



# OttoBot: Productionizing LLMs

Lukas Biewald

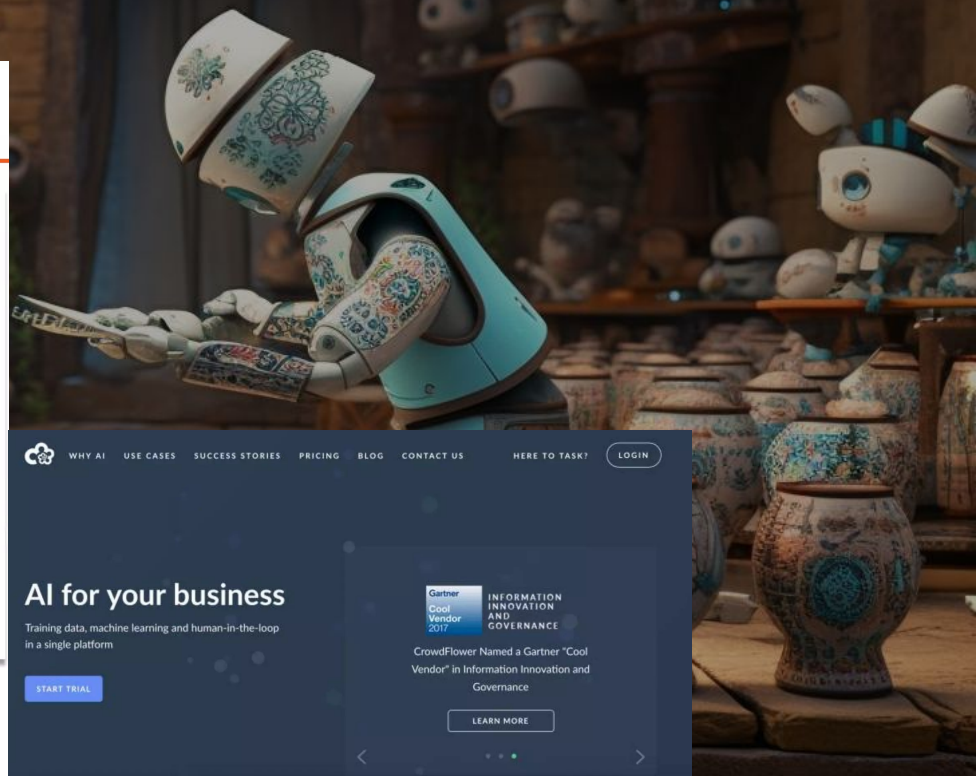
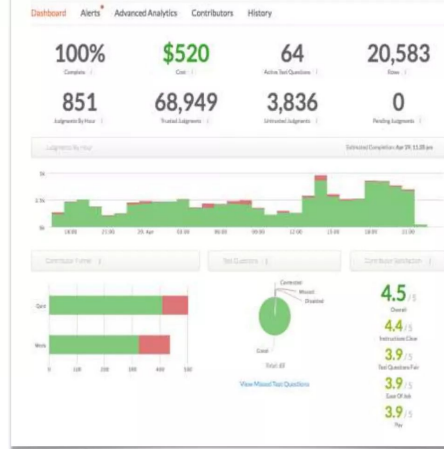
# CrowdFlower/Figure Eight

## CrowdFlower Data Enrichment Platform



Can you mark the clouds in this image?

- There are clouds in this image that can be marked
- This image is all clouds!
- This image has no clouds!
- I cannot see this image or I cannot mark the clouds.



WHY AI USE CASES SUCCESS STORIES PRICING BLOG CONTACT US HERE TO TASK? LOGIN

## AI for your business

Training data, machine learning and human-in-the-loop in a single platform

START TRIAL



CrowdFlower Named a Gartner "Cool Vendor" in Information Innovation and Governance

LEARN MORE

# A \$100 Robot That Sees

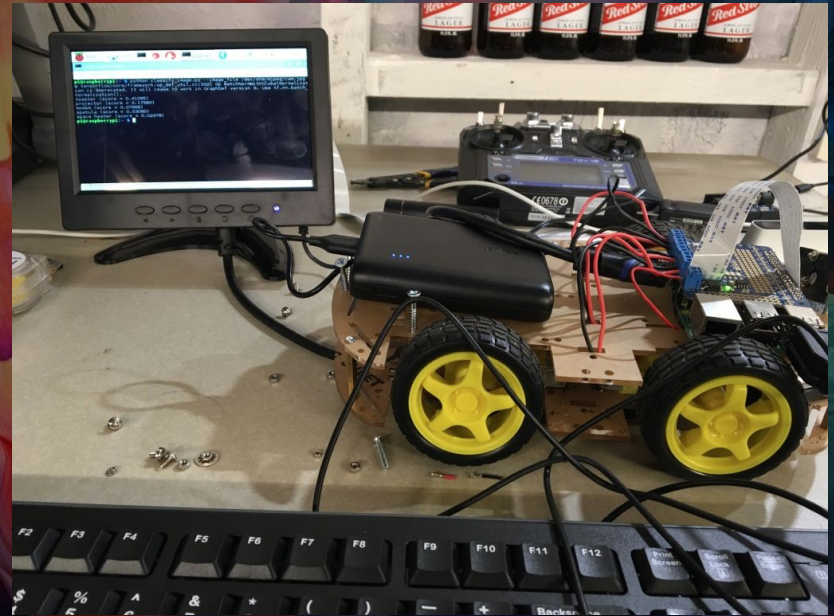
Latest Articles / Artificial Intelligence

## How to build a robot that "sees" with \$100 and TensorFlow

Adventures in deep learning, cheap hardware, and object recognition.

By [Lukas Biewald](#)

September 21, 2016





# Weights & Biases The AI Developer Platform



W&B Core: A foundational framework supporting AI developers



**Artifacts**  
Version and manage your ML pipelines



**Tables**  
Visualize and explore your ML data



**Reports**  
Document and share your ML insights

Secure enterprise deployments



SaaS Cloud

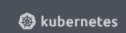


Dedicated Cloud



Customer Managed

Flexible multi-cloud support





## PROBLEM

# Massive Developer Tools Gap for ML Ops

## SOFTWARE 1.0 - WRITE CODE

Design  
Interface



Figma



Sketch

Version  
Code



GitHub



GitLab

Manage  
Projects



Jira



VS Code

Deploy to  
Infra



HashiCorp



puppet

CI/CD



circleci



Jenkins

Prod  
Monitoring



PagerDuty



DATADOG

## SOFTWARE 2.0 - TRAIN MODELS

Version data &  
pipelines

- Files in Object Store

Prep &  
visualize data

- Notebooks
- Custom apps

Experiment  
tracking

- Text files
- Screenshots

Optimize  
models

- Shell scripts

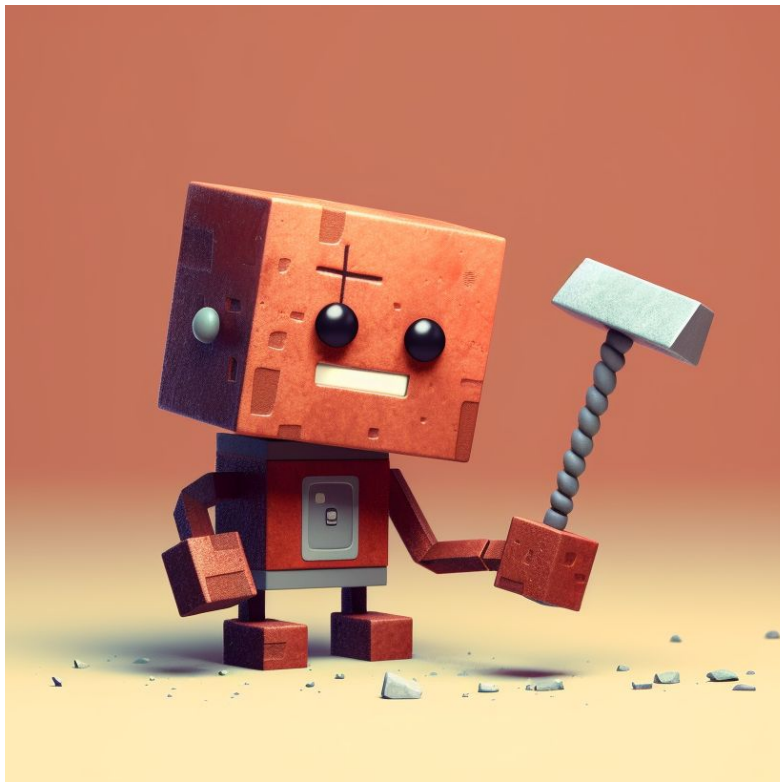
CI/CD for  
models

- Custom scripts

Production  
monitoring

- Ad-hoc

# Writing Code



# Building Models





Clemens Mewald

Jan 31, 2019 · 12 min read ·  Listen



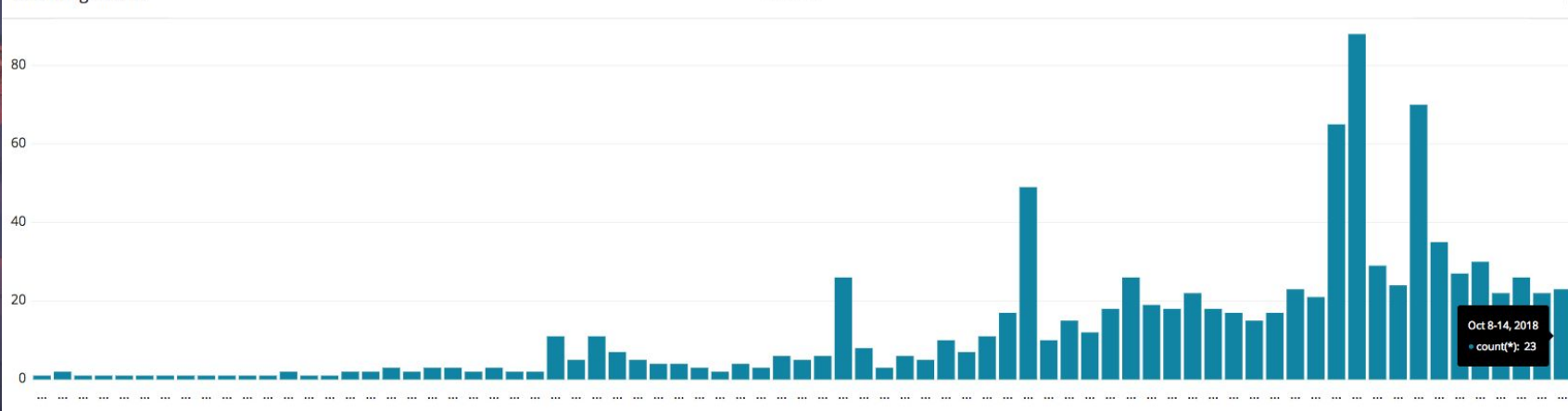
## Your Deep-Learning-Tools-for-Enterprises Startup Will Fail

I usually write about how to integrate and launch ML/AI in consumer-facing products. However, a large part of my job is building ML/AI developer tools, some of which are open sourced. In this field there is a



# First Year of Weekly Active Users

Returning Actives



# Collaborate with users





# Listen to feedback

NPS Score(last week)  
RESPONSES LAST WEEK

85

▲ 7.59%

“ It is **very easy to track the performance** of my model and also that of my coworkers.

“ It saves time! And I no longer have to keep track of logs in folders; yuck! Plus, **the API is super simple and easy to use.**



# Simple setup

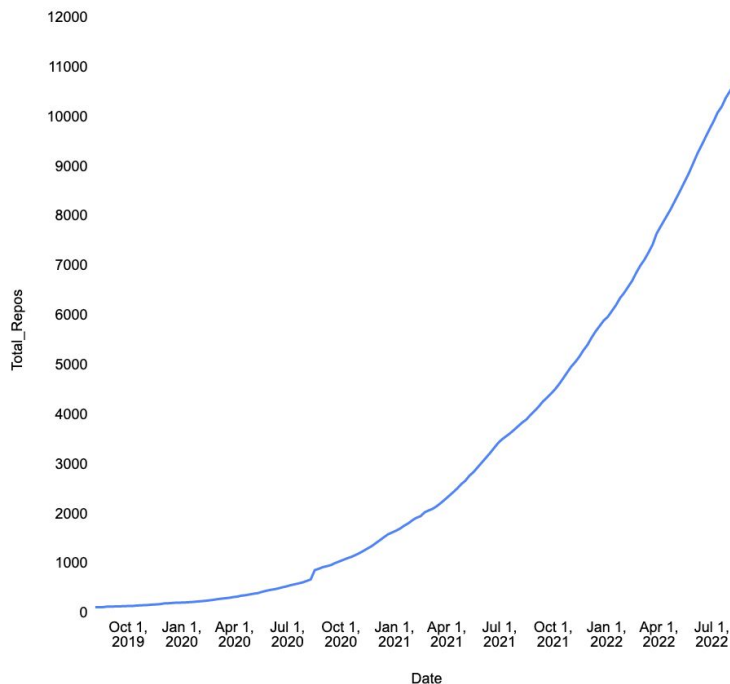
## 60 seconds to get up & running

Add a few lines of code to your ML project to immediately start seeing live charts.

```
# Start experiment  
wandb.init()  
  
# In training loop  
wandb.log(metrics)
```

# Integrate with everything

Repo Instrumentations



**spaCy** @spacy\_io · Oct 15

**spaCy** v3.0's training system also makes it easy to integrate with other tools – like [@weights\\_biases](#) for experiment tracking! The built-in logger tracks your config (and registered functions!) and lets you discover patterns & correlations.

[Details: nightly.spacy.io/usage/projects...](#)



**PyTorch** @PyTorch · Oct 22

The team [@WadhvaniAI](#) has built a multi-task network that detects pest infestations in cotton crops. This technology is being put directly in the hands of more than 18,000 farmers across India using [#PyTorch](#) Mobile, TorchServe, and **Weights & Biases**.



