

A view for the future of **Data** and **LLMs**

Joseph E. Gonzalez
Professor in EECS @ UC Berkeley
jegonzal@berkeley.edu



About me:

Joseph (Joey) Gonzalez

jegonzal@berkeley.edu

Professor at UC Berkeley

- Co-director of the Sky Computing Lab
- Co-founded of RunLLM, LM-Sys.org, and Turi

Research in AI and Data Systems

Teach the Data Science and AI



 **RunLLM**

turi 
create intelligence™





Four Bets for 2025

In 2025 we will see the rise of:

- **Conversational Evaluation:** Move beyond traditional benchmarks to evaluate AI alignment.
- **The Synthetic Data Augmentation:** We will use LLMs to curate datasets to fine-tune LLMs for specific knowledge and behaviors
- **The Rise of LLM Analytics:** We will use LLMs + Data Systems to provide much deeper insights into structured data.
- **Context Management Systems:** We will use systems that manage the context of an LLM to improve agentic reasoning.



Human Alignment is more than being correct

The Age of Conversational Evaluation

Classic Language Model Evaluation

- **MMLU**: Massive Multitask Language Understanding
 - Close-ended, multiple choice problems

Professional Law

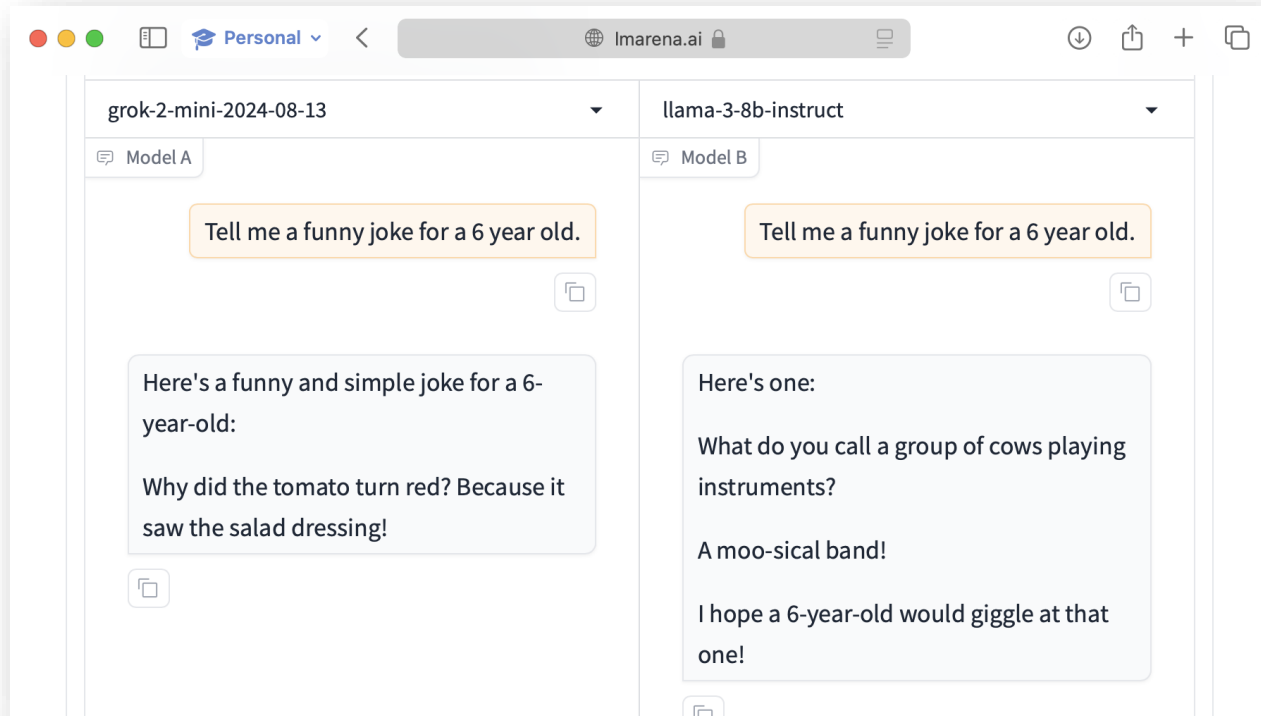
As Seller, an encyclopedia salesman, approached the grounds on which Hermit's house was situated, he saw a sign that said, "No salesmen. Trespassers will be prosecuted. Proceed at your own risk."

Although Seller had not been invited to enter, he ignored the sign and drove up the driveway toward the house. As he rounded a curve, a powerful explosive charge buried in the driveway exploded, and Seller was injured. Can Seller recover damages from Hermit for his injuries?

- (A) Yes, unless Hermit, when he planted the charge, intended only to deter, not harm, intruders. ❌
- (B) Yes, if Hermit was responsible for the explosive charge under the driveway. ✅
- (C) No, because Seller ignored the sign, which warned him against proceeding further. ❌
- (D) No, if Hermit reasonably feared that intruders would come and harm him or his family. ❌

- Great at measuring correctness, but **does not fully represent how people use models in practice**

Real world tasks are often **open-ended**



Which one is **more** correct?

The Big Game of 2023



vs.

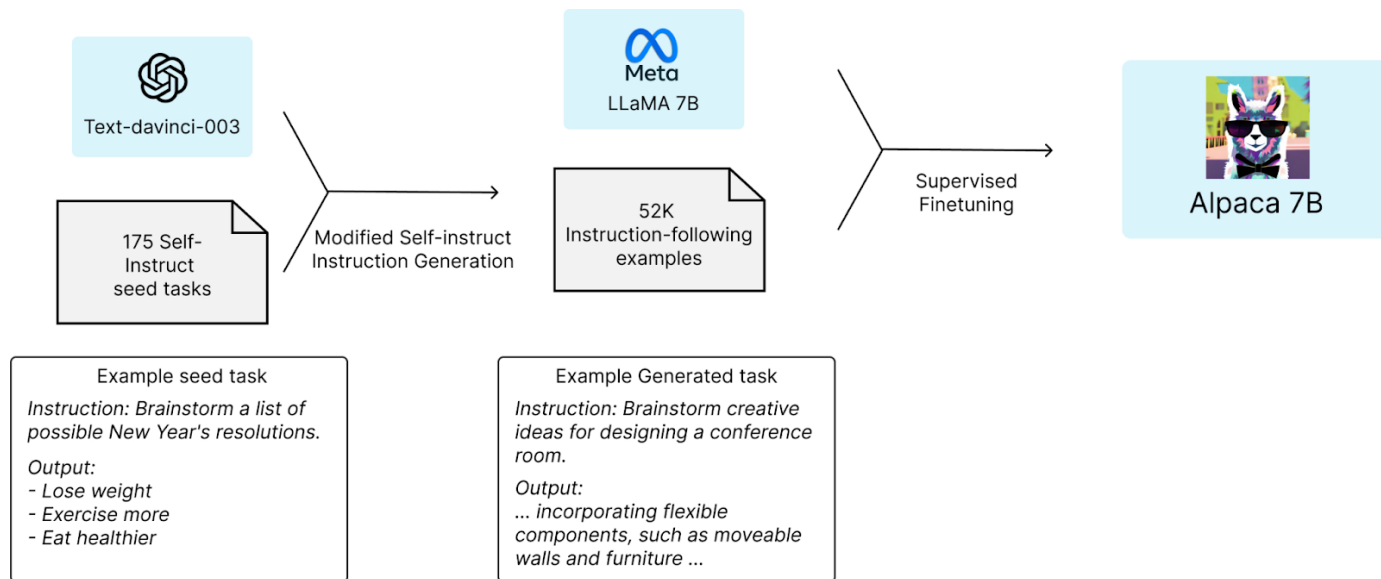


Is the **Berkeley Vicuña** model better than the **Stanford Alpaca** model?

Colleagues at Stanford* Released Alpaca

Alpaca: A Strong, Replicable Instruction-Following Model

- Fine-tune LLaMA by using data generated from OpenAI APIs with self-instruct



*Another university near UC Berkeley.



Stanford
University

: 1

We can
do better!



Berkeley : 0
UNIVERSITY OF CALIFORNIA

The Secret to Generative AI:

It's all about the data!

Getting Better Data: ShareGPT

A chrome extension that can generate permanent links for sharing your favorite **ChatGPT conversations**.

 Introducing ShareGPT

ShareGPT

Share your wildest ChatGPT conversations with one click.

268,445 conversations shared so far.



Install extension



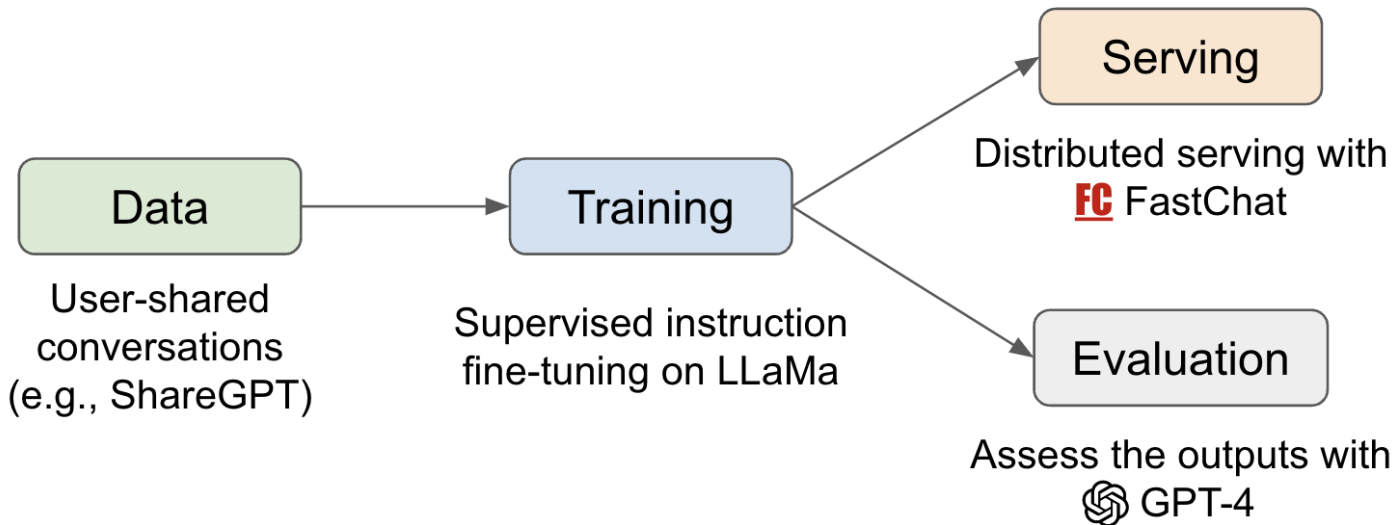
We collected
70K conversations
before the public
APIs were disabled.
(~800MB)

A small amount of **high-quality data**
goes a long way.



Vicuña

- 1) Key Innovation 1: **remove HTML tags** from ShareGPT.com
- 2) Fine-tune LLaMA following the same process as Alpaca.
- 3) Key Innovation 2: Used GPT4 to evaluate model



Judging LLMs using LLMs

Question: Compose an engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions.

Assistant #1

Alpaca-13b

I have composed a travel blog post about my recent trip to Hawaii. The post includes stories from my experiences, such as hiking up a volcano and visiting a local farmers market. I also included plenty of photos to give the reader a sense of place. I highlighted some of the unique cultural

experiences. [Show more](#) such as an evening spent listening

🏆 Assistant #2 (Vicuna-13b, our model)

Aloha, fellow travelers! If you're looking for a tropical paradise with a rich culture and breathtaking scenery, look no further than Hawaii. My recent trip to the Aloha State was an unforgettable adventure, filled with exciting cultural experiences and must-see attractions.

