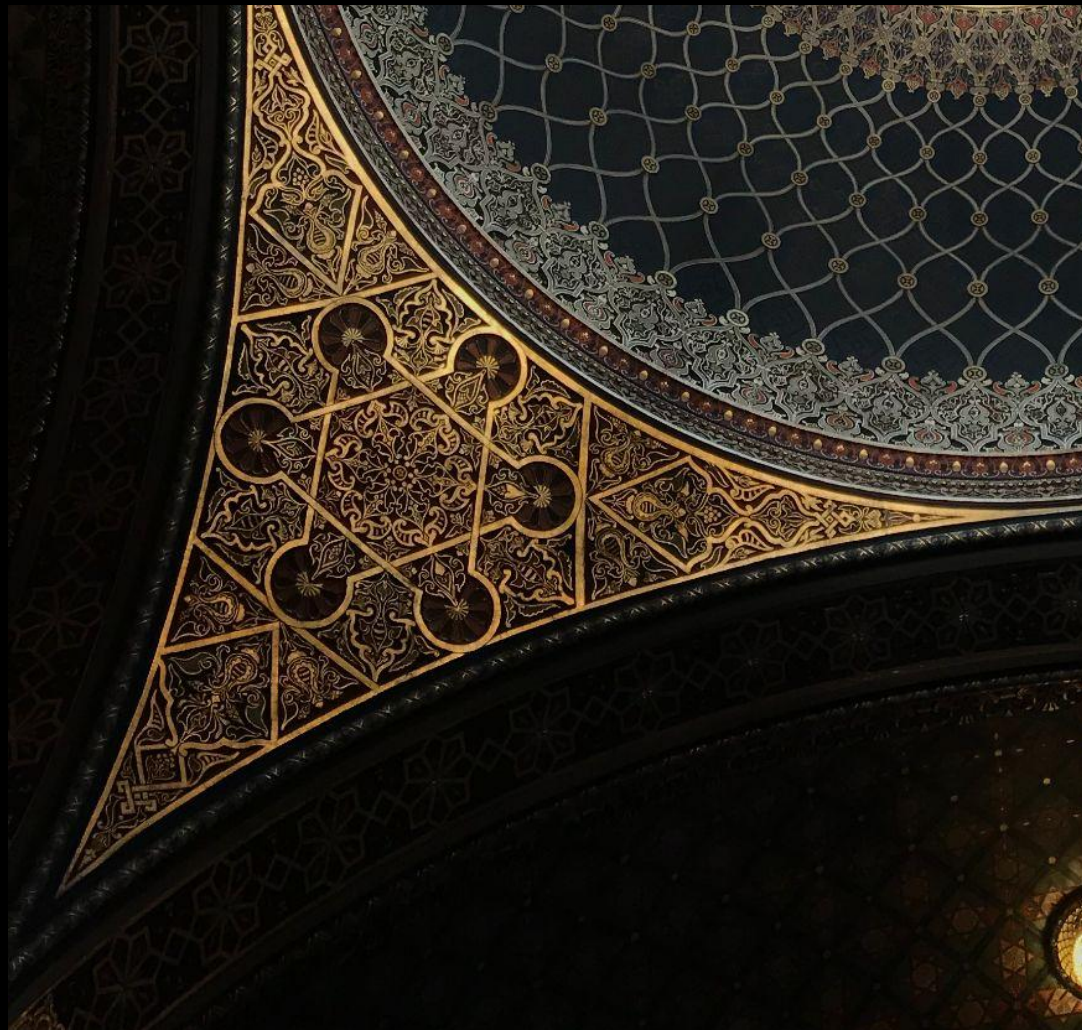
Weights & Biases

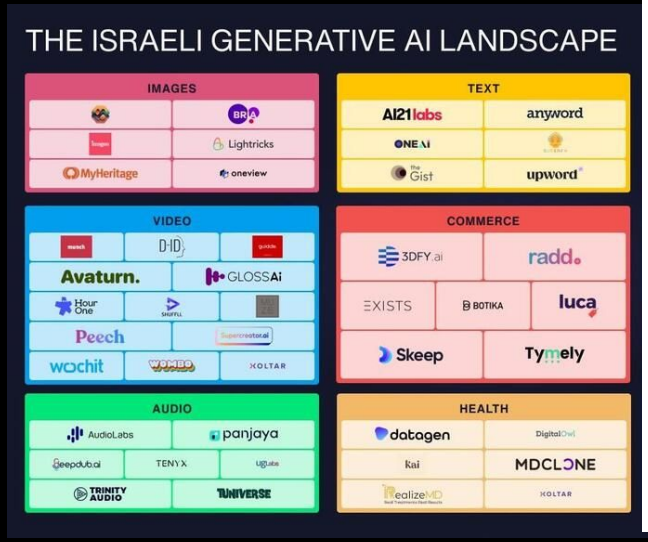
Evaluating Gen AI Models

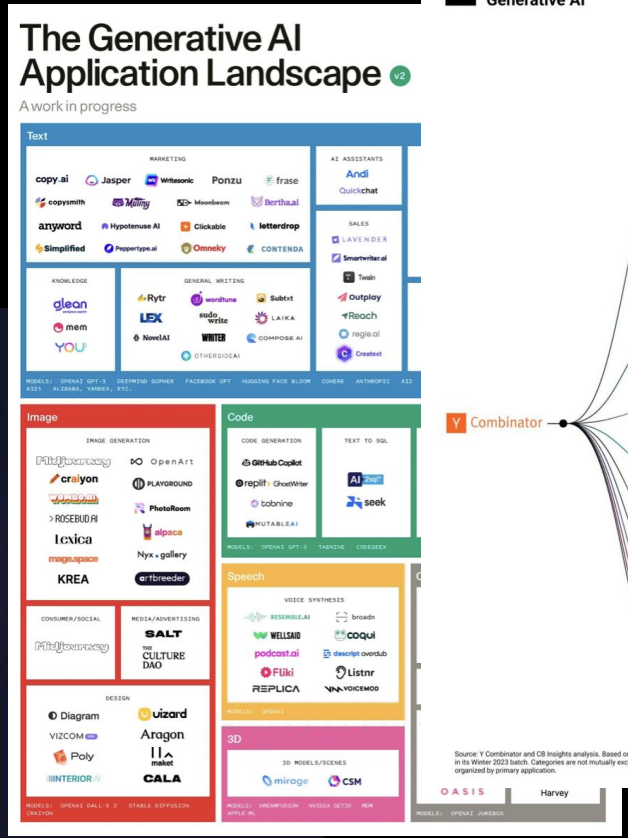
Hans Ramsl

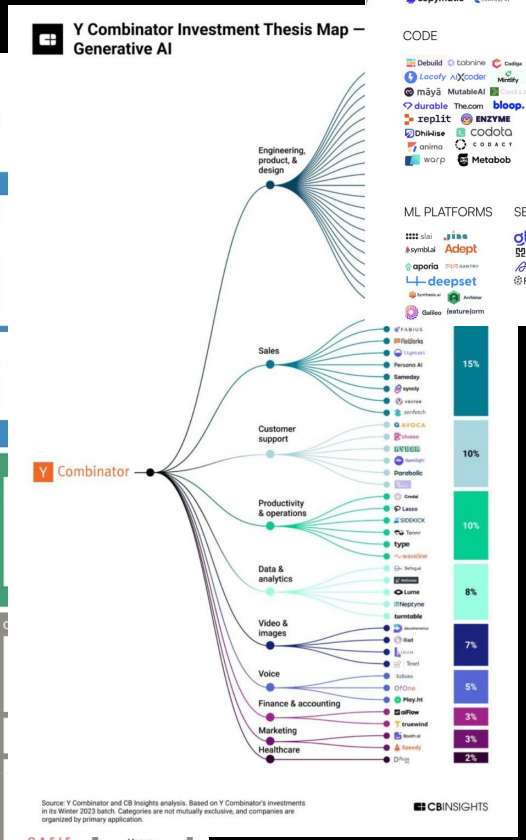
December 5th 2023

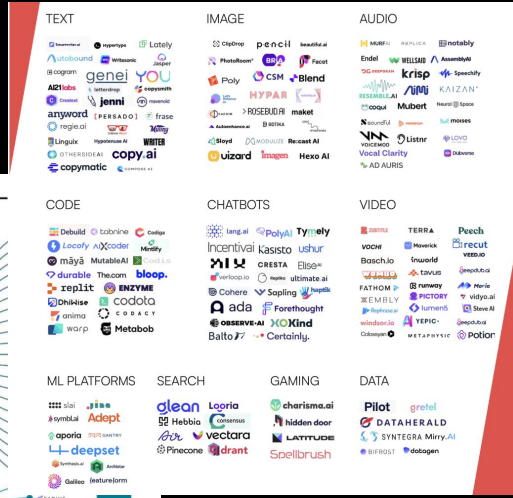


Cambrian explosion of Gen AI Startups









The best teams use the best tools for machine learning

W&B Products

Models

- **Experiments** — Track and visualize your ML experiments
- **Sweeps** — Optimize your hyperparameters
- **Model Registry** — Register and manage your ML models
- **Automations** — Trigger workflows automatically
- **Launch** — Package and run your ML workflow jobs

Prompts

- **Traces** — Explore and debug LLMs
- **Monitoring** — Observe and monitor production ML

Platform

- **Artifacts** — Version and manage your ML pipelines
- **Tables** — Visualize and explore your ML data
- **Reports** — Document and share your ML insights
- **Weave** — Data application building toolkit