**Jeremy Howard** @jeremyphoward · Aug 21

**fastai** includes many modules that add functionality, generally through callbacks. Thanks to the flexible infrastructure, these all work together, so you can pick and choose what you need (and add your own), like [@weights\\_biases](#) tracking

# Support collaboration everywhere

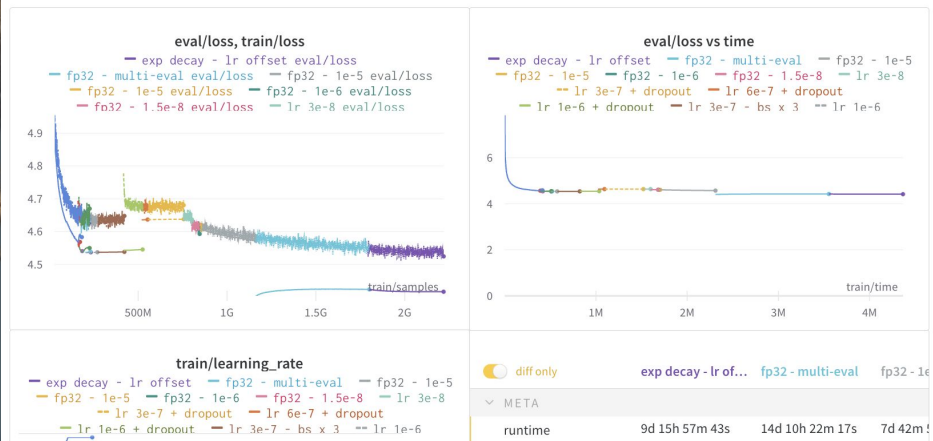


## Reports

Document and Share your  
ML Insights

Several runs are present:

- when the training seems to plateau, we lower learning rate (constant with warmup)
- we used dropout at around 500M train/samples and then stopped at around 750M train/samples (did not seem useful)
- at 500m train/samples, we have a higher eval/loss because we decided to switch validation set to a shard of training set due to some concern of different data distribution
- closer to the end of training, we started using exponential decay
- at 800M train/samples, we switched to full precision training (only 20% slowdown)
- at 1.2B train/samples, we now use 2 validation sets, the original one and the training shard we had switched to
- at 1.8B train/samples, we started exponential decay





# Make reproducibility easy



**Launch**

Package and Run your ML  
Workflow Jobs

Artifacts **iv3\_trained** 5334ab69740f9dda4fed

Model Registry →

Find matching artifacts

MODEL

- iv3\_trained
  - v2 latest
  - v1
  - v0
  - iv3

RAW\_DATA

- inat\_raw\_data

TEST\_DATA

- inat\_test\_data\_100
- inat\_test\_data\_10
- inat\_test\_data\_1

TRAIN\_DATA

- inat\_train\_data\_800
- inat\_train\_data\_80
- inat\_train\_data\_8

VAL\_DATA

- inat\_val\_data\_100

Version Metadata Usage Files **Lineage**

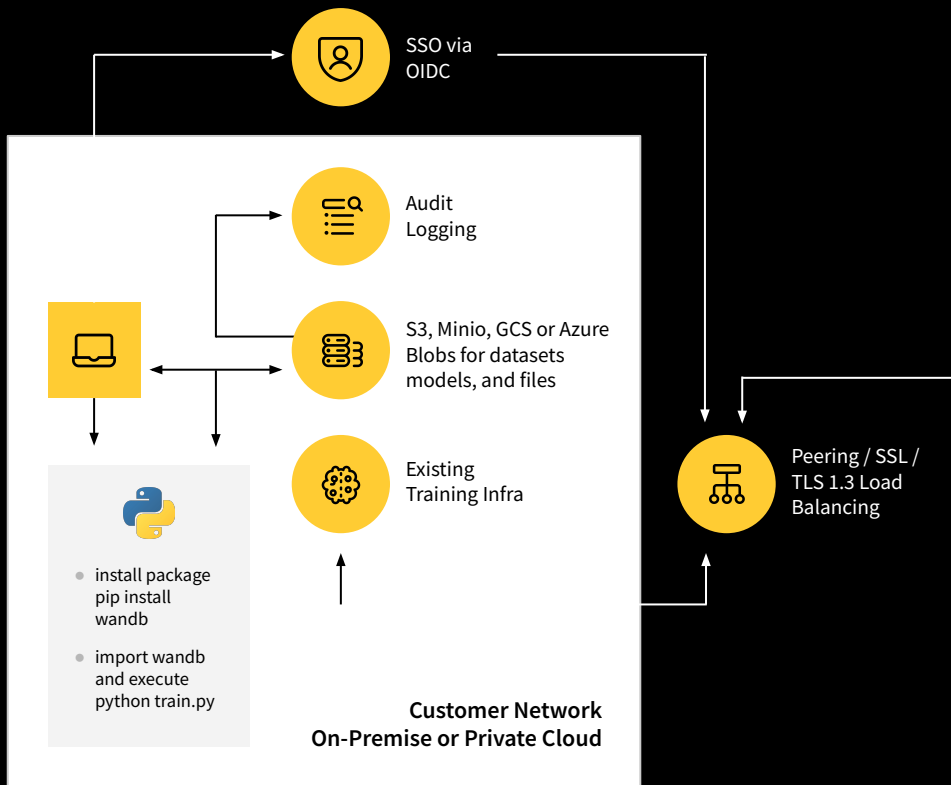
Style Direct Lineage  Include generated Artifacts

```
graph LR; R1[visionary-cloud-20] --> A1[inat_raw_data:v3]; A1 --> R2[desert-snow-21]; R2 --> A2[inat_train_data_800:v0]; R2 --> A3[inat_val_data_100:v0]; A2 --> R3[neat-hill-23]; A3 --> R3; R3 --> A4[iv3_trained:v2]; A4 --> I1[lively-waterfall-29]; A4 --> I2[fast-star-28]; A4 --> I3[fallen-field-30]; A4 --> I4[dashing-silence-32]; A4 --> I5[crisp-tree-27]; A4 --> I6[apicot-dust-31]; A4 --> I7[expert-fog-24]; A4 --> I8[golden-chape-25]; A4 --> I9[robust-fire-26];
```

# Work well with enterprise infrastructure



**Enterprise**  
Enterprise-grade scalability  
& security for production ML  
at scale



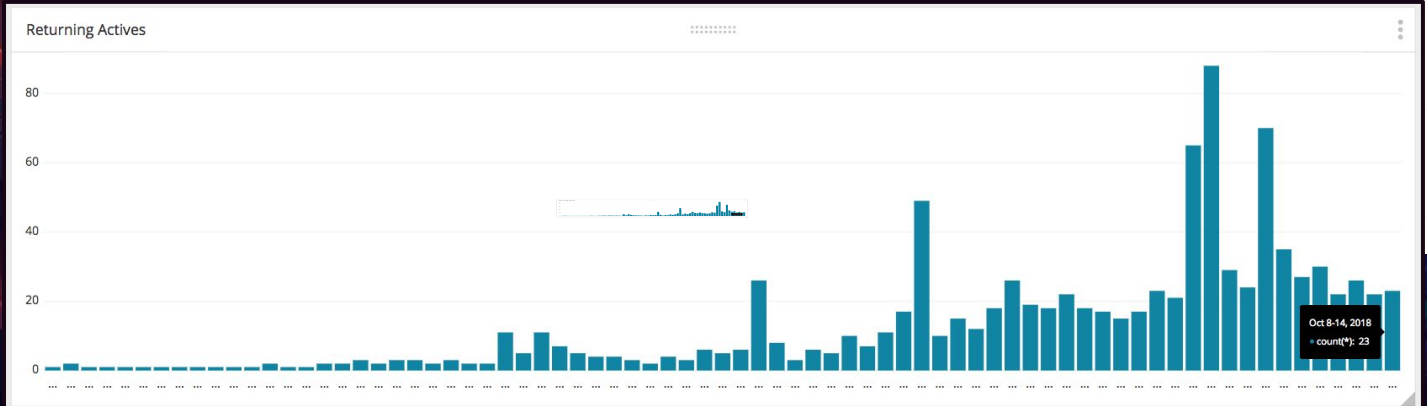
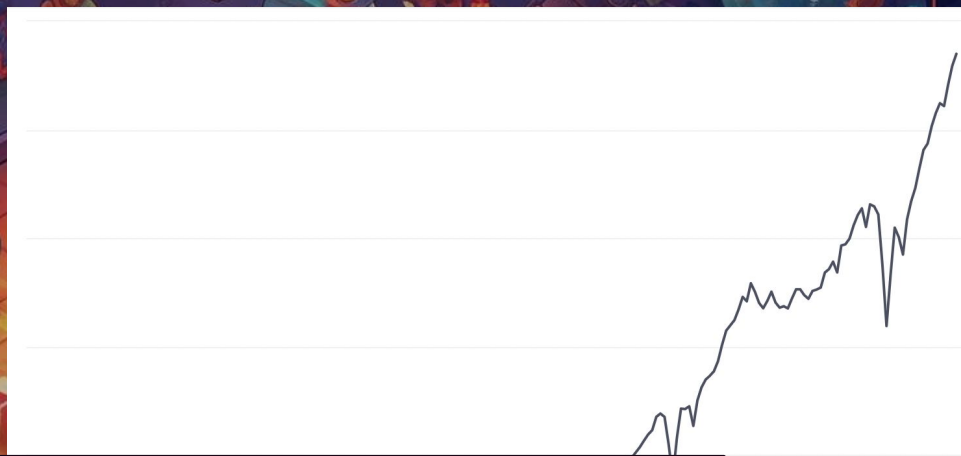
**Region chosen by customer**

- MySQL for Metadata
- Audit Logging
- wandb/local Docker image with K8s

All data Isolated and encrypted in a W&B managed private cloud account. W&B SRE staff is on call 24/7 for incident response and upgrades. 99.9% Service SLA.

aws Microsoft Azure Google Cloud

# User Growth



Jul 1 2022 Jan 1 2023



# Weights & Biases Customers

## INNOVATORS / INFLUENCERS / AI COMPANIES



## HIGH PERFORMANCE COMPUTING



## GAMING



## HEALTHCARE/LIFE SCIENCES



## AUTONOMOUS VEHICLES/ROBOTICS



JOHN DEERE



WAYVE



## FINANCIAL SERVICES



## MEDIA



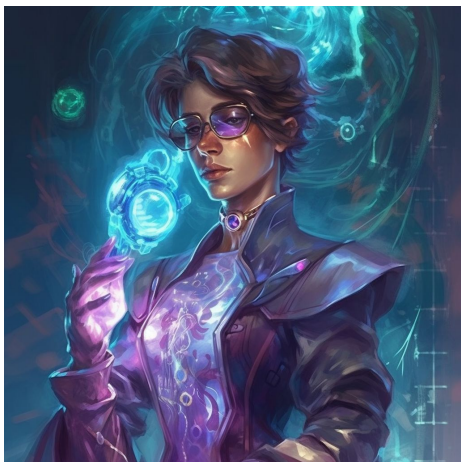
## RETAIL/CONSUMER



## HIGH TECH / SOFTWARE



# LLM Personas



*llm creator*

Trains LLMs from scratch



*llm fine-tuner*

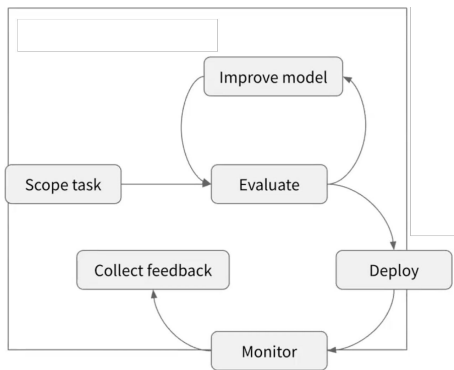
Adapts existing LLMs  
to specific tasks



*llm user*

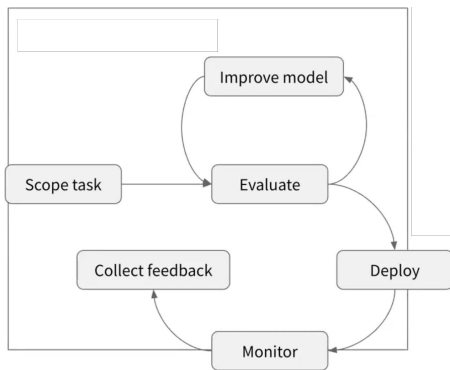
Uses LLM as API service  
("Prompt Engineer")

# Three personas, one workflow



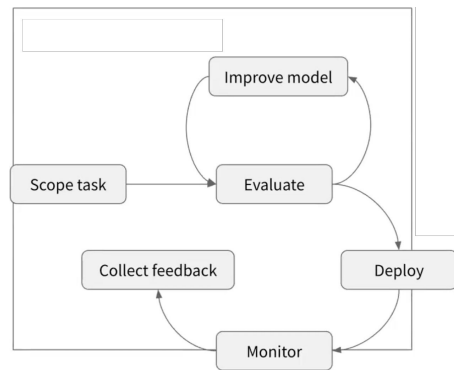
*llm creator*

Trains LLMs from scratch



*llm fine-tuner*

Adapts existing LLMs  
to specific tasks



*llm user*

Uses LLM as API service  
("Prompt Engineer")



# Building Models



# Using Models





**Otto Bot**



Alexa, play my favorite song





# Mycroft

MycroftAI / mycroft-core

Search Type to search

Code Issues 164 Pull requests 38 Actions Projects 1 Security Insights

mycroft-core Public

Watch 294 Fork 1.3k Star 6.4k

dev 69 branches 134 tags

Go to file Add file Code

forslund Handle multiple intents with the same name (#2921) ... ab242a2 on Nov 22, 2022 4,921 commits

.github	Shellcheck scripts in root folder (#3090)	last year
bin	Silence file does not exist error when clearing VK files (#3093)	last year
doc	Merge pull request #2898 from MycroftAI/bugfix/docstring-cleanup	2 years ago
mycroft	Handle multiple intents with the same name (#2921)	last year
requirements	Relax pyyaml version requirement	last year
scripts	Shellcheck scripts folder (#3063)	last year
test	Handle multiple intents with the same name (#2921)	last year
.gitignore	Add .theia to .gitignore	3 years ago
.gitmessage	Add suggestion and examples of imperative text	6 years ago
.pep8speaks.yml	Set pep8-speaks line length limit to 79 (#1953)	5 years ago
.readthedocs.yml	Fix issues with .readthedocs file	3 years ago
.shellcheckrc	Disable source lookup errors	2 years ago

About

Mycroft Core, the Mycroft Artificial Intelligence platform.

mycroft.ai

python nlp linux open-source

raspberrypi natural-language-processing

oss ai skills arch fedora

natural-language artificial-intelligence

personal-assistant hacktoberfest

voice-assistant voice-as-an-interface

Readme

Apache-2.0 license

Code of conduct

Activity

6.4k stars

294 watching

1.3k forks

Report repository

# Useful Sensors/RockPI

PRODUCTS

## See for yourself.

AI IN A BOX

PERSON SENSOR

TINY CODE READER

Will be delivered in a case with an integrated display, speaker, microphone and connectors



Rendered for illustration purposes only, final design may vary



**\$200 Device Runs llama and whisper model in real time**



...



# Extensible Library of Skills

```
def get_weather(self, city_name):  
    print("Getting weather for " + city_name)  
    params = {"q": city_name,  
             "appid": self.api_key, "units": "imperial"}  
    base_url = "http://api.openweathermap.org/data/2.5/weather"  
  
    response = requests.get(base_url, params=params)
```

Add your  
own

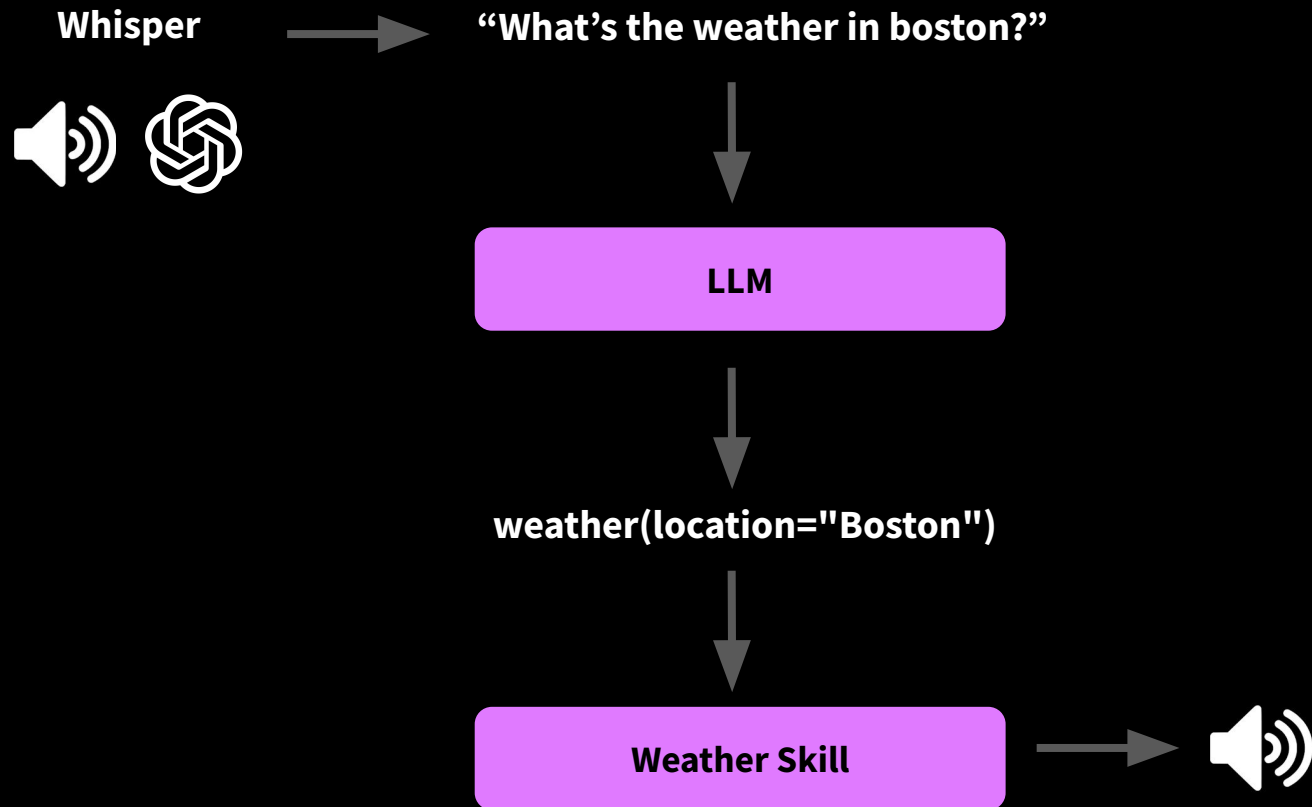
weather

math

openai

news

# How does it work



# LLaMA and Whisper: The Open Source stack

**Whisper** is an **open source model** for audio transcription by OpenAI:

- Model checkpoints in PyTorch
- Inference code in PyTorch
- Requires GPU

**LLaMA** is an **open source LLM** by MetaAI:

- Model checkpoints in PyTorch and HF formats
- Inference code in PyTorch or dedicated server engines (TGI, vLLM, etc)
- Requires GPU to serve



# Llama.cpp: Run LLaMA on a MacBook

Georgi Gerganov ported Whisper and LLaMA to C++, with no dependencies!

