query engine

TPC-H
(10TB, 30TB, 100TB)

- ✓ Up to 10 DGX Servers
- ✓ Parquet Files
- ✓ Remote File System
- ✓ Lots of Spilling

- ✘ No Sorting
- ✘ No Indexing
- ✘ No Caching
- ✘ No Warm Up (Cold Queries)

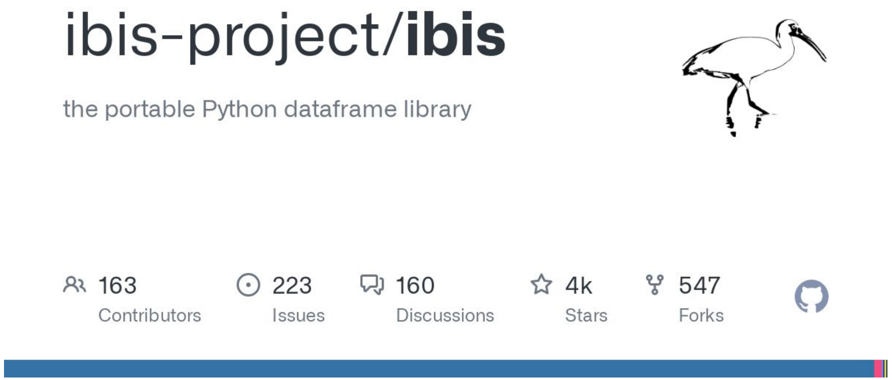
Note: Theseus: 1 Node 8 x A100 80 GB, Spark: 1 Node r5.8xlarge (AWS) 32 VCPU 32 GB



SPACE: Scale, performance, and cost efficiency

voltrondata.com/benchmarks/theseus

Open source



ibis-project/**ibis**

the portable Python dataframe library

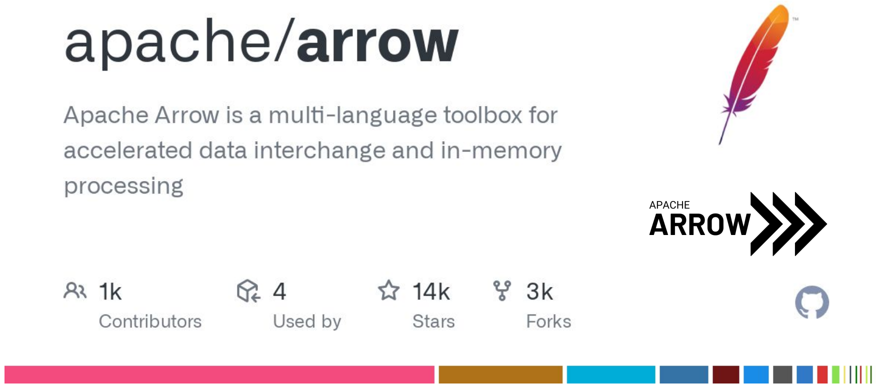
163 Contributors

223 Issues

160 Discussions

4k Stars

547 Forks



apache/**arrow**

Apache Arrow is a multi-language toolbox for accelerated data interchange and in-memory processing

1k Contributors

4 Used by

14k Stars

3k Forks

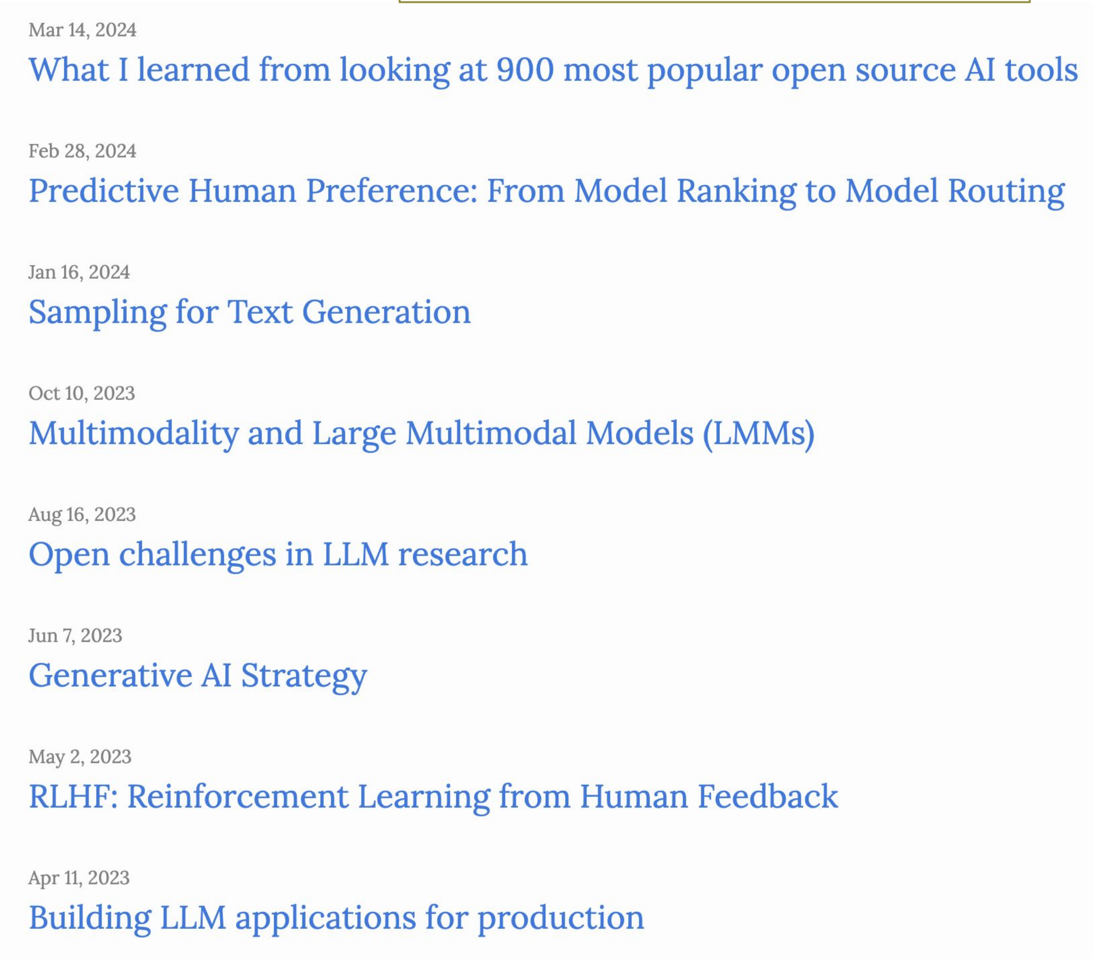
huyenchip.com/blog

Thank you!

@chipro

linkedin.com/in/chiphuyen

bit.ly/chip-mlops-discord



Mar 14, 2024

What I learned from looking at 900 most popular open source AI tools

Feb 28, 2024

Predictive Human Preference: From Model Ranking to Model Routing

Jan 16, 2024

Sampling for Text Generation

Oct 10, 2023

Multimodality and Large Multimodal Models (LMMs)

Aug 16, 2023

Open challenges in LLM research

Jun 7, 2023

Generative AI Strategy

May 2, 2023

RLHF: Reinforcement Learning from Human Feedback

Apr 11, 2023

Building LLM applications for production