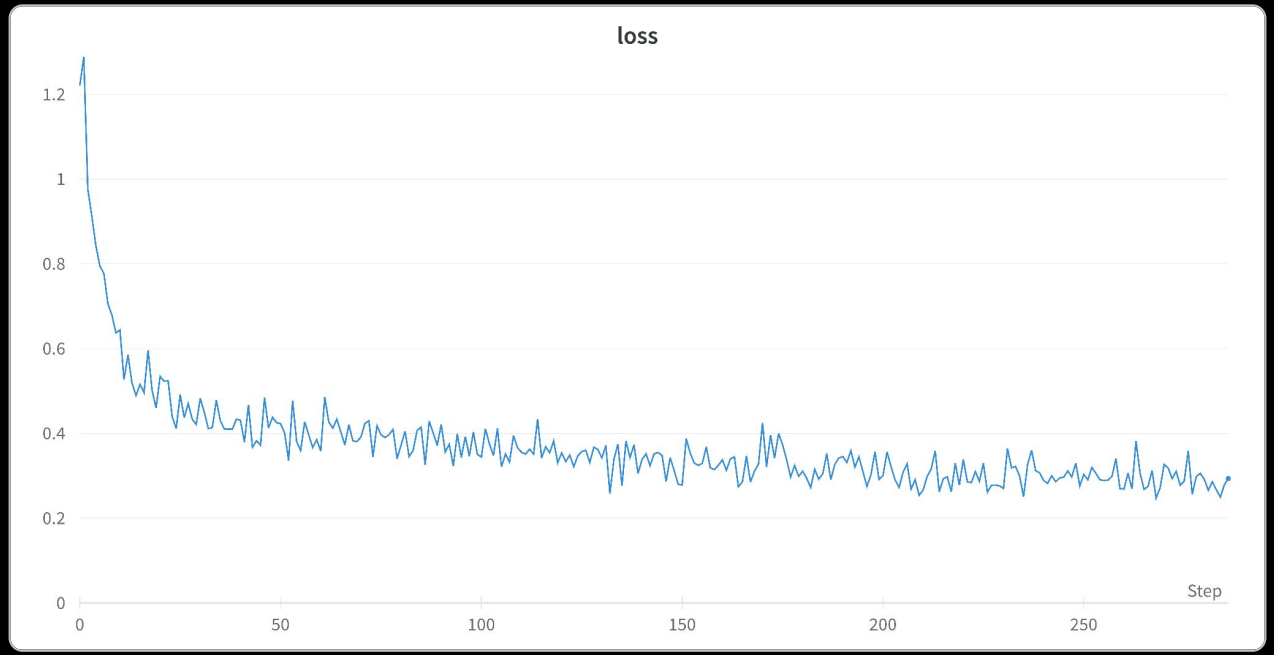
Secure Deployments

- SaaS cloud
- Dedicated Cloud
- Customer Managed

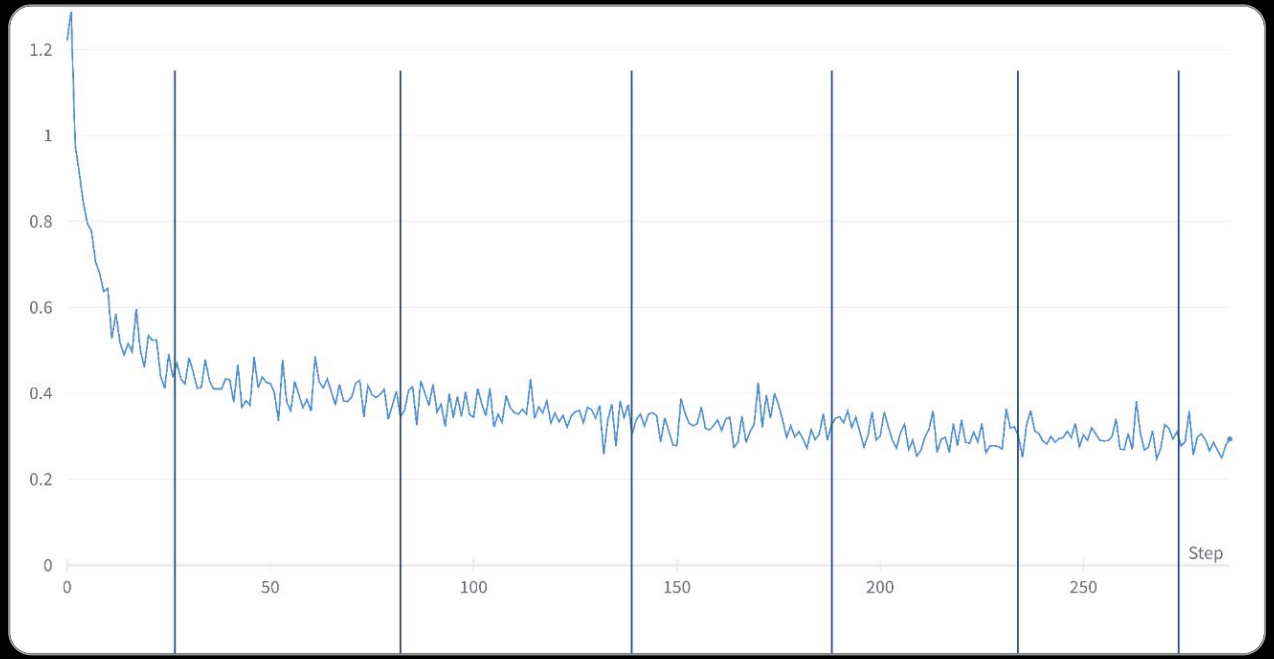
Infrastructure Integration

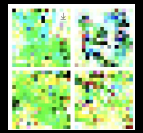
- aws
- Google Cloud
- Microsoft Azure
- NVIDIA
- kubernetes