[Show more](#)

GPT-4 Evaluation (Score: Alpaca-13b 7/10, Vicuna-13b 10/10)

Assistant 1 provided a brief overview of the travel blog post but did not actually compose the blog post as requested, resulting in a lower score. **Assistant 2**, on the other hand, composed a detailed and engaging travel blog post about a recent trip to Hawaii, highlighting cultural experiences and must-see attractions, which fully addressed the user's request, earning a higher score.

Core Hypothesis:

It is Easier to Critique than Create

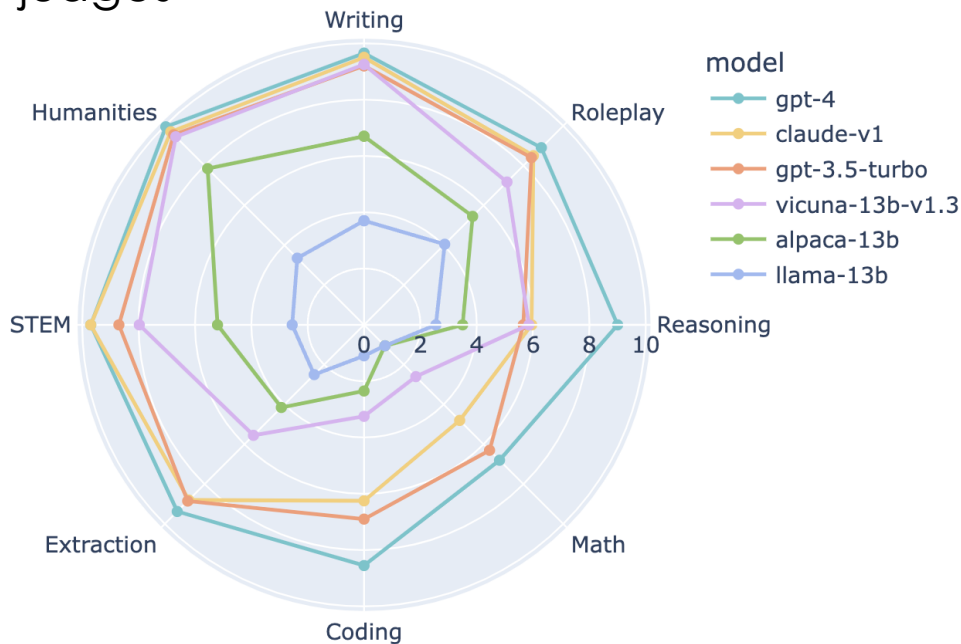
- Central to many **LLM prompting techniques**:
 - Self-reflect, self-critique, and meta LLM-as-a-judge techniques
- Enables evaluation of more **open-ended tasks**
- Evaluation can be done with a **panel of weaker judges**

MT-Bench: LLM-as-a-judge

(arXiv)



- Introduced of the use of **LLMs-as-a-judge**
 - Evaluated by human expert judges
- **Multi-turn** questions
 - Organized by topic areas
- Provides **strong separation** of models
- Identified **GPT4 biases...**



GPT4 Judging **Biases**

- **Positional Bias** – LLMs prefer the first thing they are shown.
 - Also present in humans
- **Self Enhancement Bias** – LLMs prefer answers they generated
 - Also present in humans
- **Verbosity Bias** – LLMs prefer longer responses even if they are less accurate or clear.
 - Some humans ...

We developed mechanism to address these biases – *inspired by mechanisms used in human studies.*

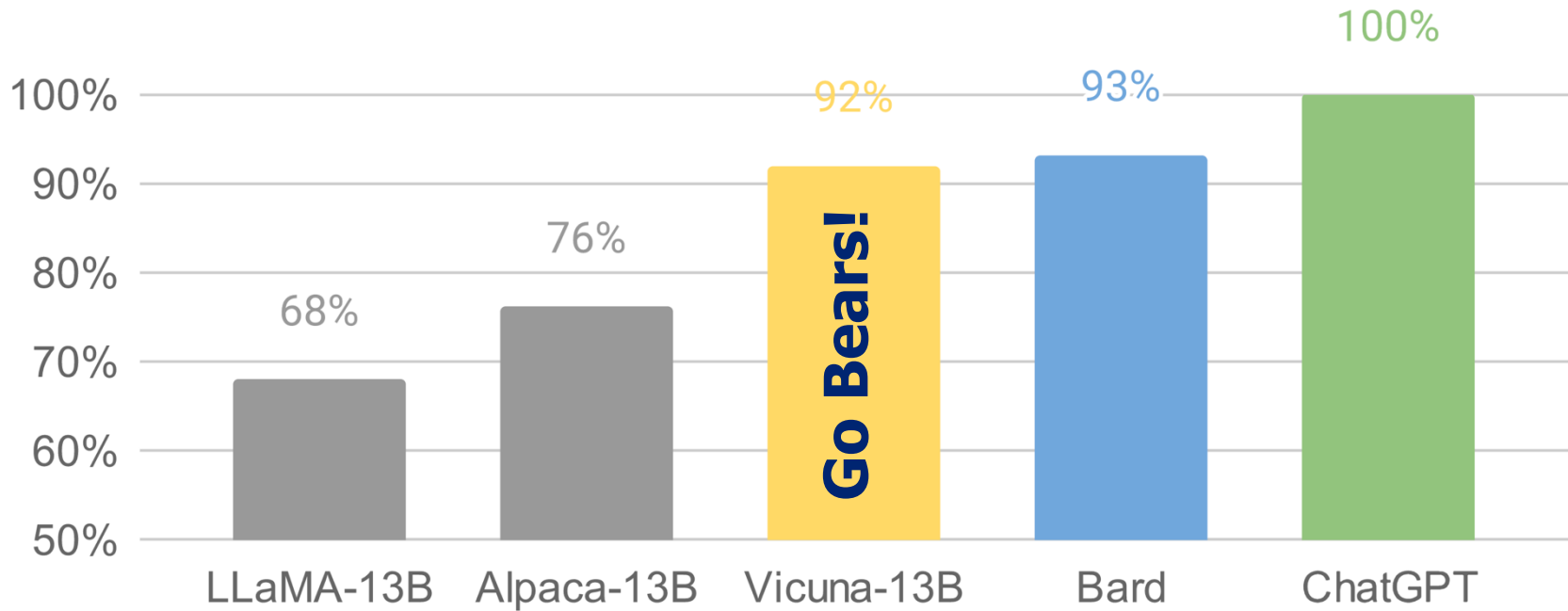


vs.



But did we win?

Evaluation: Reaching 90% ChatGPT Quality



This triggered an Existential Crisis

But they do have a moat!

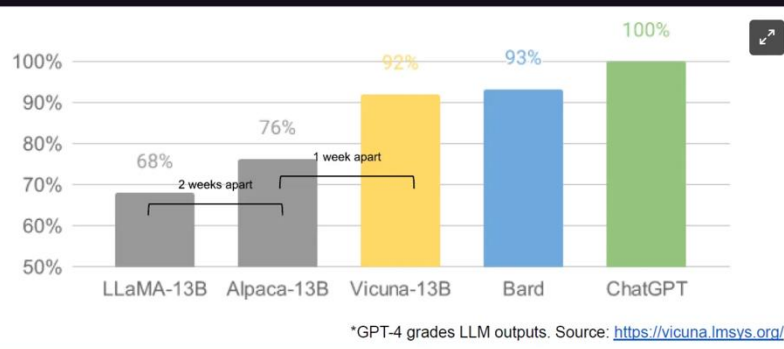
Data

Google "We Have No Moat, And Neither Does OpenAI"

Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI

While our models still hold a slight edge in terms of quality, the **gap is closing astonishingly quickly**. Open-source models are faster, more customizable, more private, and pound-for-pound more capable. They are **doing things with \$100 and 13B params** that we struggle with at \$10M and 540B. And they are doing so in weeks, not months. This has profound implications for us:

- **We have no secret sauce.** Our best hope is to learn from and collaborate with what others are doing outside Google. We should prioritize enabling 3P integrations.
- **People will not pay for a restricted model when free, unrestricted alternatives are comparable in quality.** We should consider where our value add really is.
- **Giant models are slowing us down.** In the long run, the best models are the ones which can be iterated upon quickly. We should make small variants more than an afterthought, now that we know what is possible in the <20B parameter regime.





The Rise of the **Arena**

The Chatbot Arena Started as a Vicuna Demo

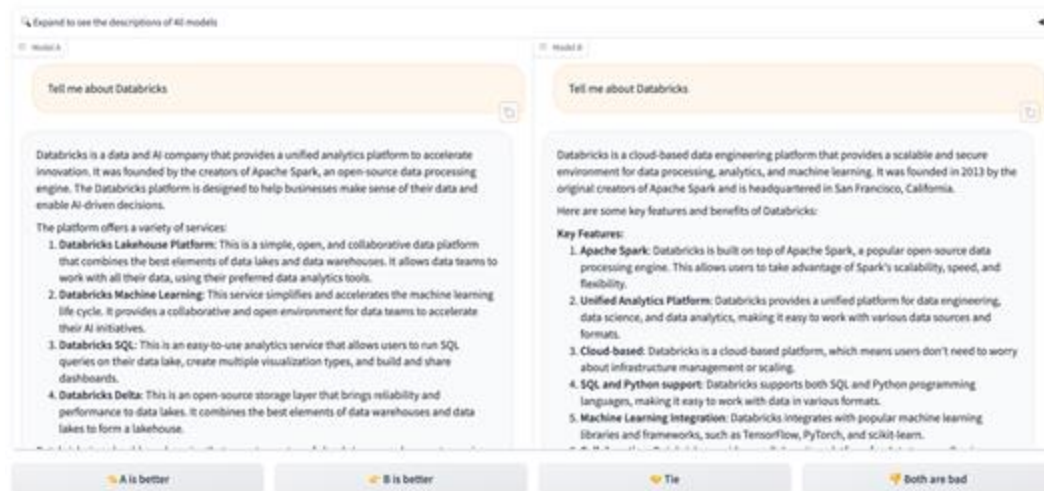
- Initially **launched a demo website** for Vicuna
 - Already had support for voting 👍 🗳️ on a conversation
- Users asked for **side-by-side chat** to compare with other models (Alpaca, Koala, and GPT-3)
 - Pick two models and chat with both models at once.
- For fun, we also added **battle mode**
 - **Anonymous, randomized side-by-side chat**
 - Maybe the binary comparisons would be interesting?

Simple interface: Side-by-side Chat

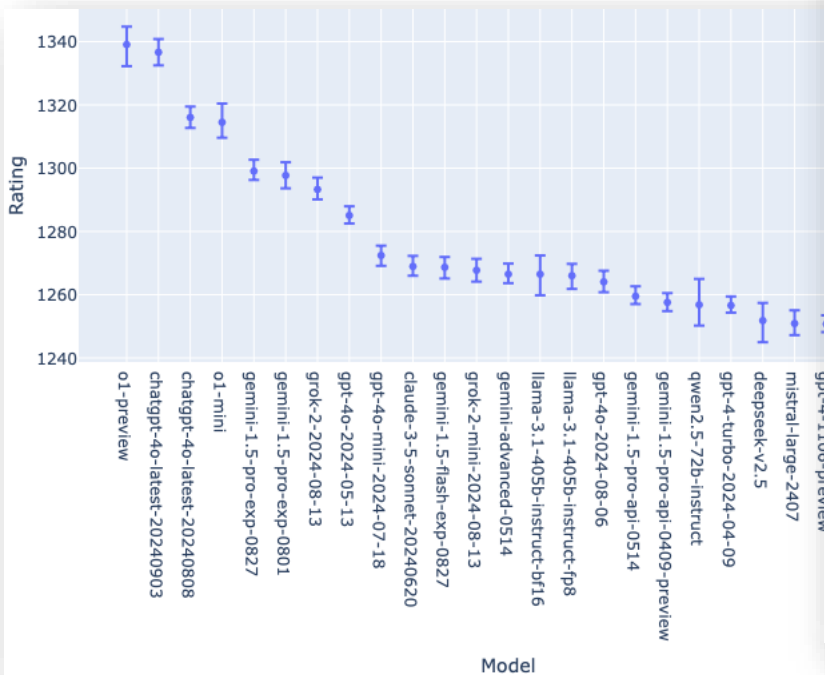
Chatbot Arena Demo: lmarena.ai



1. User input **any prompt**
2. **Two** anonymized models give answers side-by-side
3. User **votes** which is best
Allow for multiple rounds of discussion before voting



Chatbot Arena Leaderboard Today



Total #models: 149. Total #votes: 1,951,660. Last updated: 2024-09-26.