- Now you can run it on your laptop
- It has a built-in server
- Quantization and other useful techniques like speculative decoding
- It's fast!

whisper<sup>C++</sup>

LLaMA<sup>C++</sup>

## Settings

## .logs

## Listening Transcript

[\_BEG\_] [BLANK\_AUDIO] [\_TT\_500]

[\_BEG\_] . [\_TT\_300]

[\_BEG\_] [typing] [\_TT\_157]

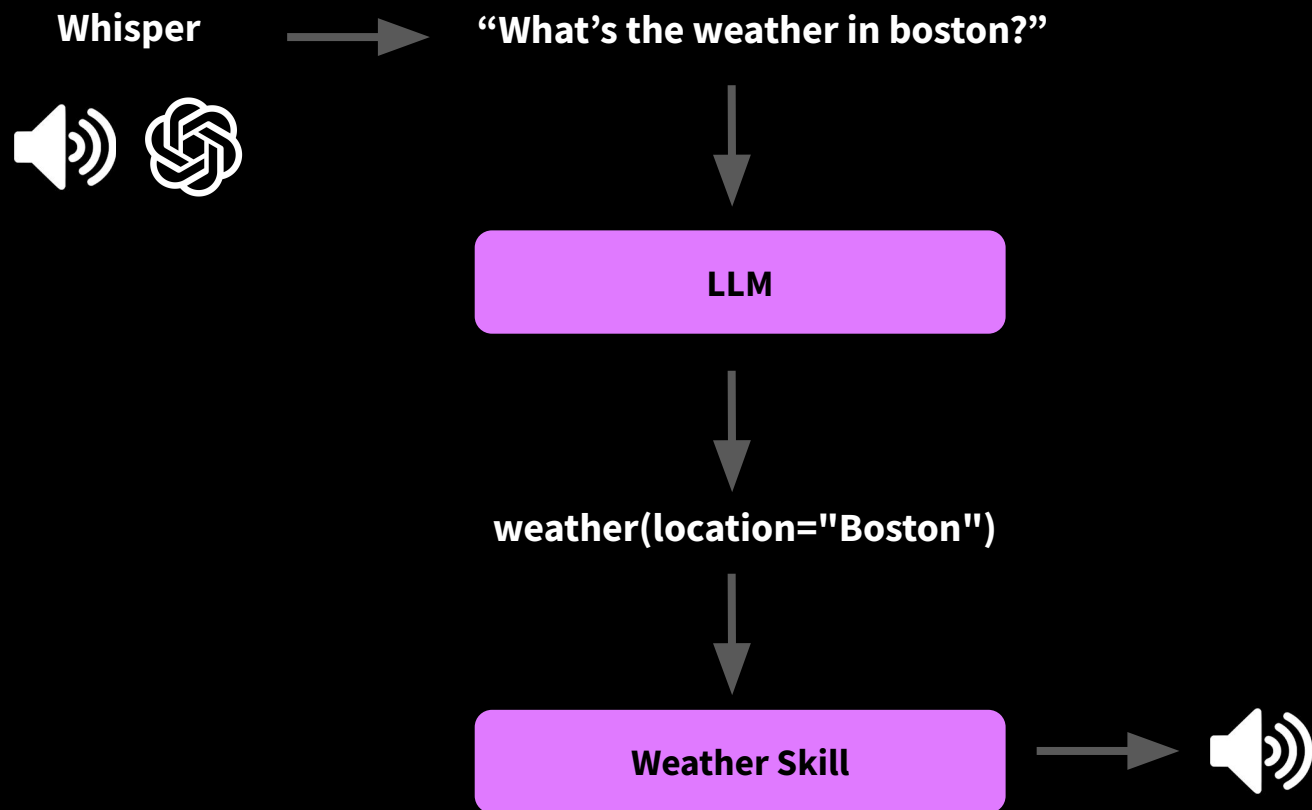
[\_BEG\_] [BLANK\_AUDIO] [\_TT\_500]

[\_BEG\_] (footsteps) [\_TT\_116] <|endoftext|>

[\_BEG\_] [clicking] [\_TT\_155]

## System Logs

# Latency Matters a Lot





# Llama doesn't work with the default prompt

USER

Below is an instruction that describes a task. Write a response that appropriately completes the request.

### User: What's the weather in boston?

### Answer:



AI

I'm happy to help you with that! Unfortunately, I'm a large language model, I don't have access to real-time weather information or current conditions in specific locations. However, I can suggest some ways for you to find out the current weather in Boston.

1. Check online weather websites: There are many websites that provide up-to-date weather information, such as AccuWeather, Weather.com, or the National Weather Service (NWS). You can enter "Boston" in the search bar and get the current weather conditions and forecast.
2. Use a mobile app: There are many mobile apps available that provide real-time weather information, such as Dark Sky (iOS, Android) or Weather Underground (Android, iOS). These apps often have features like radar and satellite imagery to help you track the weather.
3. Tune into local news: You can watch local news broadcasts or check their website for weather updates. Local news stations often provide real-time coverage of weather events in their area. Remember, it's important to stay informed about the weather and take necessary precautions to ensure your safety.



0%  
acc

# Let's make the prompt better

Tell it what functions to call and how

```
You are AI that converts human request to function calls.
You have a set of functions available to use:
-news(topic="[topic]") asks for latest headlines about a
topic.
-math(question="[question]") asks a math question in
python format.
...
-weather(location="[location]") ask for the weather at a
location. If there's no location string the location is
assumed to be where the user is.
-other() should be used when none of the other commands
apply
```

Give clear instructions and the expected output

```
Reply with the corresponding function call only
USER_QUERY: What's the weather in boston?
Assistant: weather(location="Boston")
```

# The model gets closer but...

0%  
acc

You are AI that converts human request function calls.

You have a set of functions available to use:

- news(topic="[topic]") asks for latest headlines about a topic.
- math(question="[question]") asks a math question in python format.
- notes(action="add|list", note="[note]") lets a user take simple notes.
- openai(prompt="[prompt]") asks openai a question.
- runapp(program="[program]") runs a program locally.
- story(description=[description]) lets a user ask for a story.
- timecheck(location="[location]") ask for the time at a location. If no location is given it's assumed to be the current location.
- timer(duration="[duration]") sets a timer for duration written out as a string.
- weather(location="[location]") ask for the weather at a location. If there's no location string the location is assumed to be where the user is.
- other() should be used when none of the other commands apply.

Convert the following user query to the corresponding function call:

USER\_QUERY: What's the weather in boston?



USER

```
CALL: weather(location="boston")  
```
```



AI

### Deployed Project Link

[click here](https://any-yang-ai.herokuapp.com) to access the deployed version of the project!

the correct answer

hallucination



# Let's switch to Llama2-chat

“Our fine-tuned LLMs, called Llama-2-Chat, are optimized for dialogue use cases. Llama-2-Chat models outperform open-source chat models on most benchmarks we tested, and in our human evaluations for helpfulness and safety, are on par with some popular closed-source models like ChatGPT and PaLM.”

# Llama2-chat has been trained on conversations so it performs better on instructions

```
"""<s>[INST] <<SYS>>
```

```
You are AI that converts human request into api calls.
```

```
You have a set of functions:
```

```
-news(topic="[topic]") asks for latest headlines about a topic.
```

```
-math(question="[question]") asks a math question in python format.
```

```
...
```

```
-weather(location="[location]") ask for the weather at a location. If there's no location string the location is assumed to be where the user is.
```

```
-other() should be used when none of the other commands apply
```

```
<</SYS>>
```

```
Here is a user request, reply with the corresponding function call
```

```
USER_QUERY: What's the weather in boston?[/INST]"""
```

The model has a more specific instruction format.

We give more specific instructions to the model

# We get better accuracy with LLAMA-chat the model answers but also gets the format wrong

USER

Convert the following user query to the corresponding function call:  
USER\_QUERY: What's the weather in boston?



AI

FUNCTION CALL: weather(location="Boston")



**11%**  
**acc**



# Incorporating feedback from the model improves the accuracy

```
"""<s>[INST] <<SYS>>
```

```
You are AI that converts human request into api calls.
```

```
You have a set of functions:
```

```
-news(topic="[topic]") asks for latest headlines about a topic.
```

```
-math(question="[question]") asks a math question in python format.
```

```
...
```

```
-weather(location="[location]") ask for the weather at a location. If there's no location string the location is assumed to be where the user is.
```

```
-other() should be used when none of the other commands apply
```

```
<</SYS>>
```

```
Here is a user request, reply with the corresponding function call
```

```
USER_QUERY: What's the weather in boston?[/INST]
```

```
FUCTION_CALL: """
```

**75%**  
**acc**

# LLama2-chat -> FUNCTION\_CALL

[Thomas Capelle](#)

Nothing else, just add that to the prompt...

## Results

| runs.summary["eval_predictions"] |              |        | Accuracy                             | Acc_lousy        |
|----------------------------------|--------------|--------|--------------------------------------|------------------|
| answer == generation             | prompt.count |        | moonlit-ghost-97                     | moonlit-ghost-97 |
| 1                                | False        | 15     | 0.7581                               | 0.8226           |
| 2                                | True         | 47     |                                      |                  |
| ← < 1 -2 of 2 > →                |              |        | Export as CSV Columns... Reset Table |                  |
| runs.summary["eval_predictions"] |              |        |                                      |                  |
| prompt                           | user         | answer | generation                           | answer ==        |
| <>[INST] <<SYS>>                 |              |        |                                      |                  |

Yorwe AI that converts human request

# Adding examples to the prompt reduces the accuracy

```
"""<s>[INST] <<SYS>>
```

```
You are AI that converts human request into api calls.
```

```
You have a set of functions:
```

```
-news(topic="[topic]") asks for latest headlines about a topic.
```

```
-math(question="[question]") asks a math question in python format.
```

```
""  
-weather(location="[location]") ask for the weather at a location. If there's no location string the location is assumed to be where the user is.
```

```
-other() should be used when none of the other commands apply
```

```
Some example user queries and the corresponding function call:
```

```
USER_QUERY: What is a random number under one hundred
```

```
FUNCTION_CALL: math(question="randint(100)")
```

```
USER_QUERY: Ask gpt if a leopard can swim
```

```
FUNCTION_CALL: openai(prompt="Can a leopard swim?")
```

```
USER_QUERY: So it's like...
```

```
FUNCTION_CALL: other()
```

```
USER_QUERY: to do this? Also, I've noticed that when I use
```

```
FUNCTION_CALL: other()
```

```
<</SYS>>
```

```
Here is a user request, reply with the corresponding function call
```

```
USER_QUERY: What's the weather in boston?[/INST]
```

```
FUNCTION_CALL: ""
```

51%  
acc

# Let's try MistralAI's model

**“The Mistral-7B-Instruct-v0.1 Large Language Model (LLM) is a instruct fine-tuned version of the Mistral-7B-v0.1 generative text model using a variety of publicly available conversation datasets.”**



# The same prompt with MistralAI Instruct gives us great accuracy without fine-tuning

```
"""[INST] You are AI that converts human request into api calls.
You have a set of functions:
-news(topic="[topic]") asks for latest headlines about a topic.
-math(question="[question]") asks a math question in python format.
-notes(action="add|list", note="[note]") lets a user take simple notes.
-openai(prompt="[prompt]") asks openai a question.
...
-weather(location="[location]") ask for the weather at a location. If there's no location string the location is assumed
to be where the user is.
-other() should be used when none of the other commands apply
```

Some example user queries and the corresponding function call:

```
USER_QUERY: What is a random number under one hundred
FUNCTION_CALL: math(question="randint(100)")
```

```
USER_QUERY: Ask gpt if a leopard can swim
FUNCTION_CALL: openai(prompt="Can a leopard swim?")
```

```
USER_QUERY: to do this? Also, I've noticed that when I use
FUNCTION_CALL: other()
```

Here is a user request, reply with the corresponding function call only, be brief.

```
USER_QUERY: What's the weather in boston?
FUCTION_CALL: [/INST]"""
```

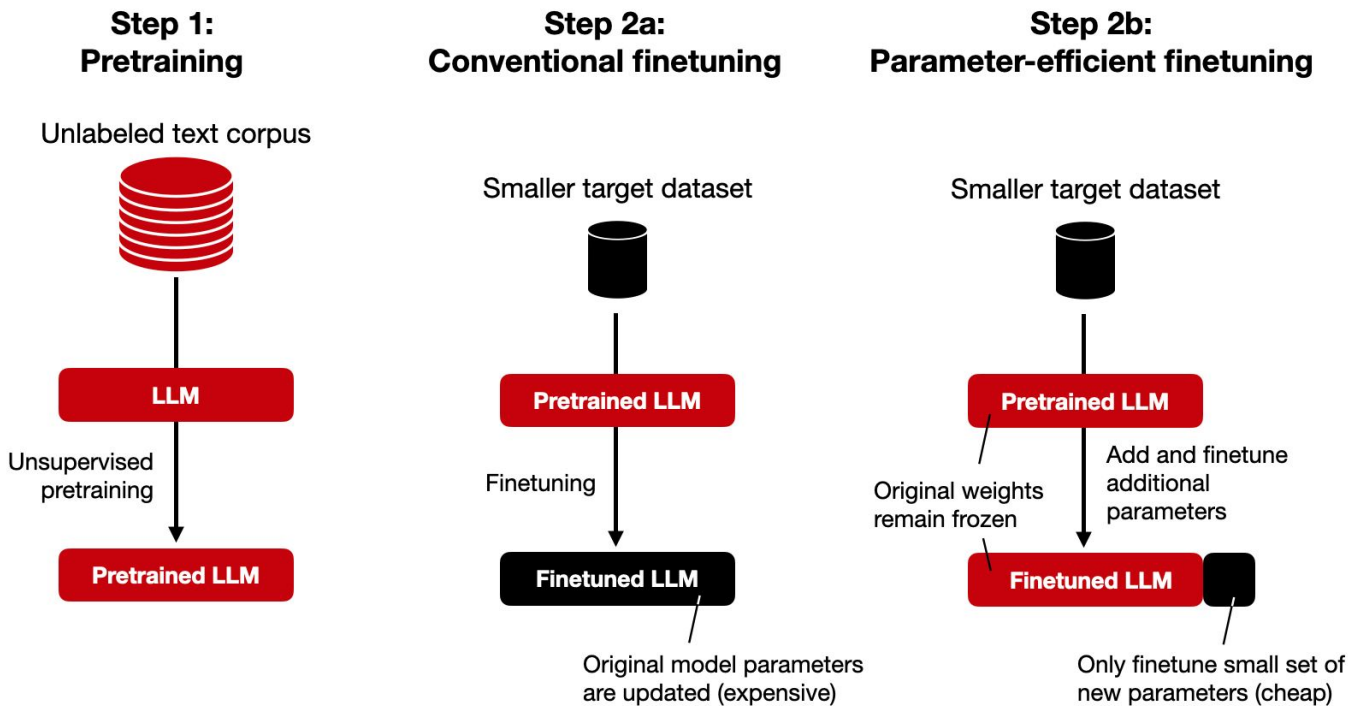
79%  
acc



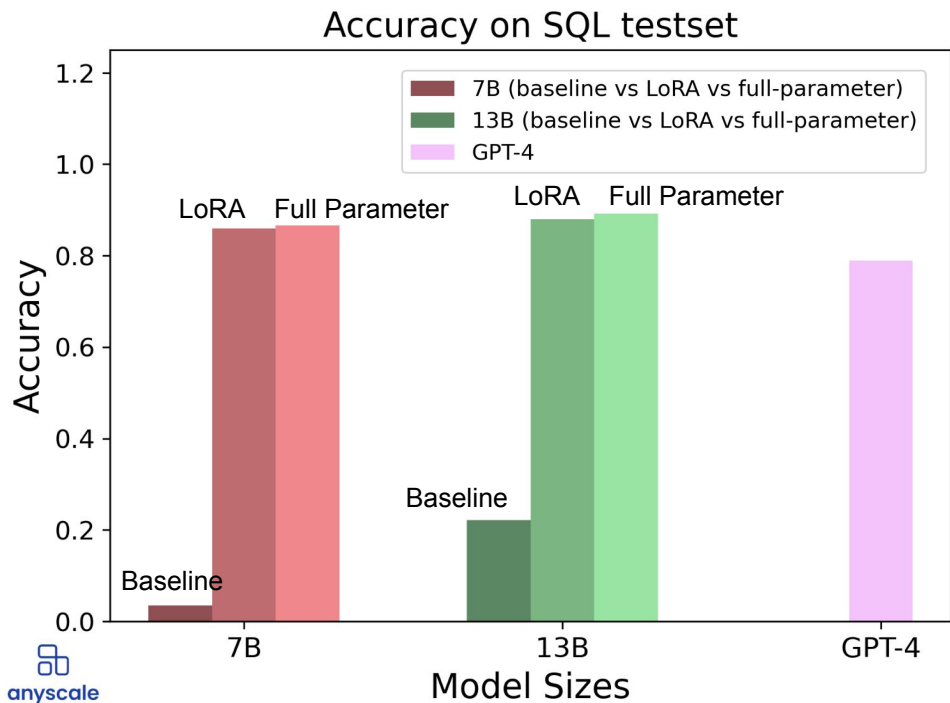
**Is a pre-trained LLM enough?**

**79% Accuracy is still pretty annoying**

# Fine-tuning



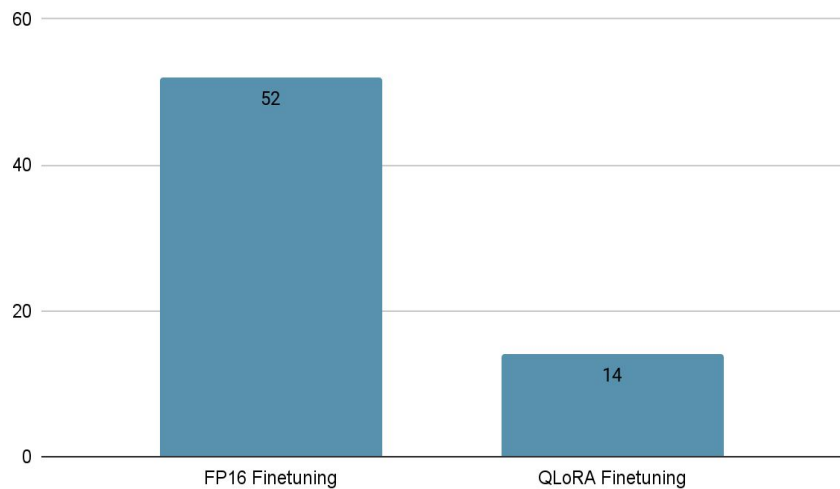
# Fine-tuning with LoRA works almost as well as full-parameter





# QLoRA allows fine-tuning large models on normal GPUs

Llama 7B GPU Memory Needs





**but you will need a dataset to fine-tune**

# Manually Creating Example

Lukas Biewald, last month | 1 author (Lukas Biewald)