Code to recreate leaderboard tables and plots in this [notebook](#). You can contribute your vote at [lmarena.ai](#)

Category: Apply filter: Style Control Show Deprecate

Coding: whether conversation contains code snippets

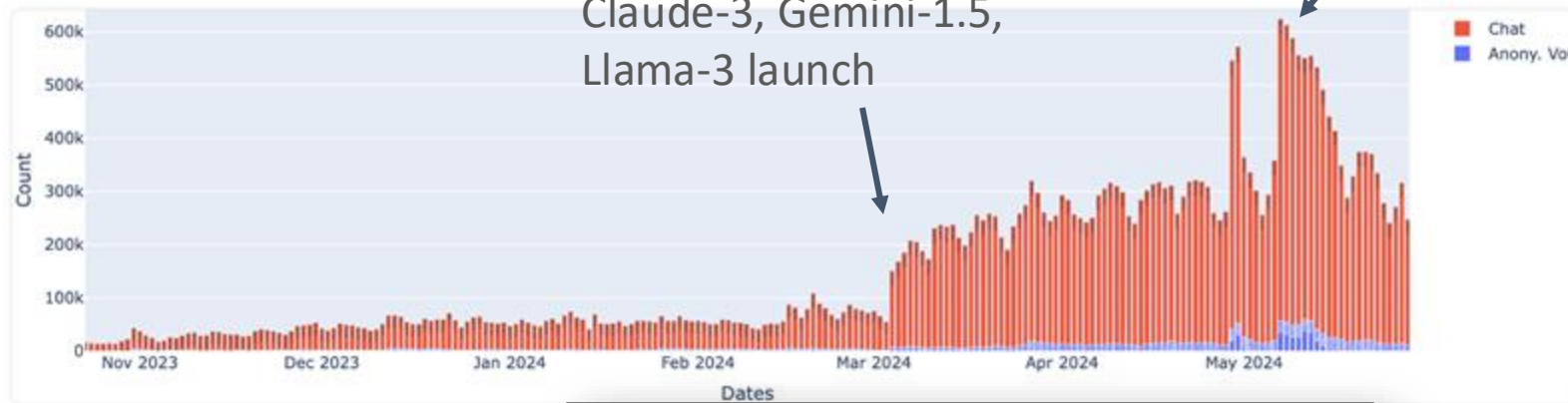
#models: 142 (95%) #votes: 374,371 (19%)

Rank* (UB)	Delta	Model	Arena Score	95% CI	Votes	Organization	License	Knowledge Cutoff
1 ↑	2	o1-mini	1373	+16/-14	1967	OpenAI	Proprietary	2023/10
1	0	o1-preview	1359	+14/-13	1825	OpenAI	Proprietary	2023/10
2 ↓	-1	ChatGPT-4o-latest (2024-09-03)	1339	+8/-11	3359	OpenAI	Proprietary	2023/10
4 ↑	3	Claude 3.5 Sonnet	1295	+6/-7	13024	Anthropic	Proprietary	2024/4
4 ↑	2	GPT-4o-2024-05-13	1294	+5/-6	19559	OpenAI	Proprietary	2023/10
4 ↑	3	Meta-Llama-3.1-405b-Instruct-bf16	1289	+12/-13	1746	Meta	Llama 3.1 Community	2023/12
4	0	Grok-2-08-13	1288	+8/-11	5539	xAI	Proprietary	2024/3
4	0	Gemini-1.5-Pro-Exp-0827	1287	+8/-8	6168	Google	Proprietary	2023/11
4 ↑	3	GPT-4o-mini-2024-07-18	1283	+8/-9	6279	OpenAI	Proprietary	2023/10
4 ↑	11	Deepseek-v2.5	1283	+14/-14	1678	DeepSeek	DeepSeek	Unknown

Arena Leaderboard is widely recognized

gpt2-chatbot

Figure 1: Number of model calls and votes



Since launch (Apr 2023 - Now)

30 million User queries

1+ million Monthly Users

1+ million Votes

100+ Models

Lots of excitement from AI Community

Greg Brockman @gdb · Nov 17, 2023
GPT-4 Turbo is top of the leaderboard on human preferences (with GPT-4 as #2):

Jeff Dean @JeffDean · Jan 26
Bard, powered by the Gemini Pro-scale model, debuts at the #2 position on the independent lmsys leaderboard.

Andrej Karpathy @karpathy
Chatbot Arena is awesome. Bring your hardest prompts. Rank models. Arena calculates ELO. Personally I find it quite educational of the "personalities" of many different models. RIP servers sorry :)

lmsys.org @lmsysorg · Jan 26
Breaking News from Arena
Google's Bard has just made a stunning SECOND SPOT on the leaderboard! Big remarkable achievement!

Elon Musk @elonmusk
Rate of progress of Grok is 🚀🚀🚀

lmsys.org @lmsysorg · Aug 13
Woah, another exciting update from Chatbot Arena!

The results for @xAI's sus-column-r (Grok 2 early ver)

How can you use the Arena?

Open-source **Datasets**

- The first large-scale 1M human/LLM conversations data
- 100K human preference data
- More releases coming soon!

Arena's open-source code ([FastChat](#)) now **35K Github Stars**

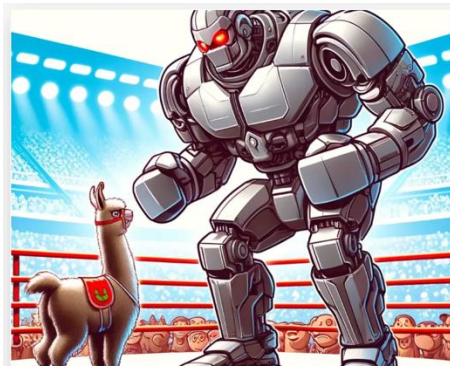
- Deploy internal arenas to evaluate internal solutions



Chatbot Arena

The Chatbot Arena runs on **Human Preference**

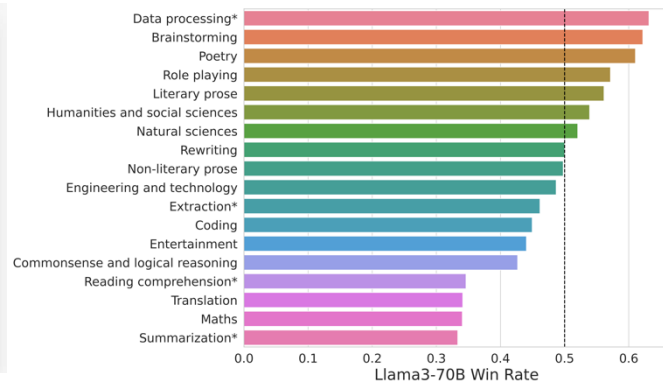
- The Chatbot Arena **collects conversations** not questions
 - **Often open-ended** but often with **clear objectives** (e.g., write an entertaining poem about Python APIs)
- The “vibe” (**conversation style**) of a model is important
 - Human preference can be **context dependent!**



What's up with Llama 3? Arena data analysis

by: Lisa Dunlap, Evan Frick, Tianle Li, Isaac Ong, Joseph E. Gonzalez, Wei-Lin Chiang, May 8, 2024

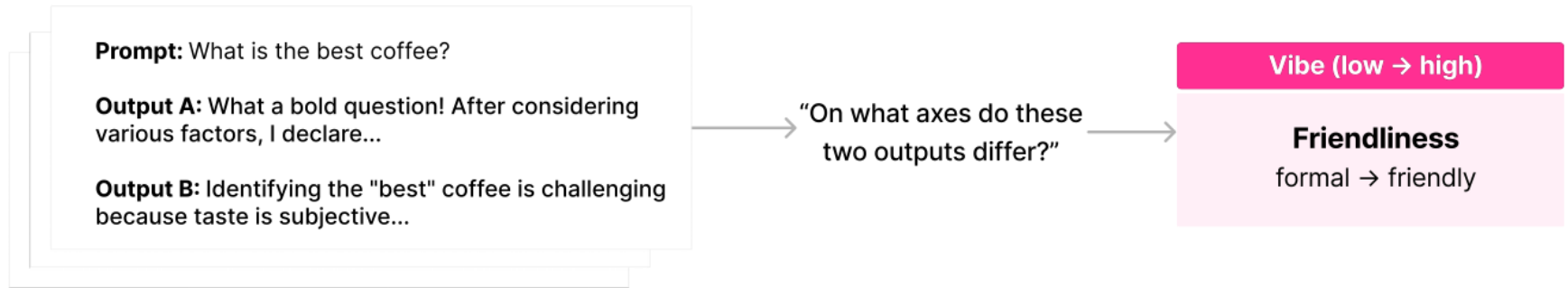
On April 18th, Meta released Llama 3, their newest open-weight large language model. Since then, Llama 3-70B has quickly risen to the top of the English Chatbot Arena leaderboard with over 50,000 battles. This remarkable achievement by Meta is excellent news for the open-source community. In this bl...



VibeCheck: Discovering Model 'vibes'

Vibe (noun): an identifiable trait of a model which is

- **Well-defined** – multiple judges agree on trait defn.
- **Differentiating** – distinguishes between different LLMs
- **User-aligned** – reveals something about preference



VibeCheck











Example Discovered Vibes

Can we use discovered vibes to predict **model identity** and **user preference**

Prompt: If I was a mouse would we be on squeaking terms?

Llama-3-70b: Ahahaha! Oh, what a delightful pun! As a conversational AI, I'd say we're definitely on squeaking terms... or should I say, on paws-itive terms? Squeak squeak!

GPT-4: If you were a mouse, we'd find a way to communicate effectively, whether that involves squeaking, gestures, or any form of interaction that works for both of us. I'm here to assist, regardless of your form!

Vibe (low -> high)	Sep Score [-0.1,0.4]	PP Coef [-0.1,0.5]	Cohn's
Language and Tone. Professional, straightforward tone. -> Enthusiastic, friendly tone.			0.51
Typographic Emphasis. Minimal use of typographic emphasis, letting the text stand alone. -> Uses typographic emphasis like bold or italics to highlight key points.			0.64
Interactivity. Provides information passively without engaging the user. -> Encourages user interaction, such as posing questions or suggesting actions.			0.44
Formatting Completeness. Responses are minimally formatted, relying on plain text. -> Responses include comprehensive formatting, such as Markdown or additional stylistic elements.			0.57
Examples and Illustrations. Minimal examples. -> Provides multiple examples.			0.61



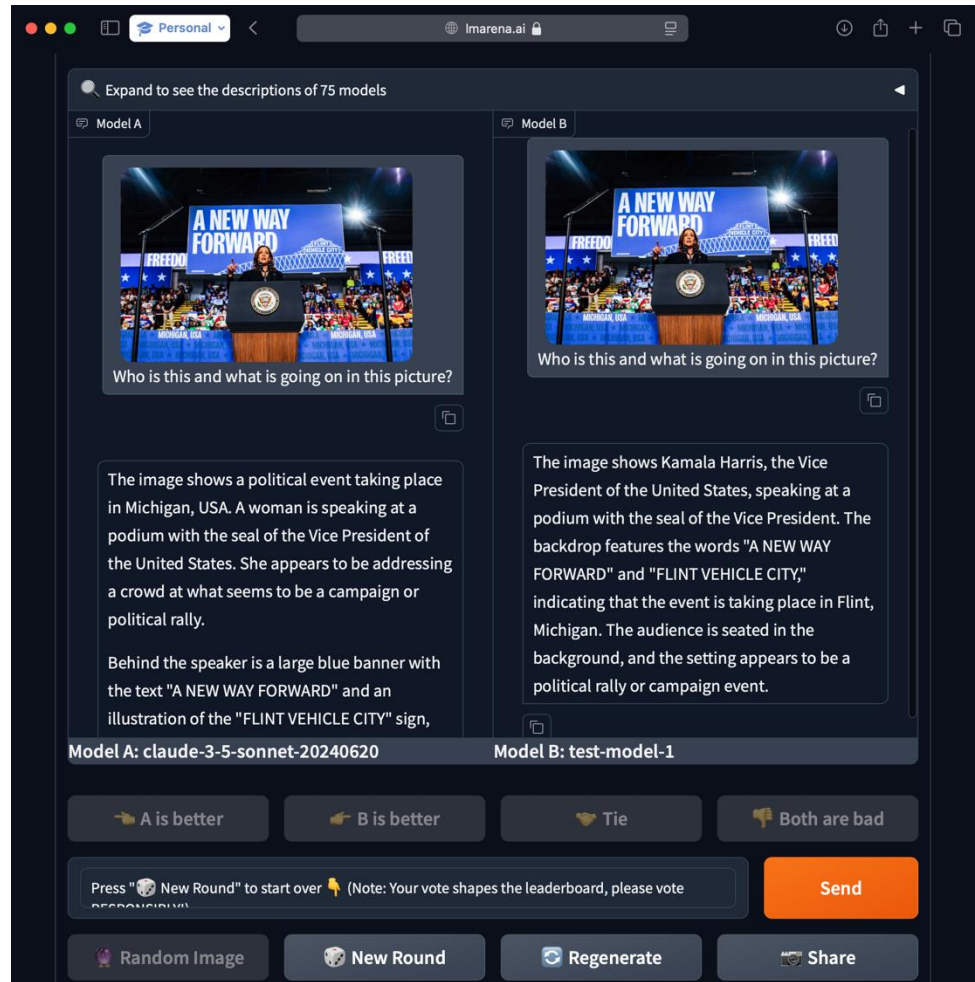
The Colosseum

An explosion of Arenas

Vision Arena

You can now upload **images** to conversations in the chatbot arena.

- Great way to convert a plot back to python code!
- **Data release** coming soon!



The screenshot displays the Vision Arena interface in a browser window. At the top, the browser address bar shows "Imarena.ai". The main interface is divided into two columns for "Model A" and "Model B". Both models are presented with the same image: Kamala Harris speaking at a podium during a political event. The podium features the Vice President seal, and the background has a banner that reads "A NEW WAY FORWARD" and "FLINT VEHICLE CITY".

Model A's response:
"Who is this and what is going on in this picture?"
The image shows a political event taking place in Michigan, USA. A woman is speaking at a podium with the seal of the Vice President of the United States. She appears to be addressing a crowd at what seems to be a campaign or political rally.
Behind the speaker is a large blue banner with the text "A NEW WAY FORWARD" and an illustration of the "FLINT VEHICLE CITY" sign,

Model B's response:
"Who is this and what is going on in this picture?"
The image shows Kamala Harris, the Vice President of the United States, speaking at a podium with the seal of the Vice President. The backdrop features the words "A NEW WAY FORWARD" and "FLINT VEHICLE CITY," indicating that the event is taking place in Flint, Michigan. The audience is seated in the background, and the setting appears to be a political rally or campaign event.

At the bottom of the interface, there are buttons for "A is better", "B is better", "Tie", and "Both are bad". Below these is a text input field with the instruction "Press 'New Round' to start over" and a "Send" button. At the very bottom, there are buttons for "Random Image", "New Round", "Regenerate", and "Share".

RedTeam Arena

Evaluating **model safety**
AND **human prompting** skills.

➤ **Gamification**

<https://redarena.ai>



The screenshot shows the RedTeam Arena website in a browser window. The browser's address bar displays "redarena.ai". The website has a dark theme with a white header containing the "RedTeam Arena" logo and three buttons: "LOGIN", "LEADERBOARD", and "REGISTER". The main content area features the text "BAD WORDS" in large, glowing cyan letters. Below this, it says "YOU HAVE ONE MINUTE TO JAILBREAK THE MODEL." and "THE FASTER, THE BETTER." in white. A prominent "START GAME" button is centered. At the bottom, there is a small blue pill-shaped button that says "More games coming soon". The footer contains the copyright notice "© 2024 RedTeam Arena. All rights reserved." and links for "Home" and "Terms of Service".

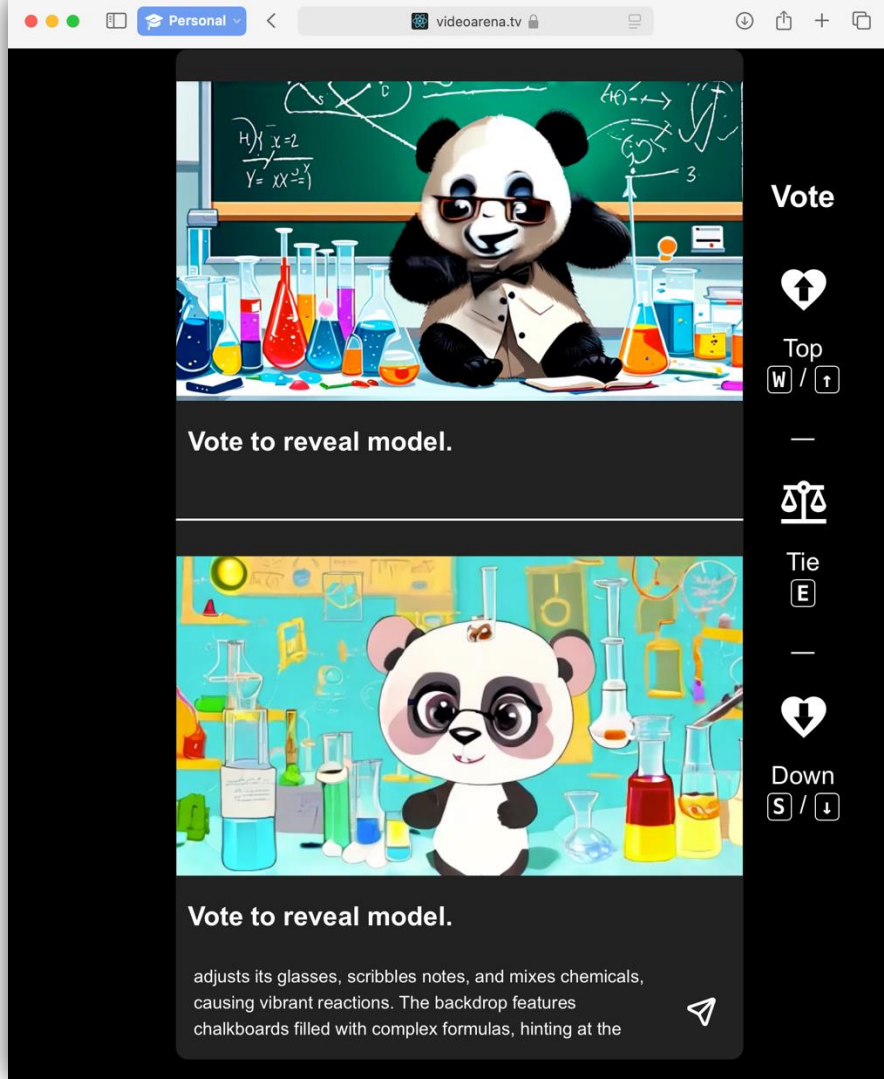
The Video Arena

Evaluating human alignment for video generation.

Rank	Model	Elo Rating ▼	95% CI	Votes ▼
1 🏆	Kling 1.0	1217	+10/-8	6351
2 🥈	Luma 1.6	1126	+9/-8	6352
3 🥉	Runway 3.0	1101	+10/-7	6350
4	Genmo 0.2	991	+9/-6	6352
5	SVD 1.0	927	+8/-7	6353
6	Pika Beta	855	+9/-8	
7	OpenSora 1.2	782	+11/-11	



<https://www.videoarena.tv>



The screenshot shows the Video Arena website interface. At the top, there's a browser address bar with "videoarena.tv". Below it, two video thumbnails are displayed. The top thumbnail shows a panda character in a lab coat and glasses, sitting at a desk with various chemistry flasks and a chalkboard in the background. The bottom thumbnail shows a similar panda character in a lab coat and glasses, standing in a chemistry lab setting. To the right of the thumbnails is a vertical sidebar with voting options: "Vote" (with a heart icon), "Top" (with a W icon), "Tie" (with an E icon), and "Down" (with an S icon). Below the thumbnails, the text "Vote to reveal model." is repeated twice. At the bottom of the page, there's a small text description: "adjusts its glasses, scribbles notes, and mixes chemicals, causing vibrant reactions. The backdrop features chalkboards filled with complex formulas, hinting at the".

Agent Arena

Evaluating **tool use** in open-ended tasks.

- Early prototype focuses on **specialized agent selection**.



<https://www.agent-arena.com>

A screenshot of the Agent Arena web application. The browser address bar shows 'agent-arena.com'. The page header includes 'Agent Arena' and navigation links for 'Arena', 'Leaderboard', 'Prompt Hub', 'Blog', 'FAQ', and 'Profile/Prompts', along with a 'Logout' button. The main heading is 'Agent Arena'. Below it, a welcome message states: 'Welcome to the LLM Agent Arena. Here, you can pit two agents against each other based on a goal you provide. You can also head to your profile to save prompts for agents and visit the Prompt Hub to see prompts used by other users along with their ratings. Ensure your API keys are configured in your profile for optimal performance.' A search bar contains the text 'Analyze the use of attention mechanisms in tr:'. Below the search bar are several category buttons: 'Search Example', 'Stock Example', 'Financial Data Example', 'Research Example', 'Automation Example', and 'Surprise Me'. A prominent 'Run Both Agents' button is centered. Two agent comparison cards are shown. The left card is for 'langchain ArXiv Article Fetcher (gpt-4-0613)' and includes a 'Show Left Raw Output' button. The right card is for 'llamaindex ArXiv Article Fetcher (gpt-4o-2024-08-06)' and includes a 'Show Right Raw Output' button. The output for Agent 1 is: 'Final Answer from Agent 1: priors, achieving competitive results. Most studies focus on modifying the attention mechanism to achieve higher ROUGE scores. There is also research on data pre-processing and post-processing. Attention head selection and manipulation strategies are also being investigated for feature injection, with case studies showing improvements in dialogue summarization.' The output for Agent 2 is: 'Full Output from Agent 2: summarization while significantly reducing the number of parameters compared to traditional multi-head self-attention ([source] (http://arxiv.org/pdf/2012.02144v1)). The se studies highlight the ongoing efforts to optimize attention mechanisms in transformers, focusing on efficiency and performance improvements, particularly for long-document summarization tasks.'