```
1 ##### Prompt:
2
3 news(topic="[topic]") asks for latest headlines about a topic.
4
5 ##### Prompt Examples:
6
7 ### User: What is the latest news?
8 ### Assistant: news()
9
10 ### User: Any news about astronomy?
11 ### Assistant: news(topic="astronomy")
12
13 ### User: Whats the latest news about apple?
14 ### Assistant: news(topic="apple")
15
```

# Using ChatGPT to create more examples

I am collecting training data for a voice assistant.

The voice assistant has the command:

`news(topic="[topic]")` asks for latest headlines about a topic.

Some examples of how a user might say this command and the response is:

### User: What is the latest news?

### Assistant: `news()`

### User: Any news about astronomy?

### Assistant: `news(topic="astronomy")`

### User: Whats the latest news about apple?

### Assistant: `news(topic="apple")`

Generate an examples of a user querying this command and the correct response.

Use the following format

User:

Assistant:



# Quick win: Give OpenAI a schema to format the answer in

```
format_function = [  
  {  
    "name": "validate",  
    "description": "validates the user assistant pair",  
    "parameters": {  
      "type": "object",  
      "properties": {  
        "user": {"type": "string", "description": "The query from the user" },  
        "assistant": {"type": "string", "description": "The command to call" }}  
      },  
    "required": ["user", "assistant"],  
  },  
]
```

# Running the flywheel to generate a dataset

```
1 [{"user": "Calculate the square root of 16", "answer": "math(question=\\"sqrt(16)\")"}]
~/work/otto/finetune/training_data_f.py ten divided by two?", "answer": "math(question=\\"10/2\")"}
3 {"user": "Give me the result of three squared", "answer": "math(question=\\"3**2\")"}
4 {"user": "What is the remainder when twelve is divided by five?", "answer": "math(question=\\"12%5\")"}
5 {"user": "Calculate the expression two plus three times four", "answer": "math(question=\\"(2+3)*4\")"}
6 {"user": "What is two raised to the power of eight?", "answer": "math(question=\\"2**8\")"}
7 {"user": "Calculate the square root of sixteen", "answer": "math(question=\\"16**(1/2)\")"}
8 {"user": "What is four to the power of three?", "answer": "math(question=\\"4**3\")"}
9 {"user": "Calculate the average of two, four, and six", "answer": "math(question=\\"(2+4+6)/3\")"}
10 {"user": "What is nine divided by three plus two?", "answer": "math(question=\\"9/3+2\")"}
11 {"user": "Evaluate the expression six minus three multiplied by two", "answer": "math(question=\\"6-(3*2)\"}
12 {"user": "What is the remainder when twelve is divided by five?", "answer": "math(question=\\"12%5\")"}
13 {"user": "What's three times five?", "answer": "math(question=\\"3*5\")"}
14 {"user": "Can you give me the most recent news?", "answer": "news()"}
15 {"user": "Please search for news about technology.", "answer": "news(topic=\\"technology\")"}
16 {"user": "I need updates on politics.", "answer": "news(topic=\\"politics\")"}
17 {"user": "Is there any news about sports today?", "answer": "news(topic=\\"sports\")"}
18 {"user": "What are the latest headlines about entertainment?", "answer": "news(topic=\\"entertainment\")"}
19 {"user": "Could you find news articles about health?", "answer": "news(topic=\\"health\")"}
20 {"user": "Can you provide me with news about fashion?", "answer": "news(topic=\\"fashion\")"}]
```

# Fine-Tuning + Mistral leads to amazing accuracy

runs.summary["eval\_predictions"]

|    | prompt                | user                    | answer                            | generation                        | answer == generation |
|----|-----------------------|-------------------------|-----------------------------------|-----------------------------------|----------------------|
| 31 | [INST]You are Al that | I need updates on       | news(topic="politics")            | news(topic="politics")            | True                 |
| 32 | [INST]You are Al that | And keep your           | other()                           | other()                           | True                 |
| 33 | [INST]You are Al that | I can go to the video   | other()                           | other()                           | True                 |
| 34 | [INST]You are Al that | which causes the        | other()                           | other()                           | True                 |
| 35 | [INST]You are Al that | Do you have the         | other()                           | other()                           | True                 |
| 36 | [INST]You are Al that | Okay, so I need to go   | other()                           | other()                           | True                 |
| 37 | [INST]You are Al that | See you in the next     | other()                           | other()                           | True                 |
| 38 | [INST]You are Al that | What's the weather like | weather(location="San Francisco") | weather(location="San Francisco") | True                 |
| 39 | [INST]You are Al that | Can you provide me      | news(topic="fashion")             | news(topic="fashion")             | True                 |
| 40 | [INST]You are Al that | What are the latest     | news(topic="entertainment")       | news(topic="entertainment")       | True                 |
| 41 | [INST]You are Al that | What's the latest news  | news(topic="entertainment")       | news(topic="entertainment")       | True                 |
| 42 | [INST]You are Al that | Just...                 | other()                           | other()                           | True                 |



← < 31 - 42 of 62 > →

Export as CSV Columns

98%  
acc

# Also works for other languages!

Examples:

### User: Set a timer for five minutes

### Assistant: timer(duration="5 minutes")

### User: set a timer for one hour and three minutes

### Assistant: timer(duration="1 hour 3 minutes")

### User: what time is it?

### Assistant: timecheck()

### User: what time is it in Tokyo?

### Assistant: timecheck(location="Tokyo")

### User: what is the weather?

### Assistant: weather()

### User: what is the weather in Boston?

### Assistant: weather(location="Boston")

Quelle est la météo à Paris?

CHAT STOP RESET

Logs

Chat

**Human** Quelle est la météo à Paris?

**Bot** weather(location="Paris")



lukas / otto

Code Issues Pull requests 1 Actions Projects Security Insights

otto Public Watch 2 Fork 2 Starred 3

main 3 branches 0 tags Go to file Add file Code

lukas Merge pull request #6 from tcapella/data 33641d1 5 days ago 102 commits

|                  |                                          |              |
|------------------|------------------------------------------|--------------|
| examples         | make calls more diverse                  | 5 days ago   |
| finetune         | Mistral-instruct                         | 5 days ago   |
| frontend         | fixes typos                              | last week    |
| server           | Merge pull request #3 from tcapella/main | last week    |
| settings         | refactors front end and back end         | last month   |
| weave            | refactors front end and back end         | last month   |
| weaveflow        | makes easy to fun without cuda           | 2 months ago |
| .gitignore       | add logs and frontend folder             | last week    |
| README.md        | add demo video                           | last week    |
| requirements.txt | Makes it easier to install               | 2 months ago |

README.md

otto

About  
No description, website, or topics provided.

- Readme
- Activity
- 3 stars
- 2 watching
- 2 forks

Report repository

Releases  
No releases published

Packages  
No packages published

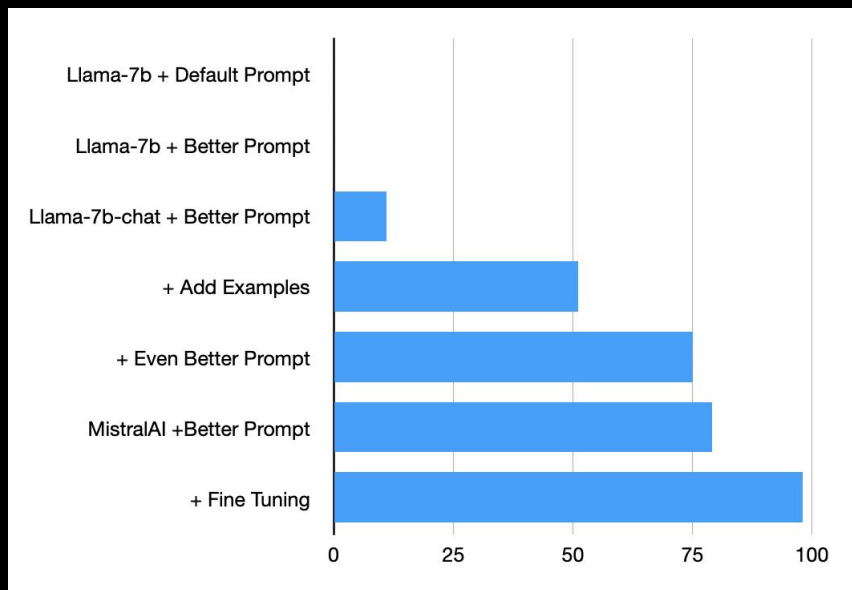
Contributors 4

lukas Lukas Biewald

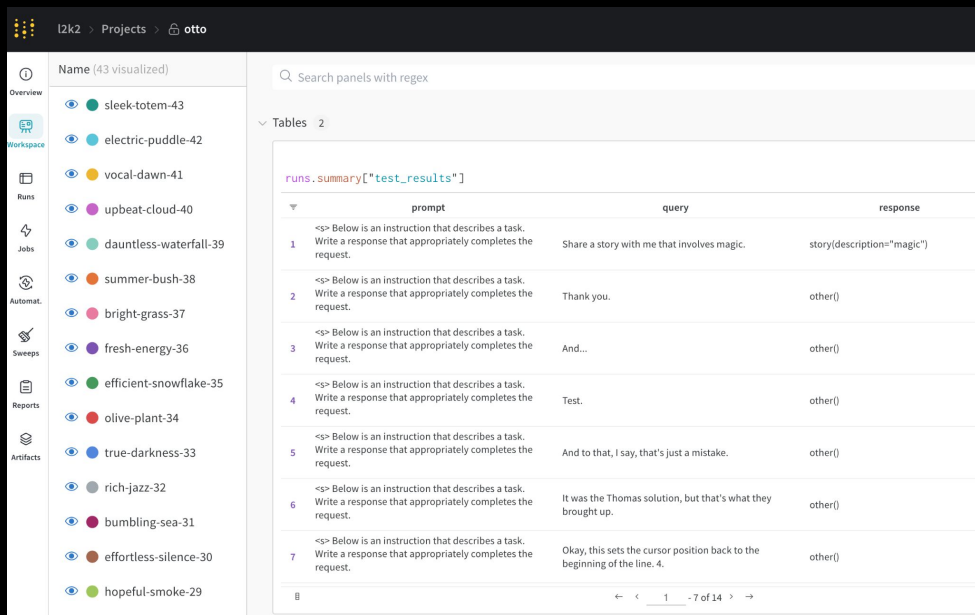
github.com/lukas/otto

# Lessons learned and next steps

- Prompt engineering improves performance a lot
- Fine-tuning improves performance a lot



# Most of the experiments didn't work



The screenshot shows the WandB interface for a project named 'otto'. The left sidebar lists 15 projects, each with a colored dot and an eye icon. The main area displays a table of test results for the selected project. The table has three columns: 'prompt', 'query', and 'response'. The data is organized into 7 rows, each representing a different test case. The 'prompt' column contains instructions for writing responses to specific queries. The 'query' column shows the input text, and the 'response' column shows the output generated by the model. The table is titled 'runs.summary["test\_results"]' and includes a search bar at the top.

|   | prompt                                                                                                        | query                                                                     | response                   |
|---|---------------------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------|----------------------------|
| 1 | <-> Below is an instruction that describes a task. Write a response that appropriately completes the request. | Share a story with me that involves magic.                                | story(description="magic") |
| 2 | <-> Below is an instruction that describes a task. Write a response that appropriately completes the request. | Thank you.                                                                | other()                    |
| 3 | <-> Below is an instruction that describes a task. Write a response that appropriately completes the request. | And...                                                                    | other()                    |
| 4 | <-> Below is an instruction that describes a task. Write a response that appropriately completes the request. | Test.                                                                     | other()                    |
| 5 | <-> Below is an instruction that describes a task. Write a response that appropriately completes the request. | And to that, I say, that's just a mistake.                                | other()                    |
| 6 | <-> Below is an instruction that describes a task. Write a response that appropriately completes the request. | It was the Thomas solution, but that's what they brought up.              | other()                    |
| 7 | <-> Below is an instruction that describes a task. Write a response that appropriately completes the request. | Okay, this sets the cursor position back to the beginning of the line. 4. | other()                    |

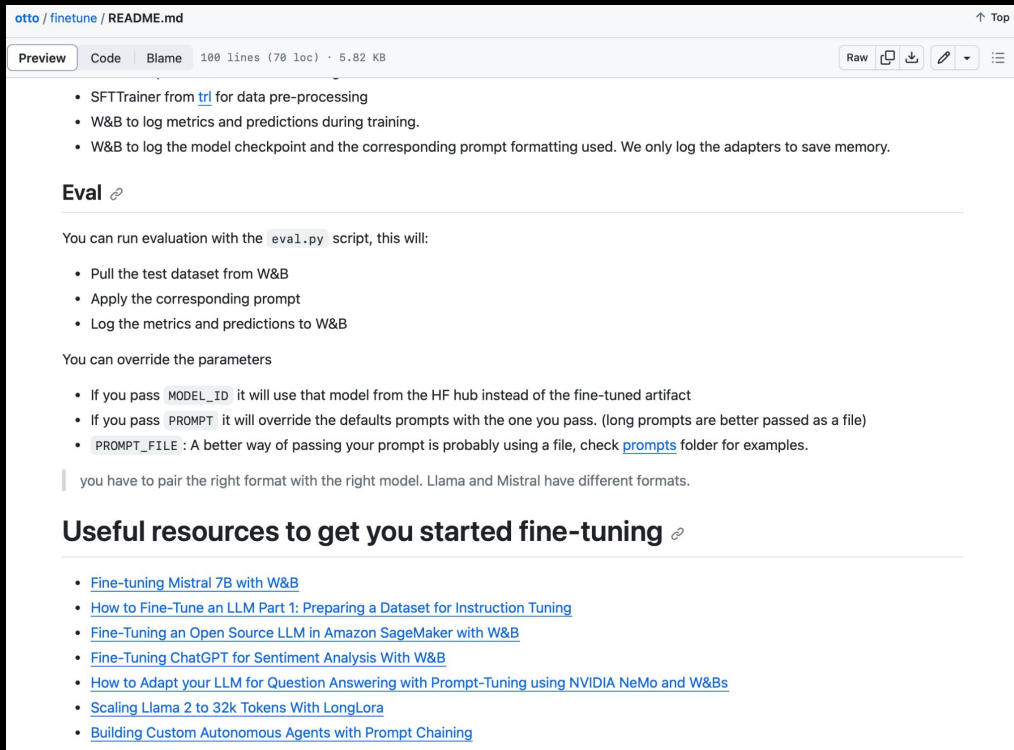
project:

<https://wandb.ai/capecape/otto>

reports:

<https://wandb.ai/capecape/otto/reportlist>

# Learn More



The screenshot shows a GitHub repository page for 'otto / finetune / README.md'. The page has a dark theme and includes a navigation bar with 'Preview', 'Code', and 'Blame' tabs. The 'Preview' tab is active, showing the README content. The content includes a list of bullet points, a section titled 'Eval', and a section titled 'Useful resources to get you started fine-tuning'.

otto / finetune / README.md ↑ Top

Preview Code Blame 100 lines (70 loc) · 5.82 KB Raw Copy Download Edit More

- SFTTrainer from [trl](#) for data pre-processing
- W&B to log metrics and predictions during training.
- W&B to log the model checkpoint and the corresponding prompt formatting used. We only log the adapters to save memory.

## Eval [↗](#)

You can run evaluation with the `eval.py` script, this will:

- Pull the test dataset from W&B
- Apply the corresponding prompt
- Log the metrics and predictions to W&B

You can override the parameters

- If you pass `MODEL_ID` it will use that model from the HF hub instead of the fine-tuned artifact
- If you pass `PROMPT` it will override the defaults prompts with the one you pass. (long prompts are better passed as a file)
- `PROMPT_FILE`: A better way of passing your prompt is probably using a file, check [prompts](#) folder for examples.

you have to pair the right format with the right model. Llama and Mistral have different formats.

## Useful resources to get you started fine-tuning [↗](#)

- [Fine-tuning Mistral 7B with W&B](#)
- [How to Fine-Tune an LLM Part 1: Preparing a Dataset for Instruction Tuning](#)
- [Fine-Tuning an Open Source LLM in Amazon SageMaker with W&B](#)
- [Fine-Tuning ChatGPT for Sentiment Analysis With W&B](#)
- [How to Adapt your LLM for Question Answering with Prompt-Tuning using NVIDIA NeMo and W&Bs](#)
- [Scaling Llama 2 to 32k Tokens With LongLora](#)
- [Building Custom Autonomous Agents with Prompt Chaining](#)





**WandBot**

AShots Today at 12:01 PM

@wandbot (beta) How can I log my input and outputs from OpenAI



wandbot (beta) BOT Today at 12:01 PM

Hi @AShots: Please note that **wandbot is currently in alpha testing** and will experience frequent updates.

Please do not share any private or sensitive information in your query at this time.

Please note that overly long messages (>1024 words) will be truncated!

Generating response...

You can log your input and output data from OpenAI using Weights & Biases by following these steps:

1. Install `wandb` and `openai` libraries:

```
!pip install wandb openai -qU
```

2. Import and call `autoLog`:

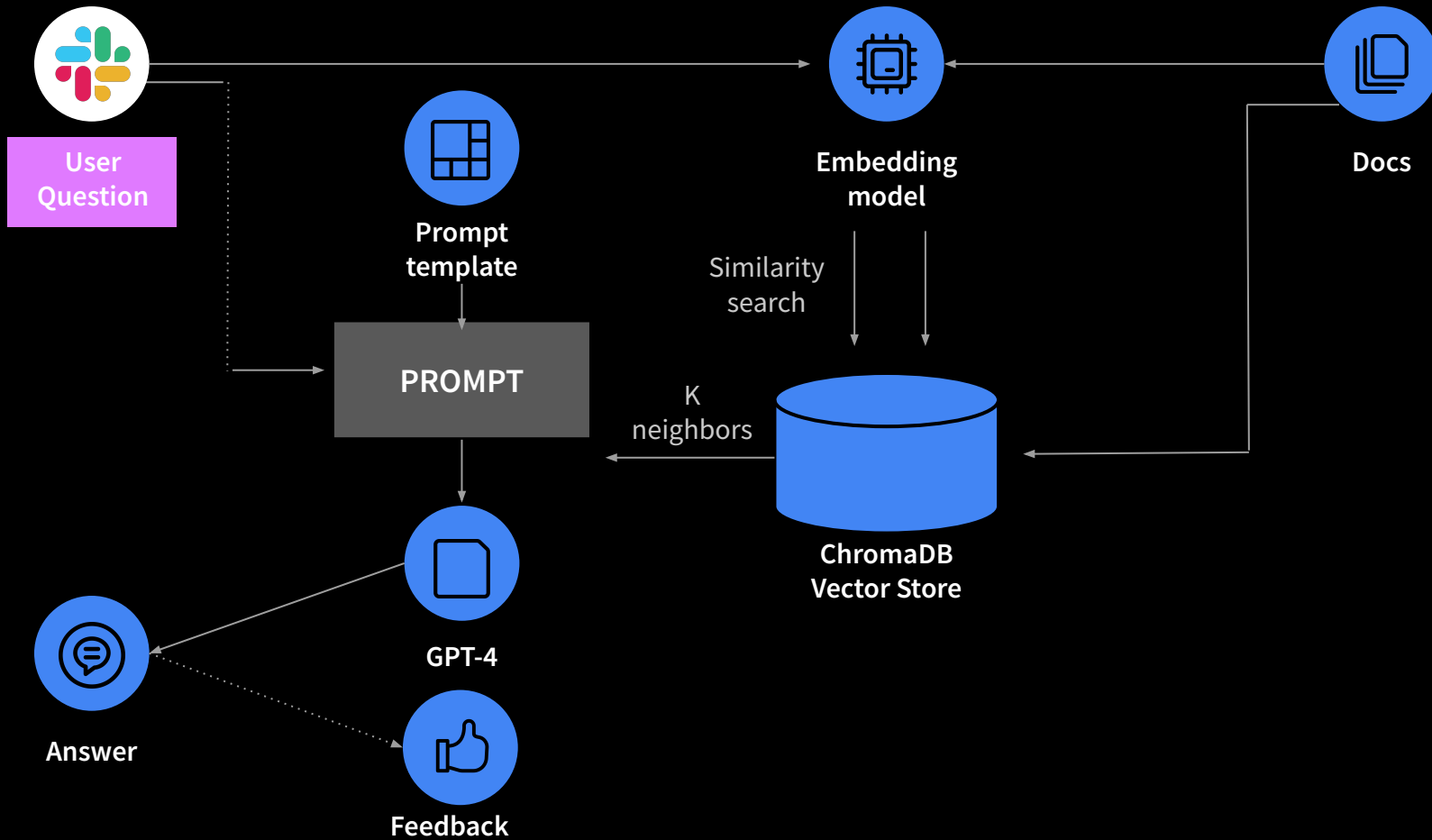
```
import openai
from wandb.integration.openai import autoLog

autoLog({"project": "my_llm_project"})
```

3. Use the OpenAI API as normal:

```
# pass your OpenAI key
import openai
openai.api_key = "sk-foo"

# make some calls to OpenAI
chat_request_kwargs = dict(
```





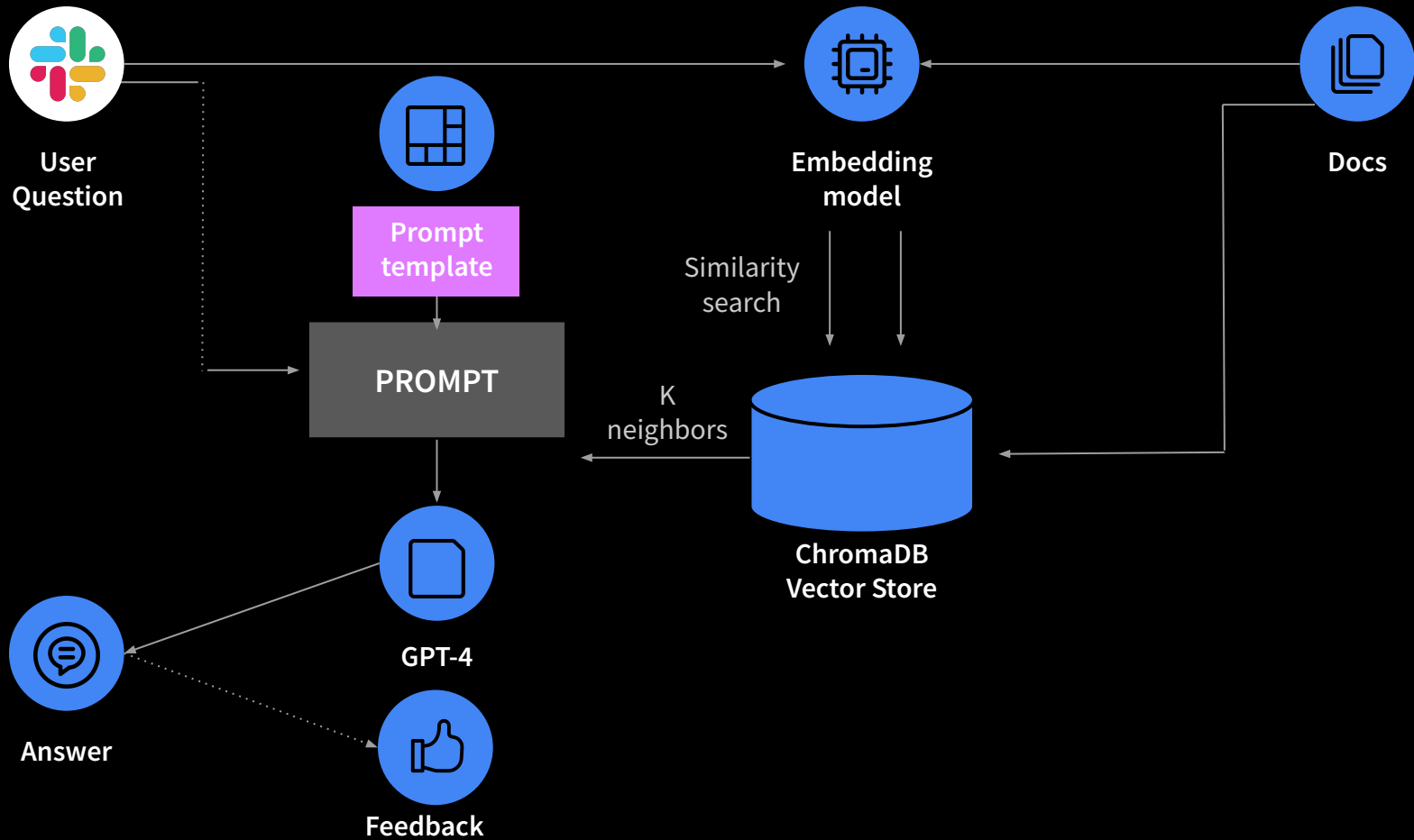
# User Question



**AShots** Today at 12:01 PM

**@wandbot (beta)** How can I log my input and outputs from OpenAI





# We started with a basic prompt template - 45% accuracy

Please answer the user's question about the Weights & Biases, W&B or wandb. Provide a detailed code example with explanations whenever possible. If the question is not related to Weights & Biases or you don't know the answer, just say that you don't know, don't try to make up an answer."

Begin

=====

Question:



**AShots** Today at 12:01 PM

[@wandbot \(beta\)](#) How can I log my input and outputs from OpenAI

Answer:

**45%**  
**acc**

# Improved prompt templates

## Setting guidelines on behavior in prompts improves accuracy

```
You are wandbot, a developer assistant designed to guide users with tasks related to Weight & Biases, its sdk `wandb` and its visualization library `weave`. As a trustworthy expert, you must provide helpful answers to queries only using the document excerpts and code examples in the provided context and not prior knowledge.
```

```
Here are your guidelines:
```

```
1. Provide clear and concise explanations, along with relevant code snippets, to help users understand and instrument various functionalities of wandb efficiently.
```

```
<!--start-example2-->  
<!--start-relevant-documents-->  
...  
<!--end-relevant-documents-->  
<!--Start-Question-->  
...  
<!--Final Answer in Markdown-->  
...  
<!--end-example2-->
```

Question:



AShots Today at 12:01 PM

Answer:

@wandbot (beta) How can I log my input and outputs from OpenAI

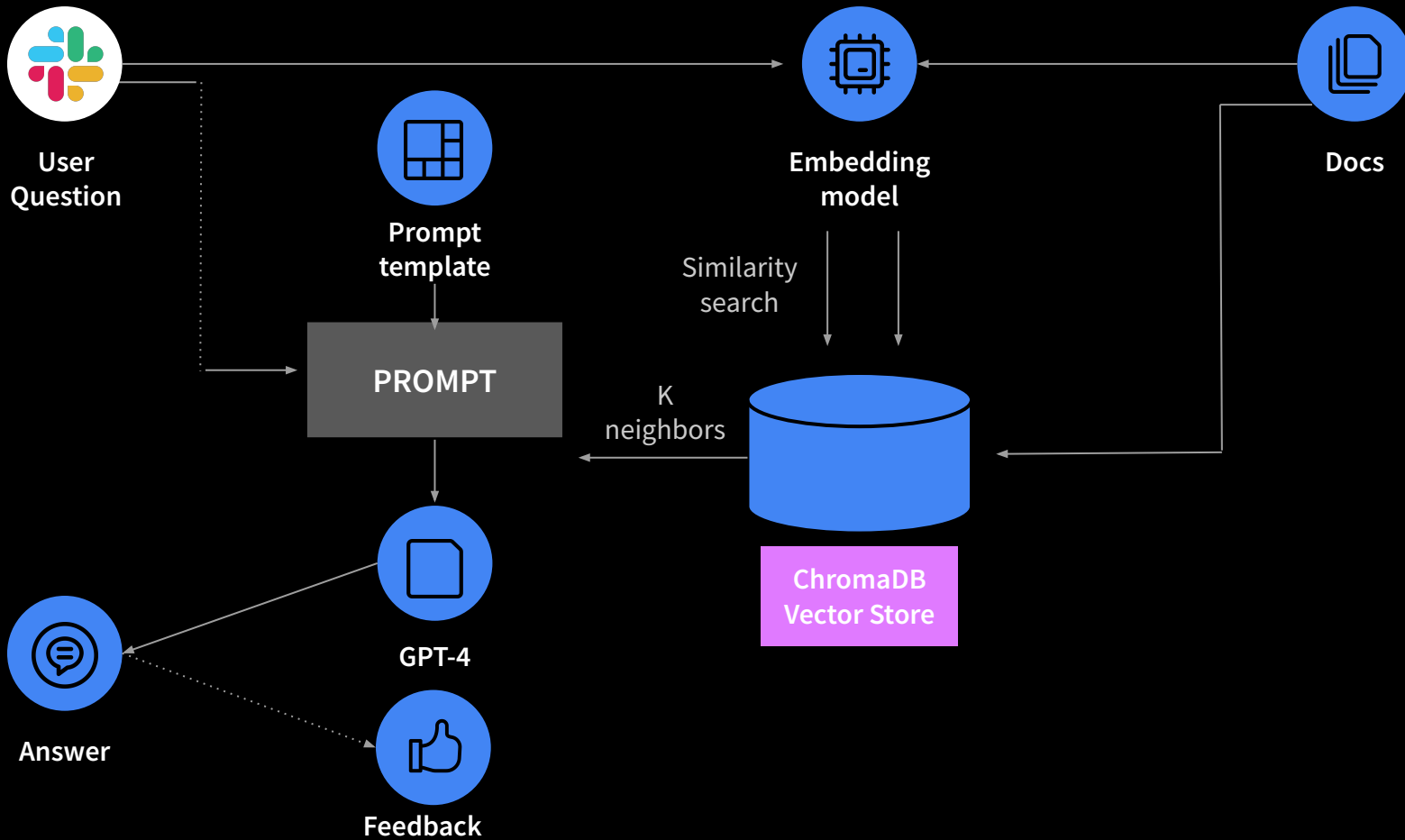
SYSTEM MESSAGE

RULES

REPRESENTATIVE  
EXAMPLES

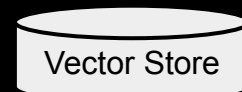
RELEVANT CONTEXT

61%  
acc





# Find the most relevant documents



How can I log my input and outputs from OpenAI?



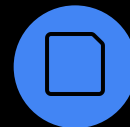
LLM

-1 3 5 -5 2 0

## 1. Grab relevant documents

### W&B Docs:

- How to log to W&B?
- W&B x OpenAI Integration
- Prompts Quickstart



LLM

0 3 3 -4 6 5

## 2. Find the most relevant one

Similarity Score

# Add documents into the prompt

```
<!--Begin-->
```

```
<!--start-relevant-documents-->
```

## How to log to Weights & Biases?

Log a dictionary of data to the current run's history.