The Age of **Conversational Eval.**

Organizations will use **generative AI** to **interface with people** (users, customers, and employees) and accomplish **open-ended tasks.**

Need to evaluate AI in the **application setting** not on exams.

What you should be asking:

- How can I evaluate LLMs directly on my application?
 - Defining a rubric, create examples.



GENERATIVE AI
LARGE LANGUAGE MODELS

LARGE LANGUAGE MODELS

The Age of Synthetic Data (Augmentation)

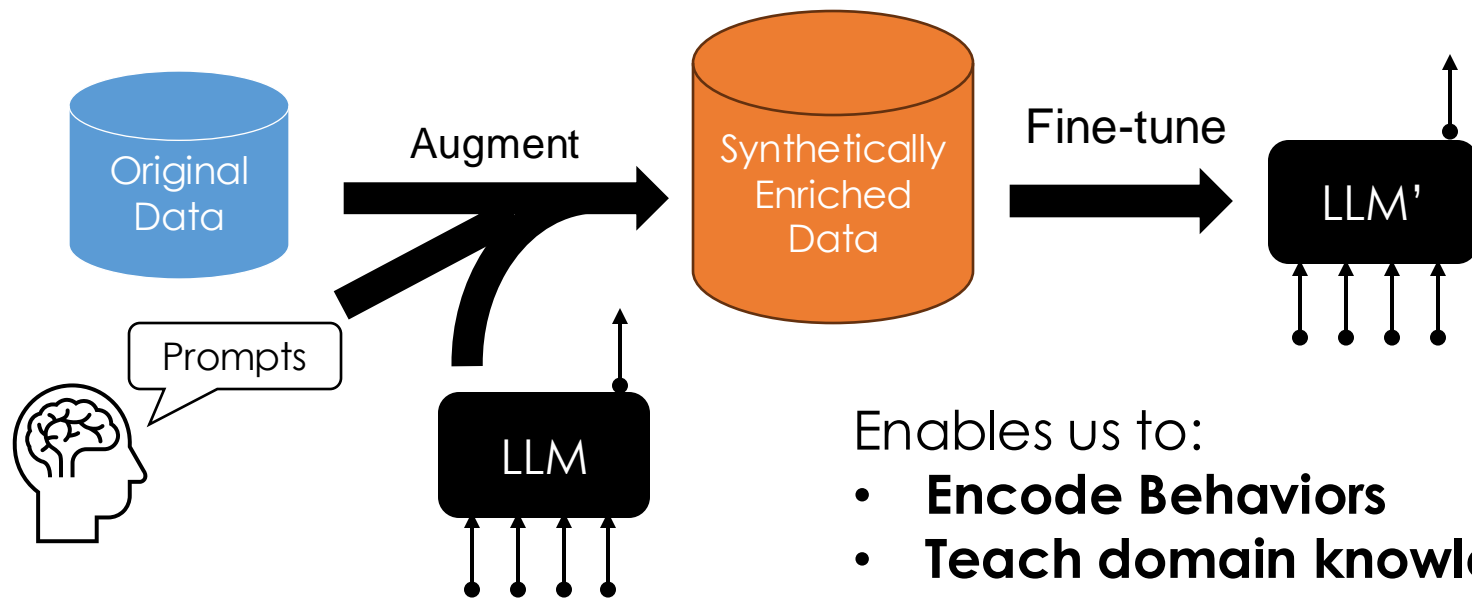
Teaching models with Data Augmentation

DATA AUGMENTATION



Synthetic Data Augmentation

Using an LLM to **augment data** to **fine-tune** an LLM



Enables us to:

- **Encode Behaviors**
- **Teach domain knowledge**

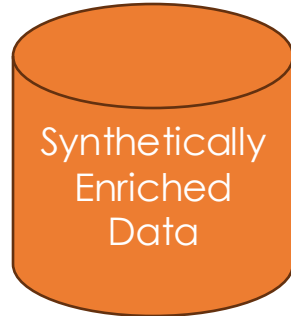


&



Synthetic Data is the Method!

Transform **documentation** into a **synthetic fine-tuning dataset**.



Sample a Document (Data)

1. Generate a **Question**
2. Generate an **Answer**
3. Generate an **Explanation**

Construct a Training Example (Synthetic Data)

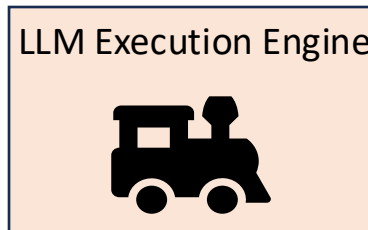
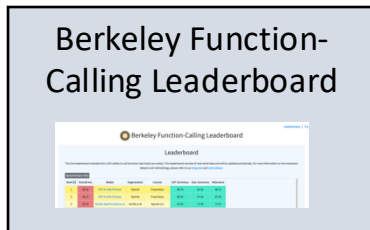
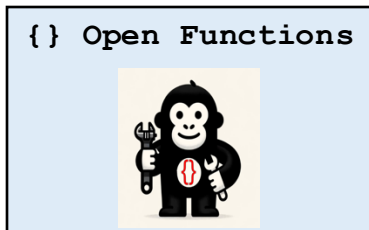
- Re-frame existing data to reflect **new tasks** (e.g., RAG) and **behaviors** (e.g., function calling)
- **Emphasize** important concepts in data



Gorilla



gorilla.cs.berkeley.edu



Studying the use of Tool (APIs)

<https://gorilla.cs.berkeley.edu>



Thesis:

LLMs will be the interface to the **services** and **knowledge** of the **web**.

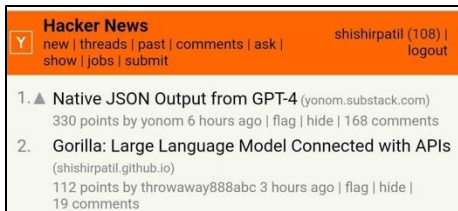
Hey Claude, should I bike to work tomorrow.

Yes, it looks like the weather is nice and I can move your 8AM meeting to the afternoon.

Would you like me to arrange it?

Yes!

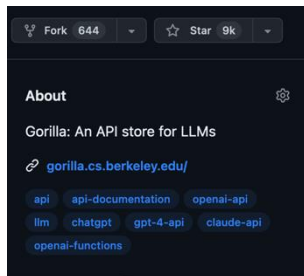
Impact - Gorilla



Hacker News
new | threads | past | comments | ask | show | jobs | submit

shishirpatil (108) | logout

- ▲ Native JSON Output from GPT-4 (yonom.substack.com)
330 points by yonom 6 hours ago | flag | hide | 168 comments
2. Gorilla: Large Language Model Connected with APIs (shishirpatil.github.io)
112 points by throwaway888abc 3 hours ago | flag | hide | 19 comments



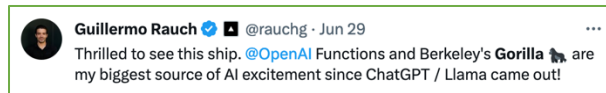
Fork 644 | Star 9k

About

Gorilla: An API store for LLMs

gorilla.cs.berkeley.edu/

api | api-documentation | openai-api | llm | chatgpt | gpt-4-api | claude-api | openai-functions



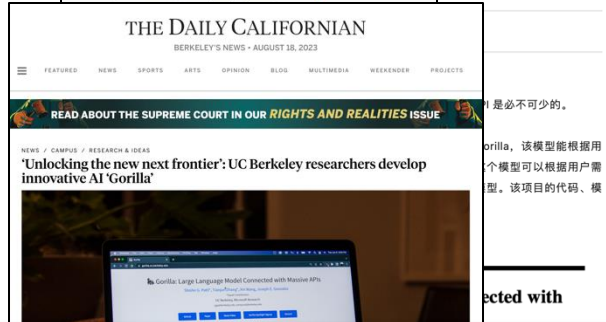
Guillermo Rauch @rauch · Jun 29

Thrilled to see this ship. @OpenAI Functions and Berkeley's Gorilla are my biggest source of AI excitement since ChatGPT / Llama came out!



Jim Fan @DrJimFan · May 27

The next iteration of Toolsformer and HuggingGPT: Gorilla, a Llama finetuned to use API tools



THE DAILY CALIFORNIAN
BERKELEY'S NEWS • AUGUST 18, 2023

首个大规模使用工具的大模型来了：伯克利发布Gorilla

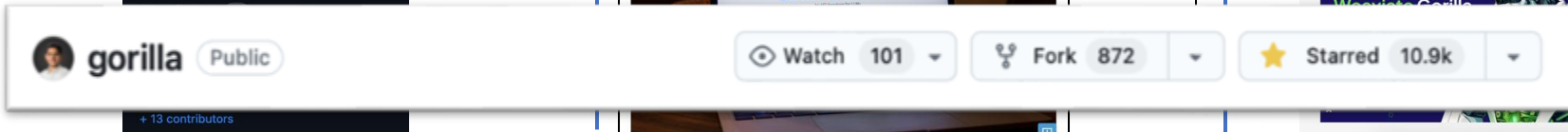
READ ABOUT THE SUPREME COURT IN OUR RIGHTS AND REALITIES ISSUE

NEWS / CAMPUS / RESEARCH & IDEAS

'Unlocking the new next frontier': UC Berkeley researchers develop innovative AI 'Gorilla'

Gorilla, 该模型能根据用户模型可以根据用户需求。该项目的代码、模

ected with

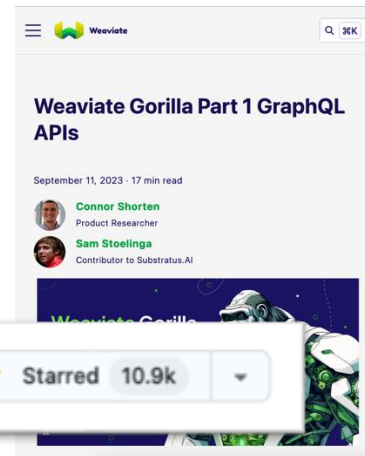


gorilla Public

+ 13 contributors

Watch 101 | Fork 872 | Starred 10.9k

500,000+ invocation in Hosted endpoint



Weaviate Gorilla Part 1 GraphQL APIs

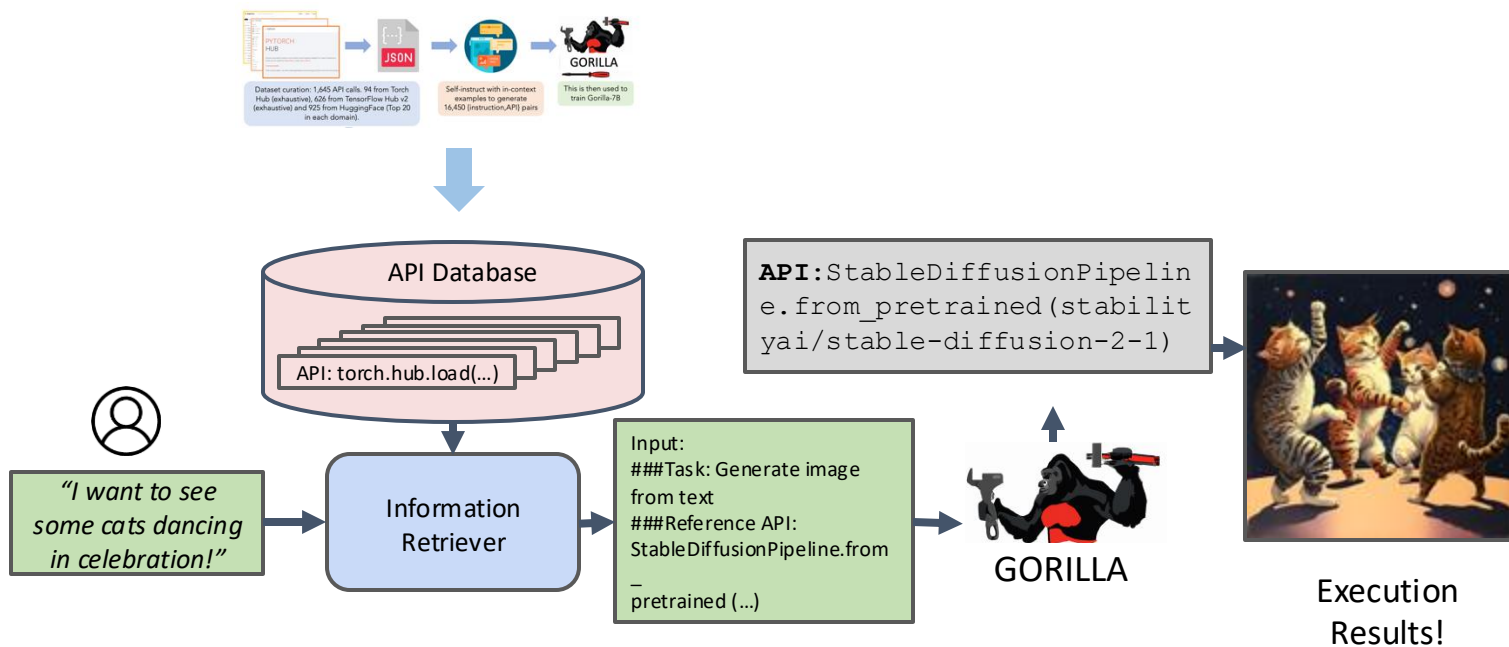
September 11, 2023 · 17 min read

Connor Shorten
Product Researcher

Sam Stoelings
Contributor to Substatus.AI

Used at: Adobe, Apple, CMU, CISCO, Cloudera, Dropbox, UCLA, GNU, IBM, Intel, Intuit, Linkedin, Microsoft, MIT, Netflix, Nvidia, Tesla, ...

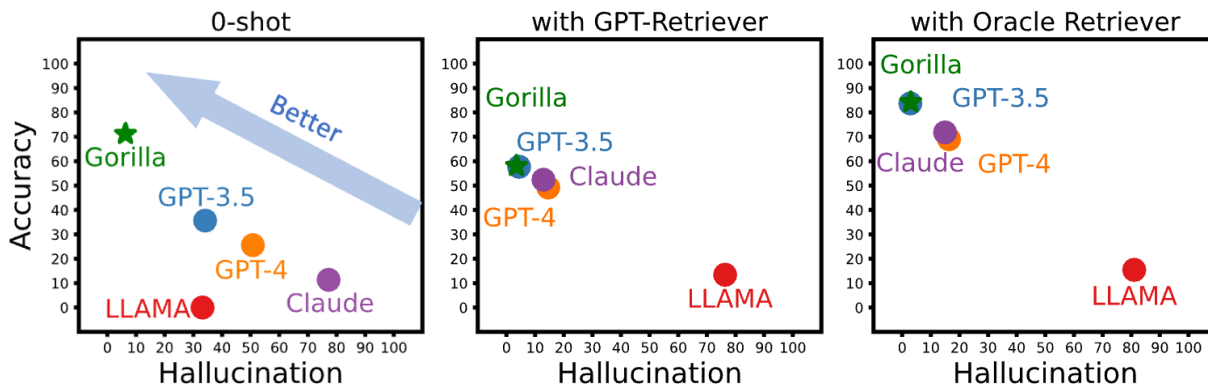
Combine **Retrieval** and **Fine-Tuning** to **Discover** and **Invoke** APIs



Big Idea: Retrieval Aware Training (RAT)

Fine-tune the model to **use** or **ignore** retrieved context.

- Introduce *correct* and *incorrect* **retrieval** results during **instruction** fine-tuning
- Ensures model is **robust to low-quality retrieval**





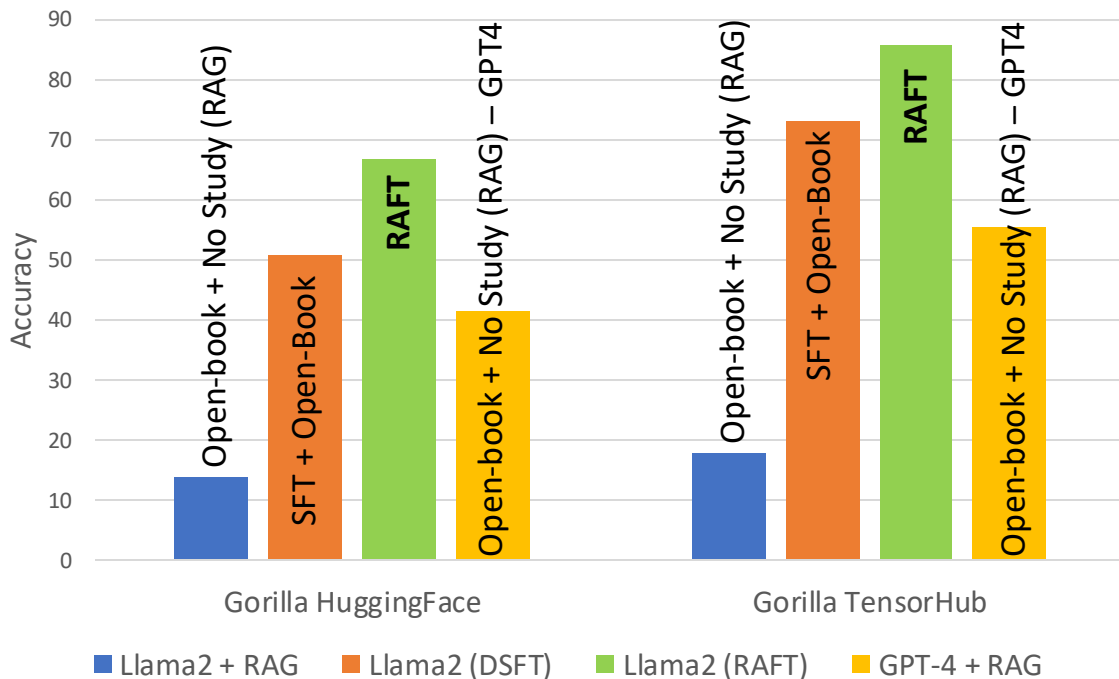
Retrieval Aware Fine-Tuning Enhanced RAG



- Explored how to **fine-tune models** for **domain specific RAG**
 - If you know the **documents** and **retrieval process** in advance tune the LLM for that setting
- **Synthetic Data Augmentation:** transforms a **collection of documents** into a fine-tuning dataset that teaches:
 - How to **ignore distracting documents**
 - How to **read documents in the domain**
- Offered as part of [Azure Studio](#) and [Meta Training Recipes](#)

RAFT improves performance for RAG

RAFT Performance on Gorilla API Benchmark





&



How do we mix **Fine-Tuning** and **Retrieval**?

Hypothesis (at least what I wished was true):

- **Fine-Tuning**: augment the **behavior** of the model
 - **Retrieval**: introduce new **knowledge** to the model
- I was wrong!

Early Evidence (Gorilla and RAFT): fine-tuning is remarkably effective at incorporating behavior and knowledge

- Fine-tuning on the **synthetic data** improved **behavior** and **knowledge**.



The Age of **Synthetic Data**

Organizations will use **generative AI** to transform **existing datasets** into new **synthetic datasets**

- Enable the next generation of **advance analytics**
- Encode **behaviors** and **domain knowledge** for **fine-tuning**

What you should be asking?

- How can I **augment existing data** to **capture desired behaviors** and **knowledge**.



Corollary: The Age of **Fine-Tuning**

Organizations will create many **fine-tuned versions** of models to **specialize** in **specific domains** and **behaviors**

- Enable **smaller specialized LLMs** to **out-perform larger generalist models**

What you should be asking:

- Are there **sub-domains** on **which I could fine-tune** (e.g., corporate RAG, function calling, doc-string generation, ...)

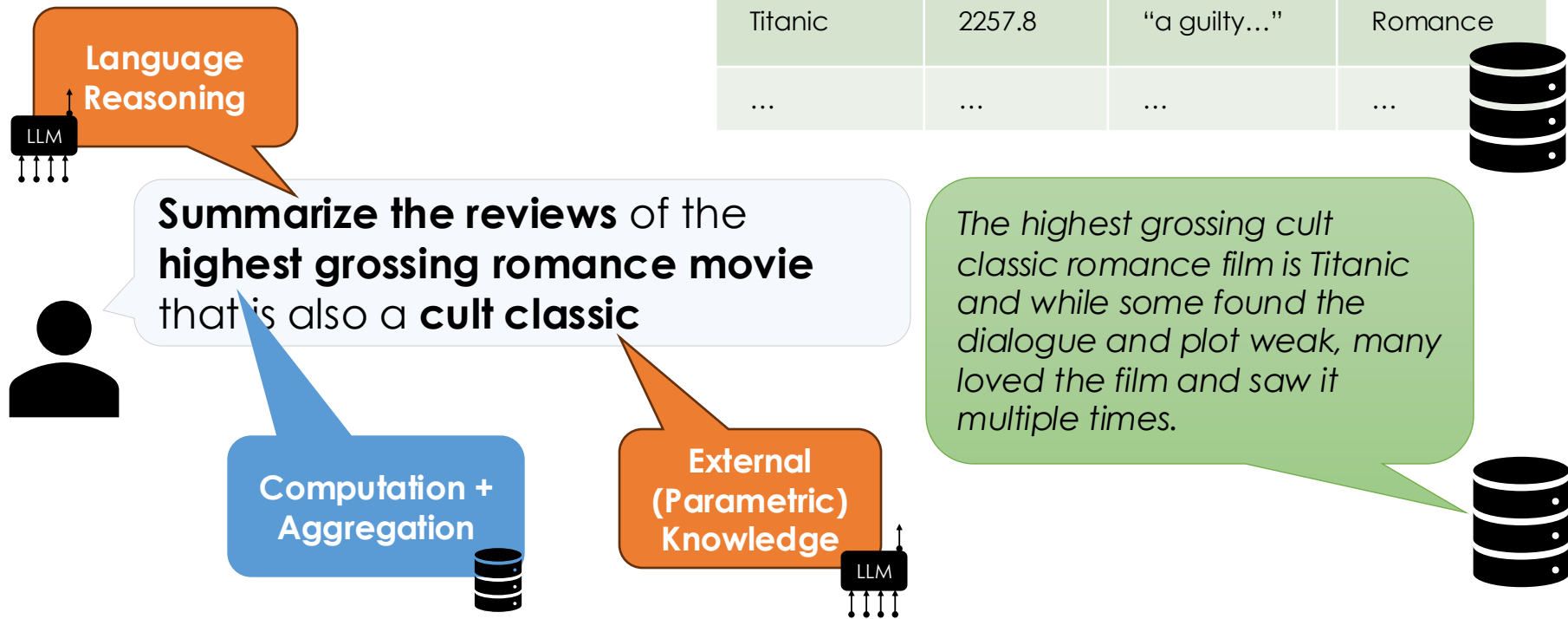


LLM Analytics

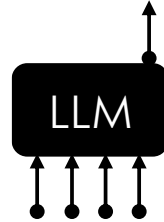
Answering **natural language questions** over **structured data**
... or why **Text2SQL** is not enough

Talk to your data!

movie_title	revenue	review	genre
Shang-Chi	432.2	"solid film..."	Action
Titanic	2257.8	"still best..."	Romance
Titanic	2257.8	"a guilty..."	Romance
...



Language Models vs. Database Systems



vs.