```
log(  
  data: Dict[str, Any],  
  ...  
)
```

## W&B x OpenAI Integration

1. Import autolog and initialize it

First, import autolog from `wandb.integration.openai` and initialise it.

```
import os  
...
```

## Prompts Quickstart

2. Log to a Trace

Now we will query OpenAI times and log the results to a W&B Trace.

```
...
```

```
<!--end-relevant-documents-->
```

```
<!--Start-Question-->
```



**AShots** Today at 12:01 PM

@wandbot (beta) How can I log my input and outputs from OpenAI

```
<!--Final Answer in Markdown-->
```

# Feed the prompt into GPT-4

How can I log my input and outputs from OpenAI?



## W&B Docs:

How to log to Weights & Biases?

```
<!--start-relevant-documents-->  
<!--end-relevant-documents-->  
<!--Start-Question-->
```

Question

Relevant Docs

Prompt Template

GPT-4

# GPT-4 responds to the user



wandbot (beta) **BOT** Today at 12:01 PM

Hi @AShots: Please note that **wandbot is currently in alpha testing** and will experience frequent updates.

Please do not share any private or sensitive information in your query at this time.

Please note that overly long messages (>1024 words) will be truncated!

Generating response...

You can log your input and output data from OpenAI using Weights & Biases by following these steps:

1. Install `wandb` and `openai` libraries:

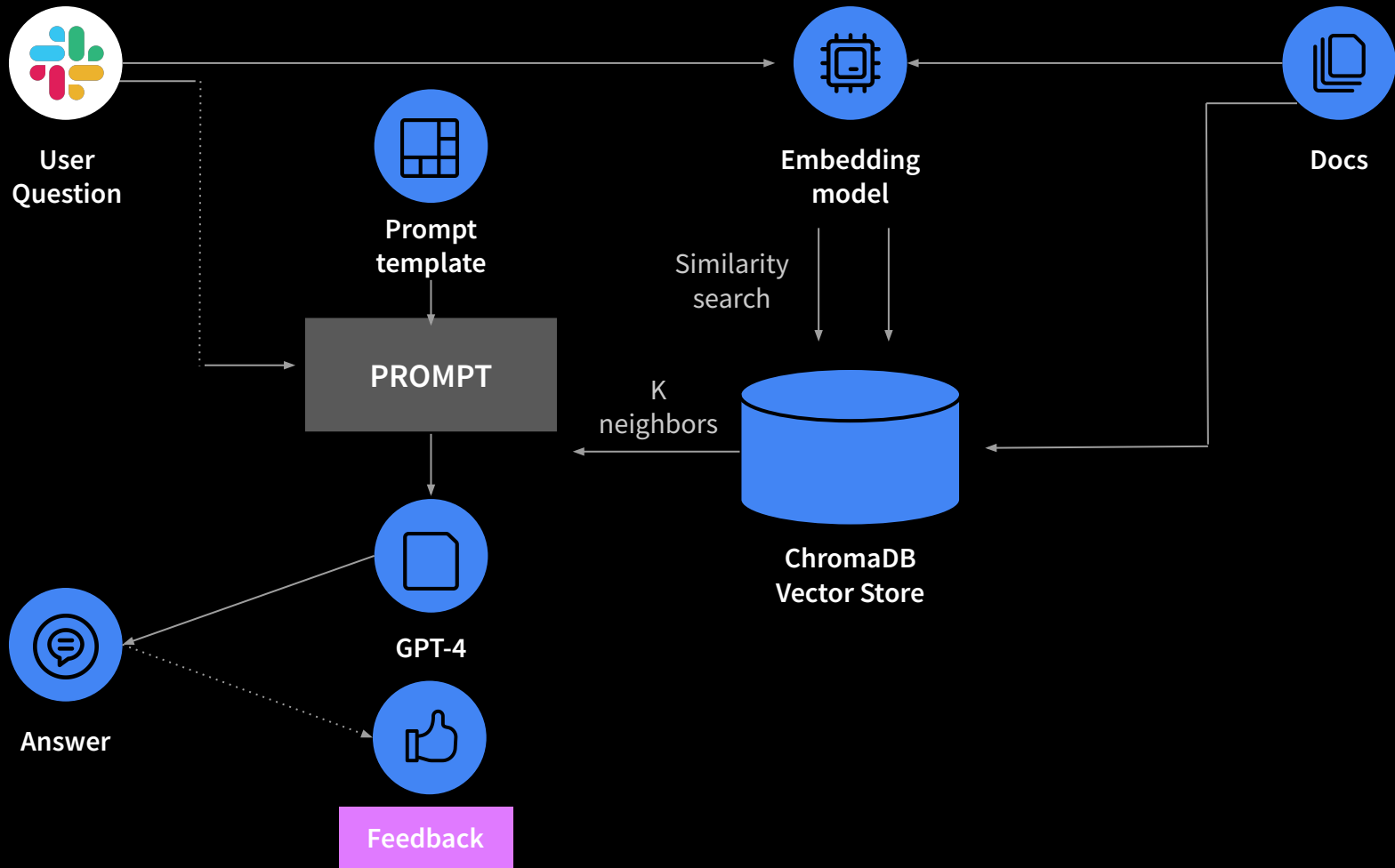
```
!pip install wandb openai -qU
```

2. Import and call `autoLog`:



```
import openai
from wandb.integration.openai
import autoLog

autoLog({"project": "my_llm_project
"})
```





# User Feedback

| Feedback                                                                          | Count |
|-----------------------------------------------------------------------------------|-------|
|  | 21%   |
|  | 8%    |
| None                                                                              | 70%   |

## Sources:


- [OpenAI API Autologger Quickstart](#)
- [OpenAI API](#)

## References

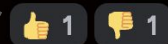
>[https://github.com/wandb/examples/tree/master/colabs/openai/OpenAI\\_API\\_Autologger\\_Quickstart.ipynb](https://github.com/wandb/examples/tree/master/colabs/openai/OpenAI_API_Autologger_Quickstart.ipynb)

<https://docs.wandb.ai/guides/integrations/openai>

<https://docs.wandb.ai/guides/prompts/openai>

 If you still need help please try re-phrase your question, or alternatively reach out to the Weights & Biases Support Team at [support@wandb.com](mailto:support@wandb.com)

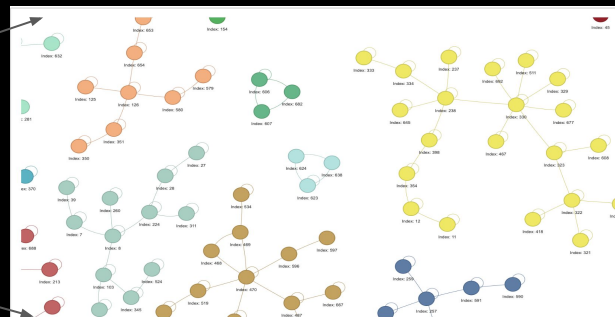
Was this response helpful? Please react below to let us know



# Create evaluation set

There's 1,000s of questions.

We cluster them into topics to efficiently evaluate on the breadth of questions asked.







# Evaluate the model: Expert Eval

Home / wc-anish / wandbot-response-annotation

Introduce a query Pending

**Pending**

**Query**

Hey I have a question about using wandb with fastapi in a prod environment. is it recommended to initialize wandb within a specific route function, ie

```
@app.route('/') def my_function(): wandb.init(...)
```

or should i initialize beforehand:

```
wandb.init(...) @app.route('/') def my_function(): ...
```

I'm getting a long list of log items in the console and many of them are empty

**Response**

The initialization of `wandb.init()` should be done as early as possible in your script to capture all console outputs, including error messages, which aids in debugging. However, in a FastAPI context, if you want to log different runs for different requests, it might be more appropriate to initialize `wandb` within your route function.

For example:

```
@app.route('/')
def my_function():
    run = wandb.init(project="my_project", reinit=True)
```

**Submit your feedback** Saved

Read the annotation guidelines

**Wandbot Response Accuracy**

Search labels

correct  incorrect  unsure

**Wandbot Link Hallucination**

YES  NO

**Is the Query related to W&B?**

YES  NO

**Note**

Clear Discard Submit

< Prev Next >

70%  
acc

Is it correct?

Is the link real

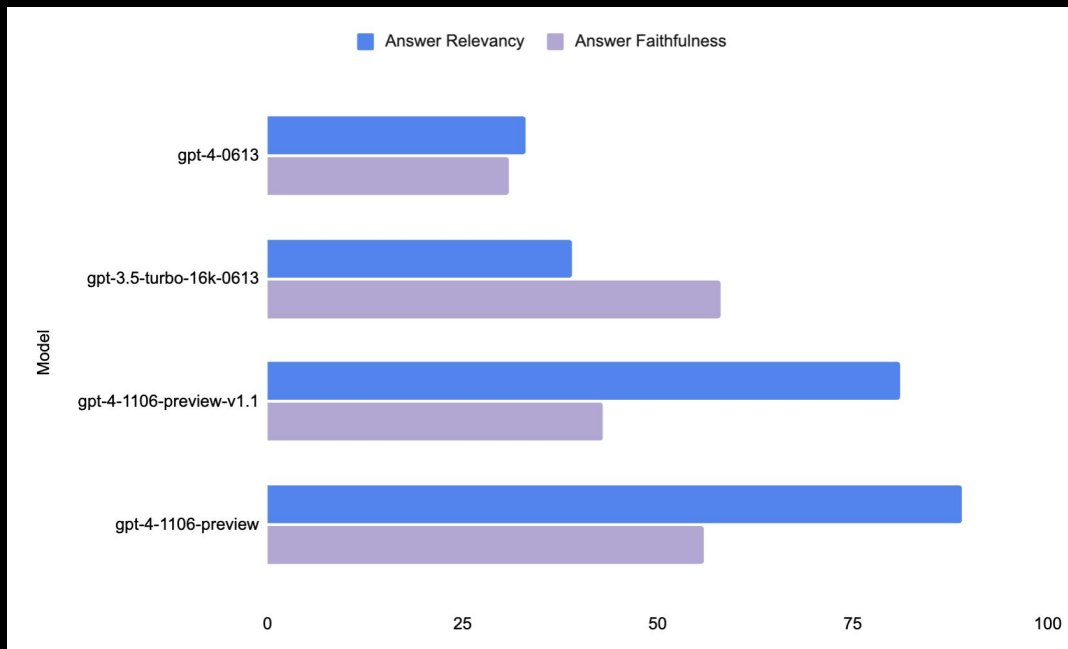
Is the question related to wandb

# Evaluate the model: LLM Eval

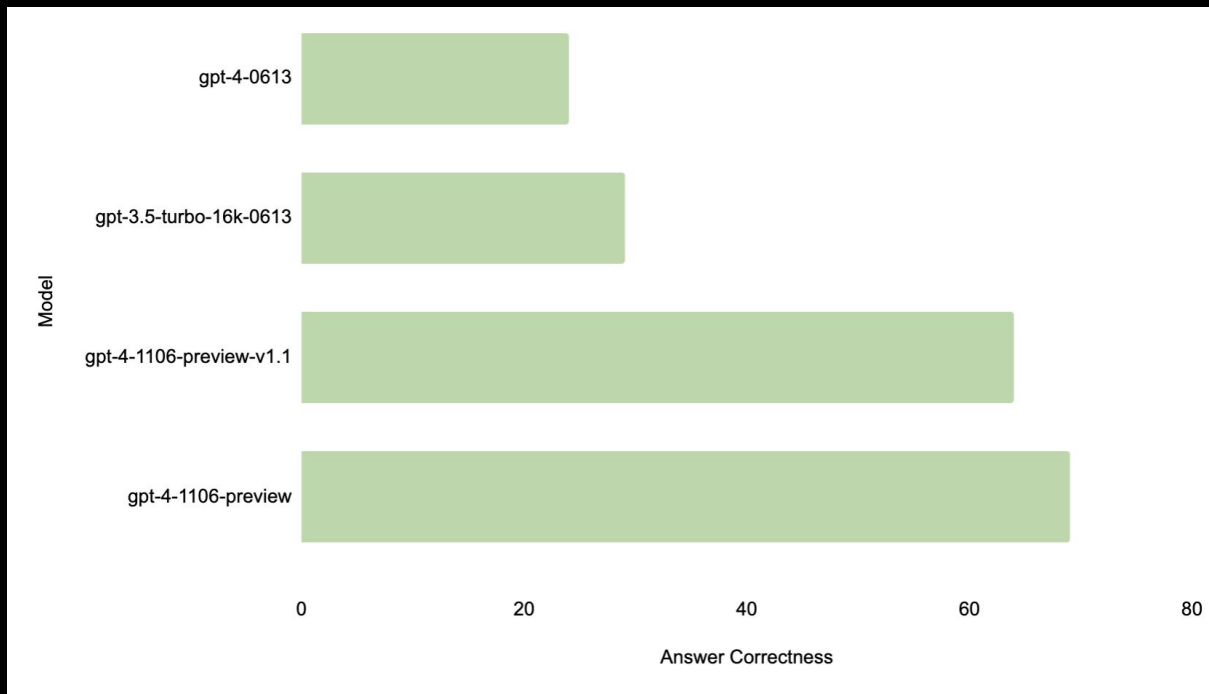
- **Faithfulness Evaluation:** does the answer **accurately** reflect the information in the source documents without introducing unverified or incorrect details?
- **Relevancy Evaluation:** Does the answer address the user's query with information **related** to the question and context provided?

| Faithfulness Accuracy (%)  | Relevancy Accuracy (%)  |
|----------------------------|-------------------------|
| baseline-faithfulness-eval | baseline-relevancy-eval |
| 53.788                     | 61.364                  |

# Try more models

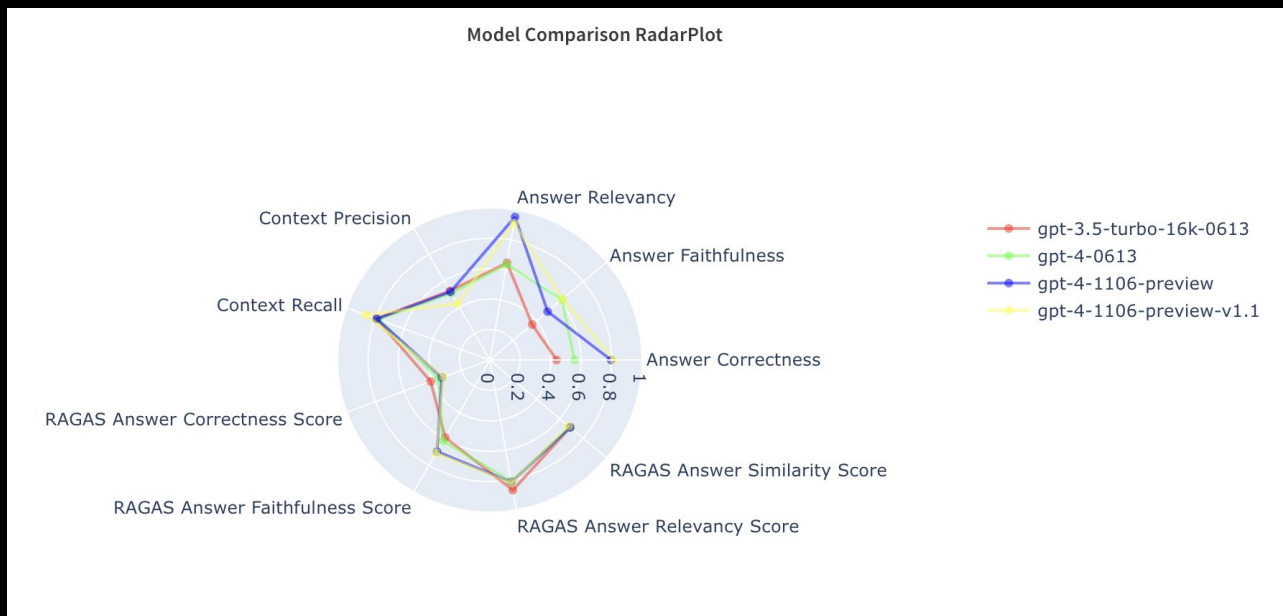


# Try more models





# Define better metrics



# Define better metrics

## ragas score

generation

### faithfulness

how factually accurate is  
the generated answer

### answer relevancy

how relevant is the generated  
answer to the question

retrieval

### context precision

the signal to noise ratio of retrieved  
context

### context recall

can it retrieve all the relevant information  
required to answer the question

# Read more at [wandb.me/wandbot-eval](https://wandb.me/wandbot-eval)

W&B Fully Connected > Articles > LLM

## Evaluation-Driven Development: Improving WandBot, our LLM- Powered Documentation App

This report describes the changes and enhancements we made to wandbot during our most recent sprint

[Bharat Ramanathan](#)

[Share](#)

[Comment](#)

[4 stars](#)

...

Last Updated: Feb 26, 2024

runs.summary["annotated\_data"]

|   | question                  | answer                    | context | correctness | is_wandb_query | notes                    |
|---|---------------------------|---------------------------|---------|-------------|----------------|--------------------------|
| 1 | Hey I have a question     | When integrating          | Source: | correct     | YES            | The answer clearly       |
| 2 | Hey with wandb is it      | Yes, with `wandb`, you    | Source: | correct     | YES            | This answer correctly    |
| 3 | Explain how I can version | Versioning datasets with  | Source: | correct     | YES            | The answer correctly     |
| 5 | Hi, can anybody help me   | The                       | Source: | correct     | YES            | This requires more data  |
| 6 | what is the difference    | `artifact.add_file` and   | Source: | correct     | YES            | The answer correctly     |
| 7 | Hi I have created an      | If you've created an      | Source: | correct     | YES            | The answer provides      |
| 8 | How can I add prediction- | To add prediction-related | Source: | correct     | YES            | While this does not take |
| 9 | On my laptop, wandb is    | To switch between your    | Source: | correct     | YES            | The provided answer      |

# Serve wandbot on Slack and Discord

## API Server

The screenshot shows a REST client interface with a POST request to `https://wandboten.bharatramanatha.repl.co/query`. The request body is a JSON object: `{ "question": "Hi @wandbot, what are you ?", "application": "slack" }`. The response is a JSON object: `{ "answer": "I am wandbot, a developer assistant designed to guide users with tasks related to Weights & Biases, its SDK 'wand', and its visualization library 'weave'. I provide helpful answers to queries using document excerpts and code examples provided in the context. I do not generate code from prior knowledge or create any methods, functions, and classes that are not found in the provided context. I always respond in the user's language and in a concise, fully formatted Markdown with the necessary code and links.", "model": "gpt-4-8613", "sources": "https://docs.wandb.ai/ja/guides/launch/sagemaker, https://docs.wandb.ai/guides/track/app, https://docs.wandb.ai/quickstart, https://github.com/wandb/examples/tree/master/colabs/prompts/Wand LLM QA_bot.ipynb, https://docs.wandb.ai/guides/integrations/dagster, https://github.com/wandb/examples/tree/master/colabs/huggingface/wandb_llm_example.ipynb, https://docs.wandb.ai/guides/integrations/huggingface, https://docs.wandb.ai/ja/guides/launch/create-launch-job, https://docs.wandb.ai/guides/app/features/pipeline/weaving/embedding-projector, https://docs.wandb.ai/app/pages/project-page, https://docs.wandb.ai/tutorials/experiments, https://docs.wandb.ai/tutorials/experiments, https://docs.wandb.ai/guides/integrations/huggingface, https://docs.wandb.ai/ja/guides/technical-faq/general, https://docs.wandb.ai/ja/guides/name/" }`

## SlackBot

The screenshot shows a Slack thread. A message from `wandbot (beta)` says: "Please note that wandbot is currently in alpha testing and will experience frequent updates. Please do not share any private or sensitive information in your query at this time. Please note that overly long messages (>1024 words) will be truncated." A reply from `Corey Strausman` says: "Hi @Corey\_S: Please note that wandbot is currently in alpha testing and will experience frequent updates." Another reply from `wandbot (beta)` says: "Sure, you can add a file from an S3 bucket to a wandb Artifact as a reference using the `add_reference()` method. Here is an example:" followed by a code block:

```
import wandb

# Initialize a wandb run
run = wandb.init()

# Create a new artifact
artifact = wandb.Artifact('my_artifact', type='dataset')

# Add a reference to a file in an S3 bucket
artifact.add_reference('s3://my-bucket/my-file')

# Log the artifact
run.log_artifact(artifact)
```

## DiscordBot

The screenshot shows a Discord thread. A message from `wandbot (beta)` says: "Hi @Corey\_S: Please note that wandbot is currently in alpha testing and will experience frequent updates." A reply from `wandbot (beta)` says: "Please do not share any private or sensitive information in your query at this time. Please note that overly long messages (>1024 words) will be truncated." Another reply from `wandbot (beta)` says: "Generating response..." followed by a code block:

```
import wandb

# Initialize a wandb run
run = wandb.init()

# Create a new artifact
artifact = wandb.Artifact('my_artifact', type='dataset')

# Add a reference to a file in an S3 bucket
artifact.add_reference('s3://my-bucket/my-file')

# Log the artifact
run.log_artifact(artifact)
```



# Feedback

Ever since we added Wandbot to our workflow, it has helped us save time writing code snippets and examples for customers, **saving us at least an hour a week** on scripts.

It has also **made onboarding easier** for our two new teammates; **instead of digging through our docs for hours**, they can simply ask Wandbot a question, and it will return a very sweet and short-to-the-point answer.

It is always a good starting point for debugging issues.

# Common questions

- 1) Which foundation model should I use?
- 2) Should I use fine tuning?
- 3) How should I set up my vector database?
- 4) What kinds of prompts should I use?
- 5) How should we build safety into the model?

# Lessons for building LLM-powered apps

## 1. Build an evaluation framework

✦ Is the Answer Correct ? \* ⌵

Search labels

correct incorrect unsure

✦ Is the Answer Correct According to GPT-4 ? ⌵

Search labels

correct incorrect unsure

✦ Is it a W&B Query? \* ⌵

YES NO

✦ Notes ⌵

Write Suggestion

The answer correctly identifies the distinction between "artifact.add\_file" and "wandb.save" in the context of Weights & Biases (W&B) based on the provided document. It explains that "artifact.add\_file" is for adding files to a versioned W&B Artifact with versioning and deduplication capabilities, while "wandb.save" is for syncing run-specific files with the W&B server. The answer also provides correct example usage for each method, aligning with the usage and intention described in the document.

✦ GPT Explanation ⌵

Suggestion

The answer correctly identifies the distinction between "artifact.add\_file" and "wandb.save" in the context of Weights & Biases (W&B) based on the provided document. It explains that "artifact.add\_file" is for adding files to a versioned W&B Artifact with versioning and deduplication capabilities, while "wandb.save" is for syncing run-specific files with the W&B server. The answer also provides correct

Clear Discard Submit

# Lessons for building LLM-powered apps

## 2. Start with a lightweight prototype

```
class GenTextOpenAIModel(weave.Model):
    model_name: str
    system_prompt: str
    user_template: str

    @weave.op()
    async def predict(self, user: str) -> dict:
        # prompt = example["text"]

        model_client = OpenAI(
            # This is the default and can be omitted
            # api_key=os.environ["OPENAI_API_KEY"]
        )

        response = model_client.chat.completions.create(
            model=self.model_name,
            messages=[
                {"role": "system", "content": self.system_prompt},
                {"role": "user", "content": user},
            ],
            max_tokens=100
        )
        print(response.choices[0].message.content)
        return {'generated_text': response.choices[0].message.content}

def evaluate_openai(dataset_name: str):
    dataset = weave.ref(dataset_name).get()

    openai_system_prompt = open("prompts/openai_system.txt", 'r').read()
    openai_user_template = open("prompts/openai_user.txt", "r").read()