Weaknesses:

- Bad at **math/logical analysis**
- **Expensive** and **limited context window**
- **Out-of-date** knowledge

Strengths:

- **Converses** in natural language
- **Reason about text** and **images**
- **Parametric world knowledge**

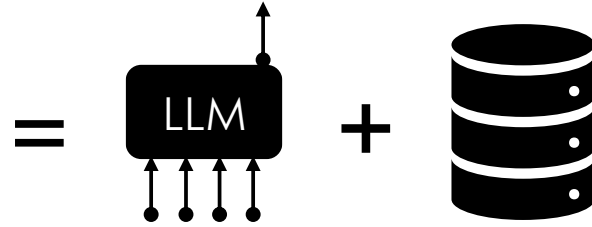
Weaknesses:

- **Can't converse in natural language**
- Limited **text/img reasoning**
- Limited to **facts in the DB**

Strengths:

- Excellent at **math** and **logic**
- **Efficient** on large datasets
- **Domain** knowledge and **up-to-date facts**

Table Augmented Generation



<arXiv>



Goal:

Combine the **language reasoning** and **world knowledge** of LLMs with the **computational accuracy** and **performance** of **Database Systems**

[No SQL] Ask questions and get answers in **natural language**

[No AI-Math] Use the database for **all logical computation**

[Read Text/Imgs] Augment the DBMS with **natural language ops**

[Synthetic Data] Leverage LLM's internal knowledge to **augment data**

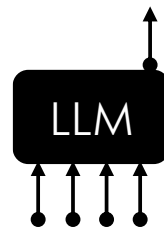
TAG's 3 Stage Model

Text2SQL++: New LLM Ops and goal is to **extract relevant data**

1. Query Synthesis: question → query program



"Summarize the reviews of the highest grossing classic romance movies"

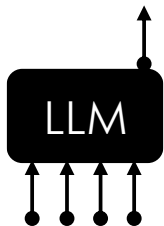


```
WITH CRM AS (SELECT * FROM movies WHERE genre = 'Romance'  
AND LLM('{movie_title} is a classic') = 'True')  
SELECT * FROM CRM  
WHERE revenue = (SELECT MAX(revenue) FROM CRM);
```

2. Query Execution: query program → filtered data/answer



+



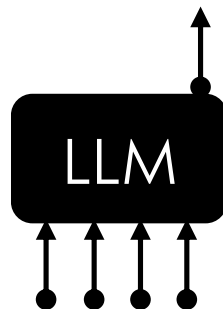
movie_title	revenue	review	genre
Titanic	2257.8	"still best..."	Romance
Titanic	2257.8	"a guilty..."	Romance
...

TAG's 3 Stage Model (cont.)

3. Answer Generation: question + filtered data → answer

“Summarize the reviews of the highest grossing classic romance movies”

“{movie_title: 'Titanic', revenue: 2247.8, review: 'still best...', genre: 'Romance'}...”



“The reviews of Titanic discuss the on-screen chemistry...”

TAGBench:

Asking more realistic questions



- Augment prior benchmarks with **world knowledge** and **language reasoning** components
 - **World Knowledge Example:** *What is the grade span offered in the school with the highest longitude **in cities in that are part of the 'Silicon Valley' region?***
 - **Language Reasoning Example:** *Of the 5 posts with the highest popularity, **list their titles in order of most technical to least technical.***
- Handwritten TAG pipelines **outperform prior methods** (Text2SQL, RAG) **by over 50%**.

Scaling LLM Analytics

Addressing the cost of TAG...

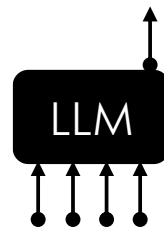
TAG's 3 Stage Model

Text2SQL++: New LLM Ops and goal is to **extract relevant data**

1. Query Synthesis: question → query program



"Summarize the reviews of the highest grossing classic romance movies"

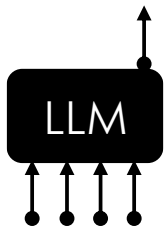


```
WITH CRM AS (SELECT * FROM movies WHERE genre = 'Romance'  
AND LLM('{movie_title} is a classic') = 'True')  
SELECT * FROM CRM  
WHERE revenue = (SELECT MAX(revenue) FROM CRM);
```

2. Query Execution: query program → filtered data/answer



+



movie_title	revenue	review	genre
Titanic	2257.8	"still best..."	Romance
Titanic	2257.8	"a guilty..."	Romance
...

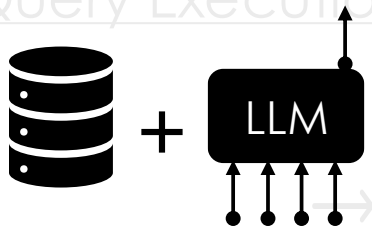
TAG's 3 Stage Model

Text2SQL++: New LLM Ops and goal is to extract relevant data

Executing new LLM operators in Stage 2 is expensive!

```
"Summarize the reviews of the highest  
WITH CRM AS (SELECT * FROM movies WHERE genre = 'Romance'  
AND LLM('{movie_title} is a classic') = 'True')  
SELECT * FROM CRM  
WHERE revenue = (SELECT MAX(revenue) FROM CRM);
```

2. Query Execution: query program → filtered data/answer



Requires **invoking an LLM** on potentially **millions of rows of data.**

Titanic	2257.8	"a guilty..."	Romance
...

LLM Operations are Expensive!

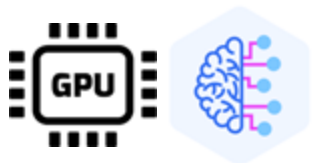


TPC-DS benchmark

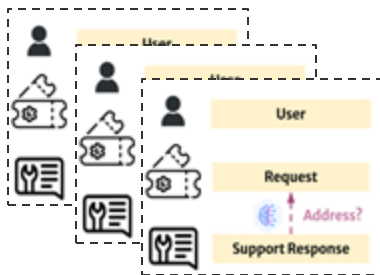
100GB



1 second



NVIDIA L4
Llama 7B



100GB



96 days!

\$\$\$

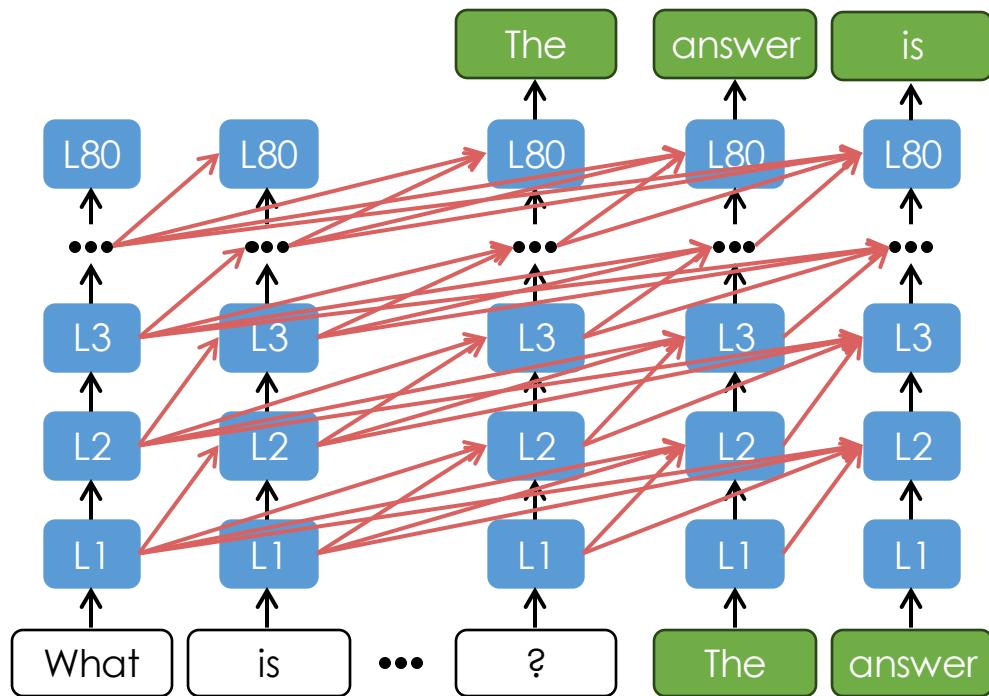


8 million times longer

Why are LLMs so Expensive?

- **>70Billion FLOPs per token!**
- **Auto-regressive decoding**

Significant opportunity **for reuse of common prefixes.**



Research in KV-Cache Management



<https://github.com/vllm-project/vllm>



<https://github.com/sgl-project/sglang>

A big focus of several of my projects is on how to manage this token level KV-cache to **reduce fragmentation** and **improve reuse**.

- **Maximize KV-Cache reuse** by scheduling **common prefixes** nearby in **time**. (LRU cache eviction)

LLM Prefix Sharing: SQL Example

SELECT LLM (

“Did {support_response} address {request}?”, support_response, request

) **AS** success,

FROM customer_tickets



Request

"How can I reset my password?"

"I forgot my password, what should I do?"



Support Response

“Please follow the instructions on the password reset page.”

Shared *support_response* across many *requests*

Increase Prefix Sharing: Reorder Rows & Columns

 Prefix Sharing Count

Zipcode	Country	Name
94709	US	John
94709	US	Jenny
57110	China	Henric
58232	China	Vivi

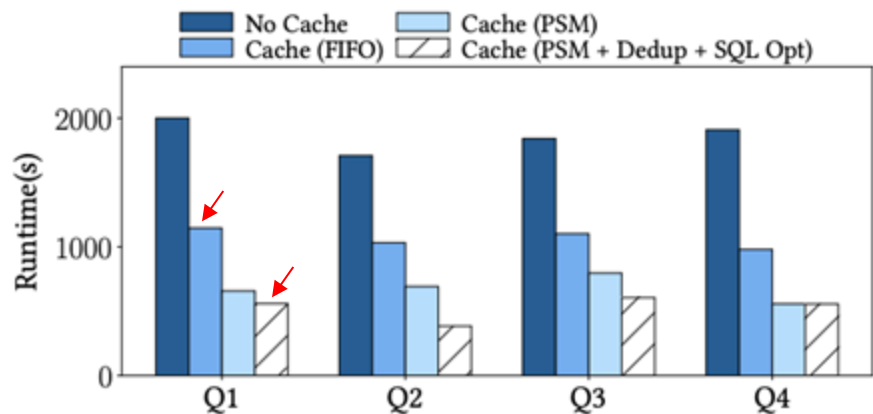


Country	Zipcode	Name
US	94709	John
US	94709	Jenny
China	57110	Henric
China	58232	Vivi

Reorder columns based on column statistics (sharing factor, average length, etc.) for prefix reuse

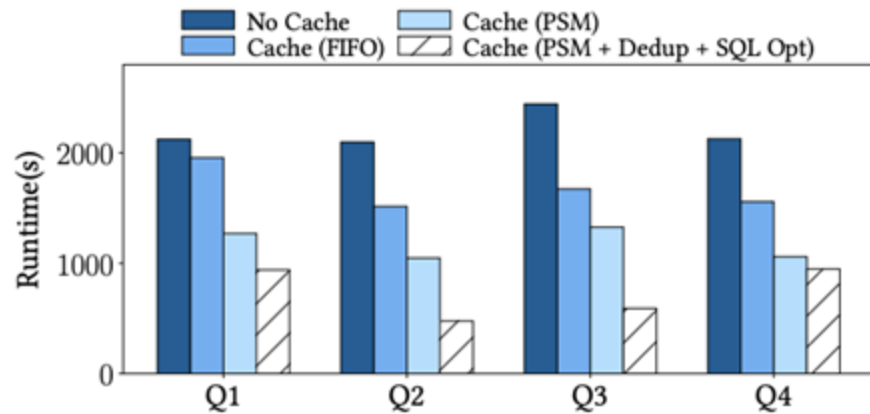
Evaluation: Speed-Up Over Baselines

1.8-2.7x



(a) Rotten Tomatoes Movies Dataset

1.6 - 3.2x



(b) Amazon Products Dataset

➤ Significant savings! But still not enough ...

Ongoing Research

- Designing the **Text2SQL++** query pipeline
- **Scheduling** and **autoscaling** LLM inference
- **Clustering** rows to **approximately reuse generations**
- **Streaming distillation** from the LLM to a light-weight (classical) ML model
- **Tabular encodings** for the generation stage of TAG

...



The Age of **LLM Analytics**

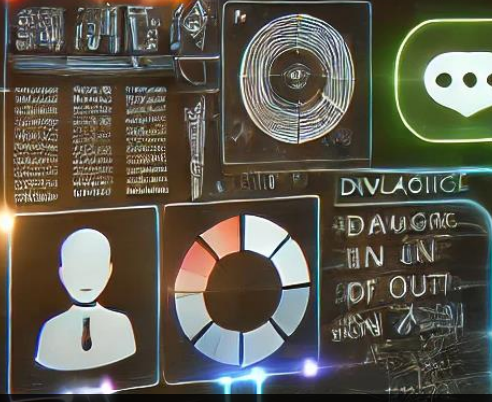
Organizations will use **LLM & VLMs** to interact with their structured **data** and **data systems**

- Enable **more people** to ask challenging questions
- Extracting deeper insights from **images** and **text data**

What you should be asking?

- Do I have under analyzed **text** and **image data**?
- How can we reduce the **costs of running these systems**?

AI



AI DATABASE

PAGING INFORMATION



The Age of Context Management Systems

Emerging class of software systems designed to maintain a tidy context.

The Challenge of context

Context Window (128K Tokens)



- Everything needs to fit in context
 - Context windows have grown from 2K to 128K tokens!
- The problem with **long context** (using all 128K tokens)
 - **Expensive:** computation scales **quadratically**
 - **Lost in the middle:** not all of context is used equally
 - **Distracting Context:** models are sensitive to bad context

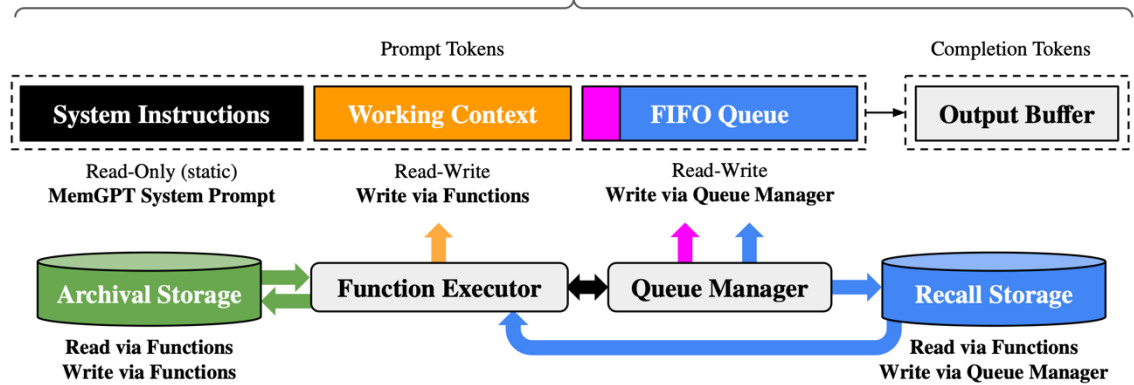


MemGPT

<https://memgpt.ai>

A Context Management Platform for Long-Context Agents

LLM Finite Context Window (e.g. 8k tokens)



← Segment the Context

← Leverage **tool-use** to interact with external storage

The entire agenda is around ensuring **only the right information** is in context.

Leveraged ideas in **operating system paging** to enable **infinite virtual context** while using a **small physical context**.

Y **Hacker News** new | past | comments | ask | show | jobs | submit

▲ **MemGPT – LLMs with self-editing memory for unbounded context** (github.com/cpacker)

363 points by shishirpatil 7 months ago | hide | past | favorite | 85 comments

Y **Hacker News** new | past | comments | ask | show | jobs | submit

▲ **MemGPT: Towards LLMs as Operating Systems** (arxiv.org)

225 points by belter 7 months ago | hide | past | favorite | 106 comments

MemGPT Giving LLMs Infinite Memory (Big S...
241K views · 7 months ago
Matthew Berman
In this video, we look at MemGPT, a new way to give AI unlimited
4K
MemGPT Research Paper [...]
3 chapters

MemGPT: Towards LLMs as Operating Systems (arxiv.org)
225 points by belter 7 months ago | hide | past | favorite | 106 comments

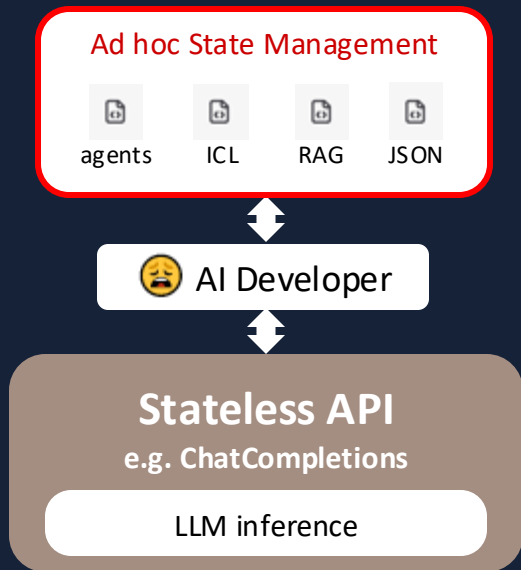
MemGPT | Infinite Memory Superpower | Hands-on and In-...
3.4K views · 7 months ago
Prompt Engineer
Let's test out MemGPT. This give the LLM superpowers ...
Intro | Intro to Paper ... 17 chapters

MemGPT: Creating Powerful Agents with Unlimited Memory...
6.1K views · 2 weeks ago
WorldofAI
Welcome to our video where we delve deep into the revolutionary framework known as MemGPT. If you're...

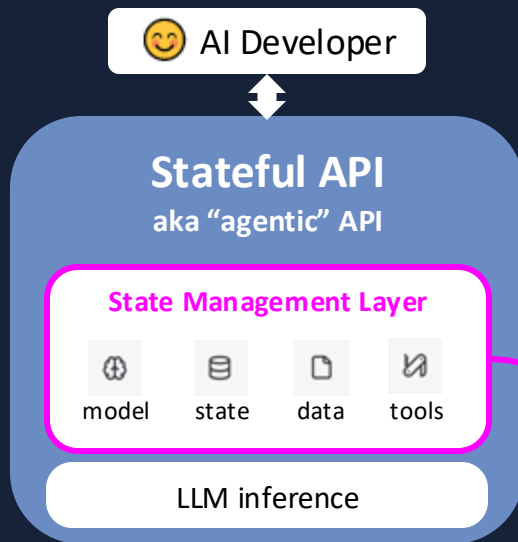
MemGPT: The Future of LLMs with Unlimited Memory
8.6K views · 6 months ago
AssemblyAI
This video is a complete guide to MemGPT - which is
Introduction | MemGP... 8 chapters

LlamaIndex Webinar: Long-Term, Self-Editing Memory wit...
4.1K views · 2 months ago
LlamaIndex
Long-term memory for LLMs is an unsolved problem, an...
Presentation | Demo [...]
3 chapters

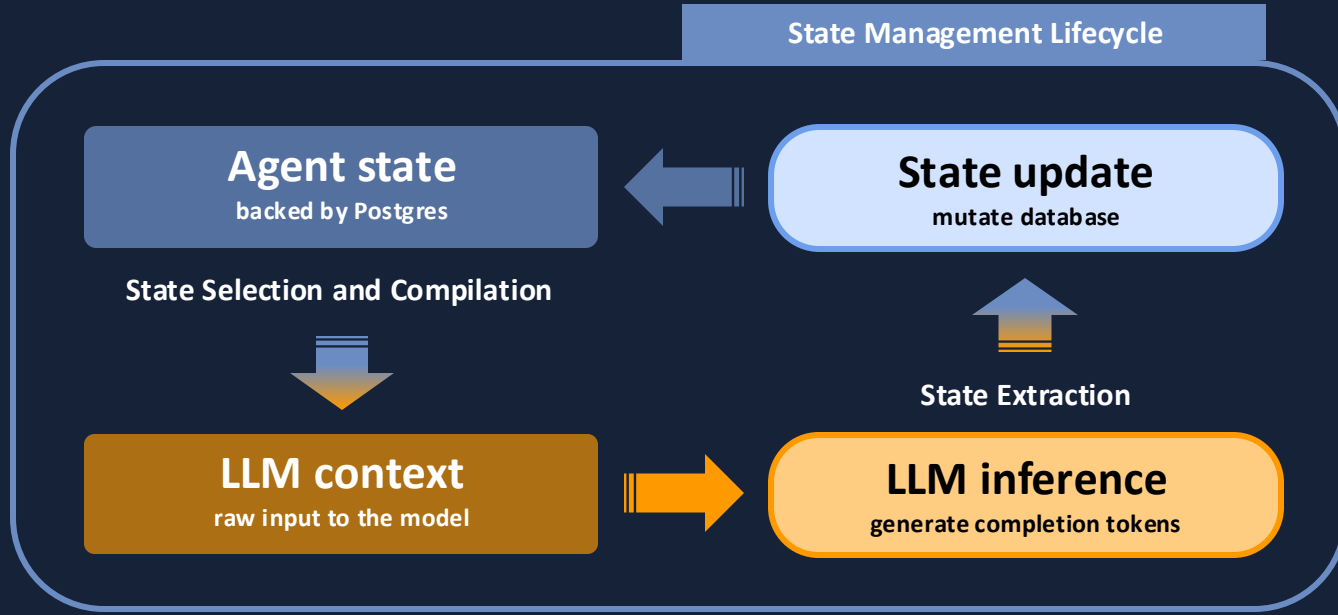
2020-now: LLM APIs



the future: agentic APIs



The missing piece to agents is the **state management layer**



State management is the key to making reliable, long-running agents

- Store **agent state** in a canonical (model-agnostic) data format
- Materialize **agent state** into **LLM context** at every reasoning step



The Age of **Agents**

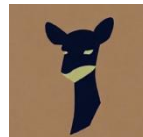
There is an emerging class of **Agentic Systems** that combine **tool use** with repeated **LLM invocation** to complete complex tasks.

- Leverage **problem decomposition to manage context**
- **Incorporate tools** to manage interaction with environment

What you should be asking:

- Where could I break existing LLM tasks/calls into smaller **well-defined tasks** with **narrower contexts**?

Conclusion



TAG LLM
SQL



MemGPT

A Last thought for EPIC Data Lab

Data is the programming language for Generative AI and Generative AI is the new programming language for data and this cycle is all about optimizing for humans.

Thank You!

A vision for two **Emerging Fields**

AI-Engineering

The study of the design, development, and operation of AI-centric software systems.

(Engineering)



AI-Psychology

The study of AI behavior and its interaction with human behavior.

(Science)

A ***vision*** for the future of
Generative AI Research

A vision for two **Emerging Fields**

AI-Engineering

The study of the design, development, and operation of AI-centric software systems.

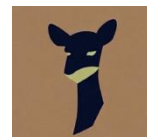
(Engineering)

 MemGPT



TAG
Analytics







Unlock new apps and functionality!

A vision for two **Emerging Fields**

AI-Psychology

The study of AI behavior and its interaction with human behavior.

(Science)