    weave_model = GenTextOpenAIModel(model_name="gpt-4", system_prompt=openai_system_prompt,
    eval = weave.Evaluation(dataset=dataset, scorers=[match])
    asyncio.run(eval.evaluate(weave_model))
    # table, acc, acc_lousy = create_predictions_table(model, tokenizer, test_dataset, 64)
```



# Lessons for building LLM-powered apps

## 3. Incorporate end-user feedback



**Zachary Blasczyk** 11:31 AM

@Wandbot use <https://learn.microsoft.com/en-us/python/api/overview/azure/storage?view=azure-python> to write a sample python example that uses W&B to log an artifact by reference.



4 replies Last reply 2 months ago



**Morgan** 2 months ago

haha it doesn't have browsing capabilities, although we do use a search api as part of retrieval... maybe we could make it work.... 🤔



**Zachary Blasczyk** 2 months ago

Good to know. I assumed it could pick up on the context based on the URL even if it couldn't brows the docs.

Just wanted to point it in the right direction given that I have no idea what library to use.

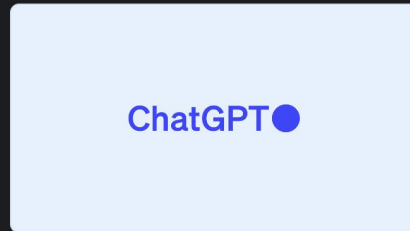
Here is a GPTy response if you are interested:

<https://chat.openai.com/share/e5895bfb-9480-43da-bf1d-f745ff28d996>

 chat.openai.com

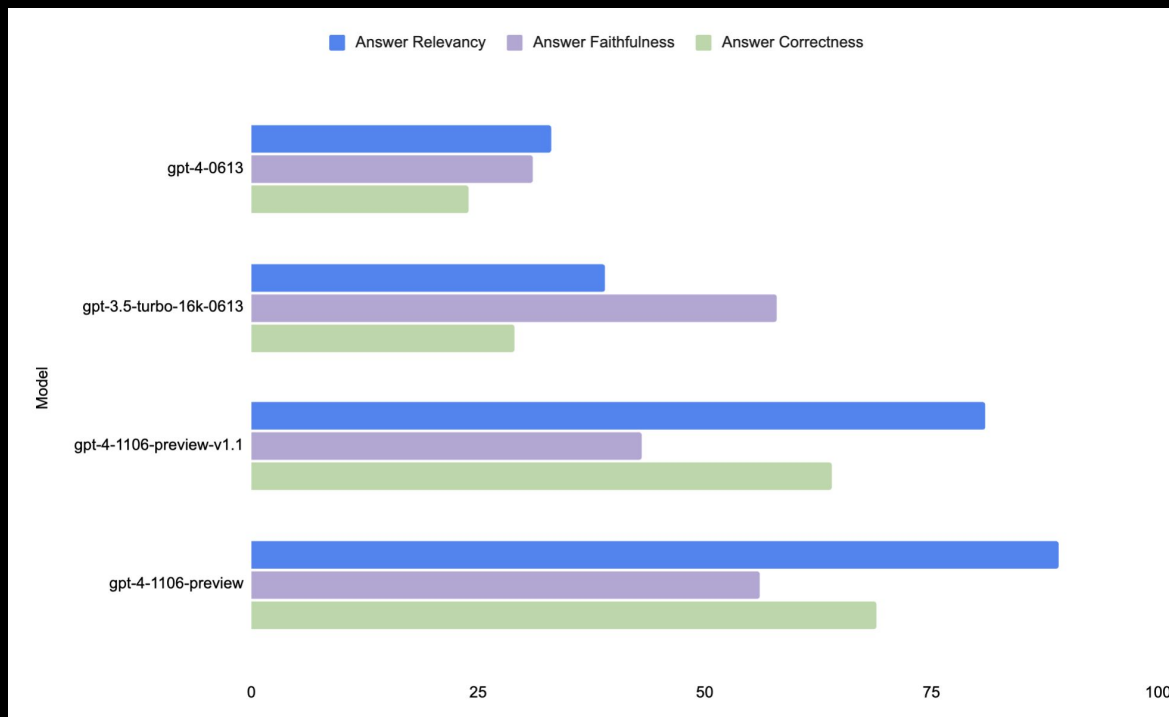
ChatGPT

A conversational AI system that listens, learns, and challenges (15 kB)



# Lessons for building LLM-powered apps

## 4. Iterate



# Lessons for building LLM-powered apps

- 1. Build an evaluation framework**
2. Start with a lightweight prototype
3. Incorporate end-user feedback
4. Iterate

# Evaluation Best Practices

- 1) Use multiple evaluation sets/techniques
- 2) Make sure metrics correlate with user experience
- 3) Actually do it

Weave Documentation API Reference GitHub

Getting Started

- Introduction
- Quickstart
- Tutorial: Build an Evaluation pipeline
- Using Weave
- Core Types
  - Models
  - Datasets
  - Evaluation
- Tracking
  - Objects
  - Ops
  - Tracing
- Ecosystem
  - OpenAI
- Tools
  - Serve
  - Deploy
- API Reference

## Introduction

*This version of Weave is pre-release software.*

Weave is a toolkit for developing AI-powered applications, built by [Weights & Biases](#).

Our goal is to bring rigor, best-practices, and composability to the inherently experimental process of developing AI-based software, without introducing cognitive overhead.

[Get started](#) by decorating Python functions with `@weave.op()`.

Seriously, try the [quickstart](#) first.

You can use Weave to:

- Log and debug language model inputs, outputs, and traces
- Build rigorous, apples-to-apples evaluations for language model use cases
- Organize all the information generated across the LLM workflow, from experimentation to evaluations to production

### Key concepts

Weave's **core types** layer contains everything you need for organizing AI projects, with built-in lineage, tracking, and reproducibility.

- **Datasets:** Version, store, and share rich tabular data.
- **Models:** Version, store, and share parameterized functions.
- **Evaluations:** Test suites for AI models.

wandb.me/weave



llm-play · otto11 · Weave · calls Lavanya Shukla wandb

### Evaluation.predict\_and\_score Traces

Category  Evaluation.predict\_and\_score:v0 Parent: Evaluation.evaluate (c67...e0b) Roots only

Common values: Category Predict input.self test-labels-evaluation:v0 input.model GenTextAnthropic:v0

| Trace                             | Status | Called ↓      | Ref | inputs             |                  |                       | output |                |           |     | Latency |
|-----------------------------------|--------|---------------|-----|--------------------|------------------|-----------------------|--------|----------------|-----------|-----|---------|
|                                   |        |               |     | user               | answer           | generated_text        | acc    | correct format | other acc |     |         |
| Evaluation.predict_and_score cc52 | 🟢      | 4 minutes ago | 🔗   | attached fro...    | other()          | other()               | ✓      | ✓              | ✓         | 1s  |         |
| Evaluation.predict_and_score 01e8 | 🟢      | 4 minutes ago | 🔗   | Do you think ...   | other()          | other()               | ✓      | ✓              | ✓         | 3s  |         |
| Evaluation.predict_and_score 345c | 🟢      | 4 minutes ago | 🔗   | Any updates ...    | news()           | news()                | ✓      | ✓              | N/A       | 5s  |         |
| Evaluation.predict_and_score 1ba1 | 🟢      | 4 minutes ago | 🔗   | Um... I think i... | other()          | other()               | ✓      | ✓              | ✓         | 7s  |         |
| Evaluation.predict_and_score 8712 | 🟢      | 4 minutes ago | 🔗   | That was a go...   | other()          | other()               | ✓      | ✓              | ✓         | 3s  |         |
| Evaluation.predict_and_score 6d02 | 🟢      | 4 minutes ago | 🔗   | Oh my, Omi.        | other()          | other()               | ✓      | ✓              | ✓         | 4s  |         |
| Evaluation.predict_and_score b883 | 🟢      | 4 minutes ago | 🔗   | Snored, oh w...    | other()          | other()               | ✓      | ✓              | ✓         | 6s  |         |
| Evaluation.predict_and_score 579e | 🟢      | 4 minutes ago | 🔗   | That should b...   | other()          | other()               | ✓      | ✓              | ✓         | 7s  |         |
| Evaluation.predict_and_score 2a68 | 🟢      | 4 minutes ago | 🔗   | I would like to... | notes(action=... | notes(action="add"... | ✓      | ✓              | N/A       | 10s |         |
| Evaluation.predict_and_score 29b7 | 🟢      | 4 minutes ago | 🔗   | That's so goo...   | other()          | other()               | ✓      | ✓              | ✓         | 12s |         |
| Evaluation.predict_and_score b6a4 | 🟢      | 4 minutes ago | 🔗   | Okay, Okay. O...   | other()          | other()               | ✓      | ✓              | ✓         | 14s |         |

Total Rows: 62

wandb.me/weave

llm-play · otto11 · Weave · calls

Lavanya Shukla  
wandb

### Traces

Category Op Roots only

Common values: Category Evaluate input.self test-labels-evalua

| Trace               | Status | Called ↓      | model >        |
|---------------------|--------|---------------|----------------|
| Evaluation.evaluate | 3e0b   | 5 minutes ago | GenTextAnthr.. |
| Evaluation.evaluate | 4e36   | 9 minutes ago | gpt-4-v0       |

Total Rows: 2

### Trace tree

- Evaluation.evaluate
  - Evaluation.predict\_and\_score
    - GenTextAnthropic.predict
    - match
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score
  - Evaluation.predict\_and\_score

### Evaluation.evaluate 3e0b

Call Code Summary

**Inputs**

| Path  | Value                     |
|-------|---------------------------|
| self  | test-labels-evaluation:v0 |
| model | GenTextAnthropic:v0       |

**Outputs**

| Path           | Value    |
|----------------|----------|
| match          |          |
| acc            |          |
| true_count     | 58       |
| true_fraction  | 0.935484 |
| correct format |          |
| true_count     | 62       |
| true_fraction  | 1        |
| other acc      |          |
| true_count     | 47       |

wandb.me/weave



**Thank you**

**@l2k**





More Slides



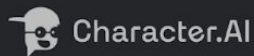
**LLMs**



# LLM Creators



# LLM Creators Using W&B





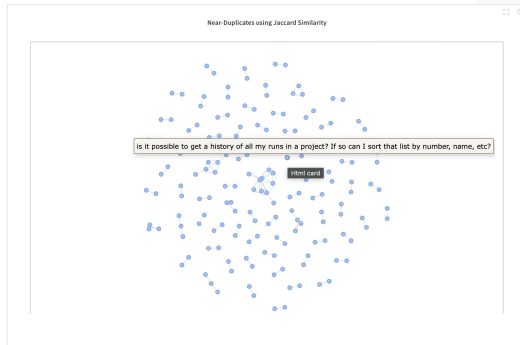
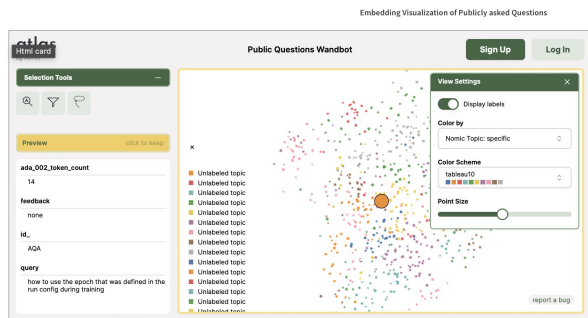
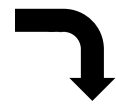
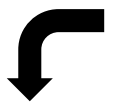
# LLM Users



# We can then take snapshots of the feedback-question-answer data for analysis such as topic analysis to determine what data we want to score the effectiveness of the model

2 none  
**Questions coming in from our users**

611  
**Visualize it**



@wandbot (beta) how to use wandb log with a flask app

@wandbot (beta) why is the best run from the sweep api sweep = api.sweep(f"qmarila/mlops-

@wandbot (beta) I am training a spacy textcat model. This process produces two model directories: model-

< 1-3 of 74 >

@wandbot (beta) how can i log anthropic responses in wandb

@wandbot (beta) are you there?

@wandbot (beta) Hey I have a question about using wandb with fastapi in a prod environment. is it

< 1-3 of 611 >

@wandbot (beta) How is langsmith different from W&B prompting?

@wandbot (beta) How is langsmith different from W&B prompting?

@wandbot (beta) Can I make wandb only record a video of a single environment rather than one for every CPU

< 1-3 of 187 >





# Creating our advanced prompt - Create System Message

You are wandbot, a developer assistant designed to guide users with tasks related to Weight & Biases, its sdk `wandb` and its visualization library `weave`.

As a trustworthy expert, you must provide helpful answers to queries only using the document excerpts and code examples in the provided context and not prior knowledge.

| Faithfulness Accuracy (%)  | Relevancy Accuracy (%)  | Hit Rate (sub-section)                 | Mean Reciprocal Ranking (sub-section)  |
|----------------------------|-------------------------|----------------------------------------|----------------------------------------|
| baseline-faithfulness-eval | baseline-relevancy-eval | baseline-retriever-eval-prod@launch-en | baseline-retriever-eval-prod@launch-en |
| 53.788                     | 61.364                  | 0.7927                                 | 0.7439                                 |

Define a **system message** to set the overall tone and behavior of the LLM

# Creating our advanced prompt - Define Guidelines

Here are your guidelines:

1. Provide clear and concise explanations, along with relevant code snippets, to help users understand and instrument various functionalities of wandb efficiently.
2. Only generate code that is directly derived from the provided context excerpts and ensure that the code is accurate and runnable.
3. Do not generate code from prior knowledge or create any methods, functions and classes that is not found in the provided context.
4. Always cite the sources from the provided context in your response.
5. Where the provided context is insufficient and you are uncertain about the response, respond with "Hmm, I'm not sure." and direct the user to the [Weights & Biases \[support\]\(support@wandb.com\)](mailto:support@wandb.com) or [\[community forums\]\(http://wandb.me/community\)](http://wandb.me/community)
6. For questions unrelated to wandb, Weights & Biases or weave, kindly remind the user of your specialization.
7. Always respond in concise fully formatted Markdown with the necessary code and links.
8. For best user experience, always respond in the user's language. For instance, if the query is in Japanese, you should respond in Japanese

| Faithfulness Accuracy (%)  | Relevancy Accuracy (%)  | Hit Rate (sub-section)                 | Mean Reciprocal Ranking (sub-section)  |
|----------------------------|-------------------------|----------------------------------------|----------------------------------------|
| baseline-faithfulness-eval | baseline-relevancy-eval | baseline-retriever-eval-prod@launch-en | baseline-retriever-eval-prod@launch-en |
| 53.788                     | 61.364                  | 0.7927                                 | 0.7439                                 |

Set a clear set of guidelines to define what behavior is available and NOT available to the bot.

- We can define instructions to prevent hallucinations
- We can define how the LLM should respond to uncertainty

# Creating our advanced prompt - Add Strictly Formatted Examples

Here are some examples:

```
<!--start-example1-->
<!--start-relevant-documents-->
Weights & Biases allows logging of audio data arrays or
files for playback in W&B.
You can use the `wandb.Audio()` to create audio instances
and log them to W&B using `wandb.log()`.
Source: 28-pl
```

~~~

```
# Log an audio array or file
wandb.log({"my whale song": wandb.Audio(array_or_path,
caption="monterey whale 0034", sample_rate=32)})
Source: 29-pl
```

```
# Log multiple audio files
# Log audio within a W&B Table
my_table = wandb.Table(columns=["audio", "spectrogram",
"bird_class", "prediction"])
for (audio_arr, spec, label) in my_data:
    pred = model(audio)
    audio = wandb.Audio(audio_arr, sample_rate=32)
    img = wandb.Image(spec)
    my_table.add_data(audio, img, label, pred)
```

```
# Log the Table to wandb
wandb.log({"validation_samples" : my_table})
Source: 30-pl
```

~~~

```
<!--end-relevant-documents-->
```

| Faithfulness Accuracy (%)  | Relevancy Accuracy (%)  | Hit Rate (sub-section)                     | Mean Reciprocal Ranking (sub-section)      |
|----------------------------|-------------------------|--------------------------------------------|--------------------------------------------|
| baseline-faithfulness-eval | baseline-relevancy-eval | baseline-retriever-eval-<br>prod@launch-en | baseline-retriever-eval-<br>prod@launch-en |
| 53.788                     | 61.364                  | 0.7927                                     | 0.7439                                     |

Use specific formatting to provide representative examples to the LLM; simulating an ideal user/data/response grouping



After analysis, we use the heuristics we set to create a filtered set of data for evaluation

### Filter the questions

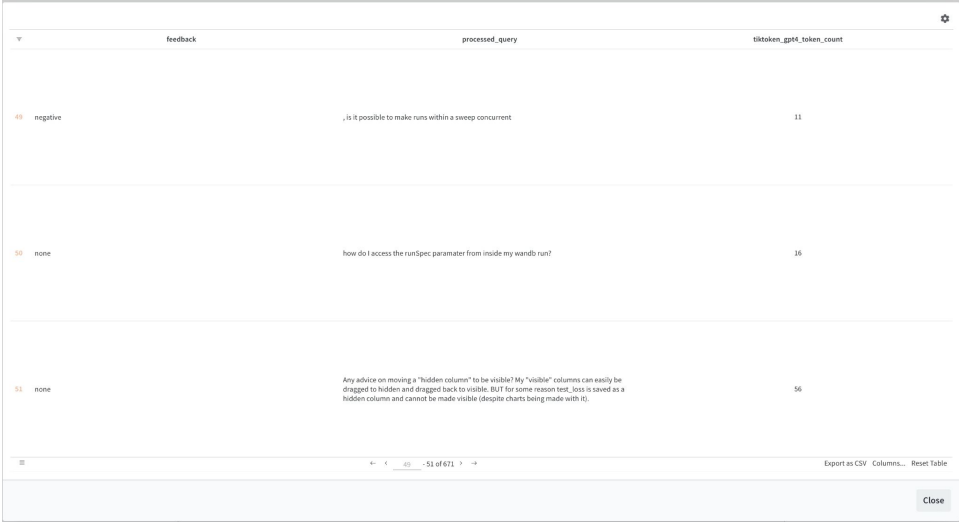


The screenshot shows a table with columns: feedback, count, and query. The data is as follows:

| feedback   | count | query                                                                                                                                                                                                                                                                      |
|------------|-------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| 1 negative | 74    | @wandbot (beta) how to use wandb log with a flask app<br>@wandbot (beta) when I get the best run from the sweep and sweep = spt caseqpt/ptwvta/mjss...<br>@wandbot (beta) I am training a spacy neural model. This process produces two model directories: model-1.3 of 74 |
| 2 none     | 611   | @wandbot (beta) how can I log anthropic responses in wandb<br>@wandbot (beta) are you there?<br>@wandbot (beta) Hey I have a question about using wandb with fastapi in a prod environment, is it...<br>1.3 of 611                                                         |
| 3 positive | 187   | @wandbot (beta) How is langsmith different from W&B prompting?<br>@wandbot (beta) How is langsmith different from W&B prompting?<br>@wandbot (beta) Can't make wandb only record a video of a single environment rather than one for every CPU...<br>1.3 of 187            |



### Golden eval set

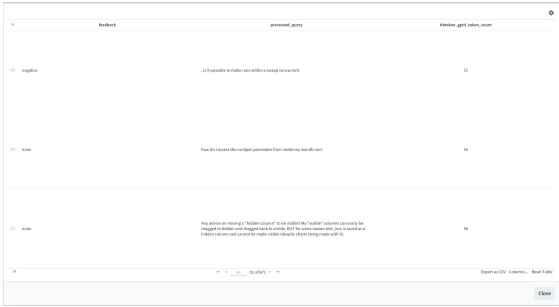


The screenshot shows a filtered table with columns: feedback, processed\_query, and tiktoken\_gpt4\_token\_count. The data is as follows:

| feedback    | processed_query                                                                                                                                                                                                                                                   | tiktoken_gpt4_token_count |
|-------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|
| 49 negative | , is it possible to make runs within a sweep concurrent                                                                                                                                                                                                           | 11                        |
| 50 none     | how do I access the runSpec parameter from inside my wandb run?                                                                                                                                                                                                   | 16                        |
| 51 none     | Any advice on moving a "hidden column" to be visible? My "visible" columns can easily be dragged to hidden and dragged back to visible. BUT for some reason test_logs is saved as a hidden column and cannot be made visible (despite charts being made with it). | 56                        |

The cheapest, and quickest form of evaluation is to use a separate LLM to score the question answer pairs for us

**Golden eval set**



| text                                                                                                                                                                                                                               | ground_truth | dataset_name |
|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------|--------------|
| ... possible to make one without using an internet                                                                                                                                                                                 | 11           |              |
| How do I create the bridge connection from my desktop to my laptop                                                                                                                                                                 | 12           |              |
| Any other one that might be better than the one I have? I have the one that is the best one I have found so far. I have the one that is the best one I have found so far. I have the one that is the best one I have found so far. | 13           |              |



Evaluation LLM

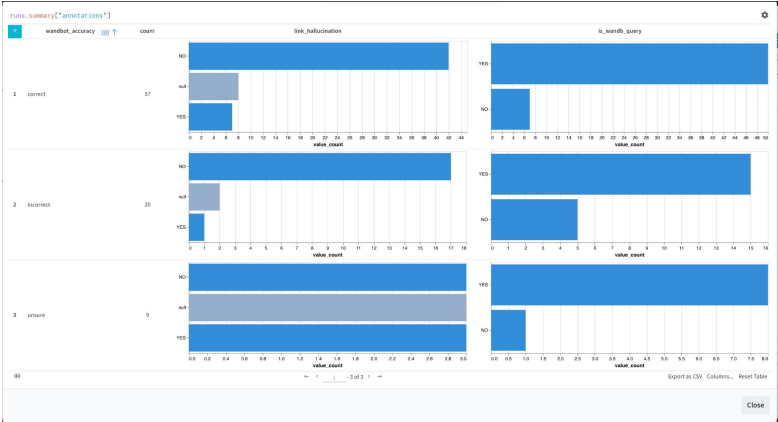


|                            |                         |
|----------------------------|-------------------------|
| Faithfulness Accuracy (%)  | Relevancy Accuracy (%)  |
| baseline-faithfulness-eval | baseline-relevancy-eval |
| 53.788                     | 61.364                  |

However, LLM based evaluation has its own set of problems that may not accurately represent human quality. Manual evaluation is still essential for determining if the LLM is operating correctly

|    | feedback | processed_query                                                                                                                                                                                            | tiktoken_gpt4_token_count |
|----|----------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------------|
| 49 | negative | , is it possible to make runs within a sweep concurrent                                                                                                                                                    | 11                        |
| 50 | none     | how do i access the runSpec paramater from inside my wandb run?                                                                                                                                            | 16                        |
| 51 | none     | Any advice on moving a "hidden column" to be visible? My "visible" columns can easily be dragged to hidden and dragged back to visible. BUT for some reason there is a "wandb" hidden column and cannot be | 56                        |
| 52 | none     | how can i make a hidden column visible?                                                                                                                                                                    | 9                         |
| 53 | negative | how can i use sweep = wandb.controller(sweep_id)?                                                                                                                                                          | 13                        |
| 54 | none     | how to run another run after using sweep in jupyterlab?                                                                                                                                                    | 13                        |

Manual Human Evaluation



Current accuracy stands at 65.26%.