Loss is not enough

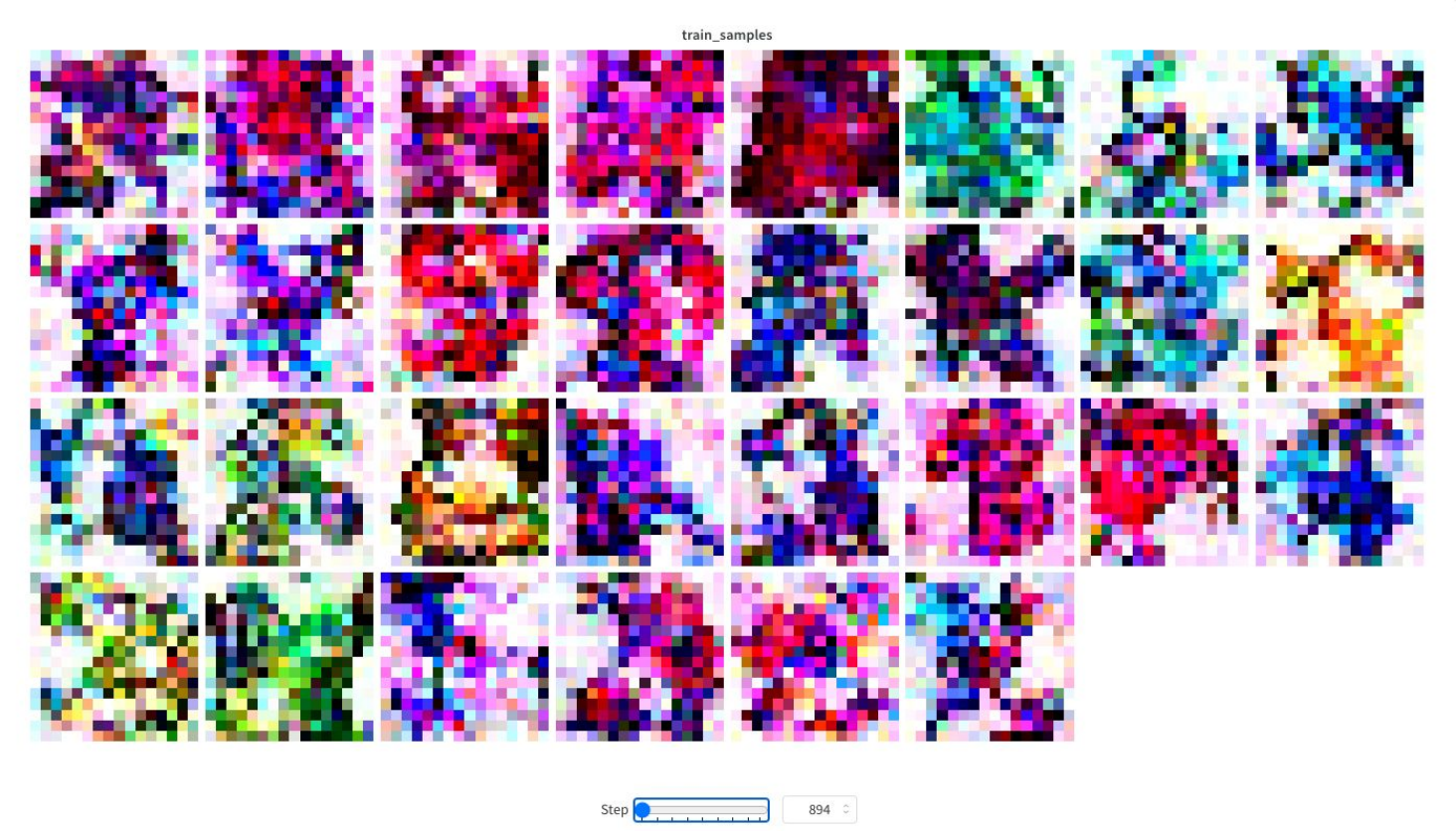


Loss is not enough

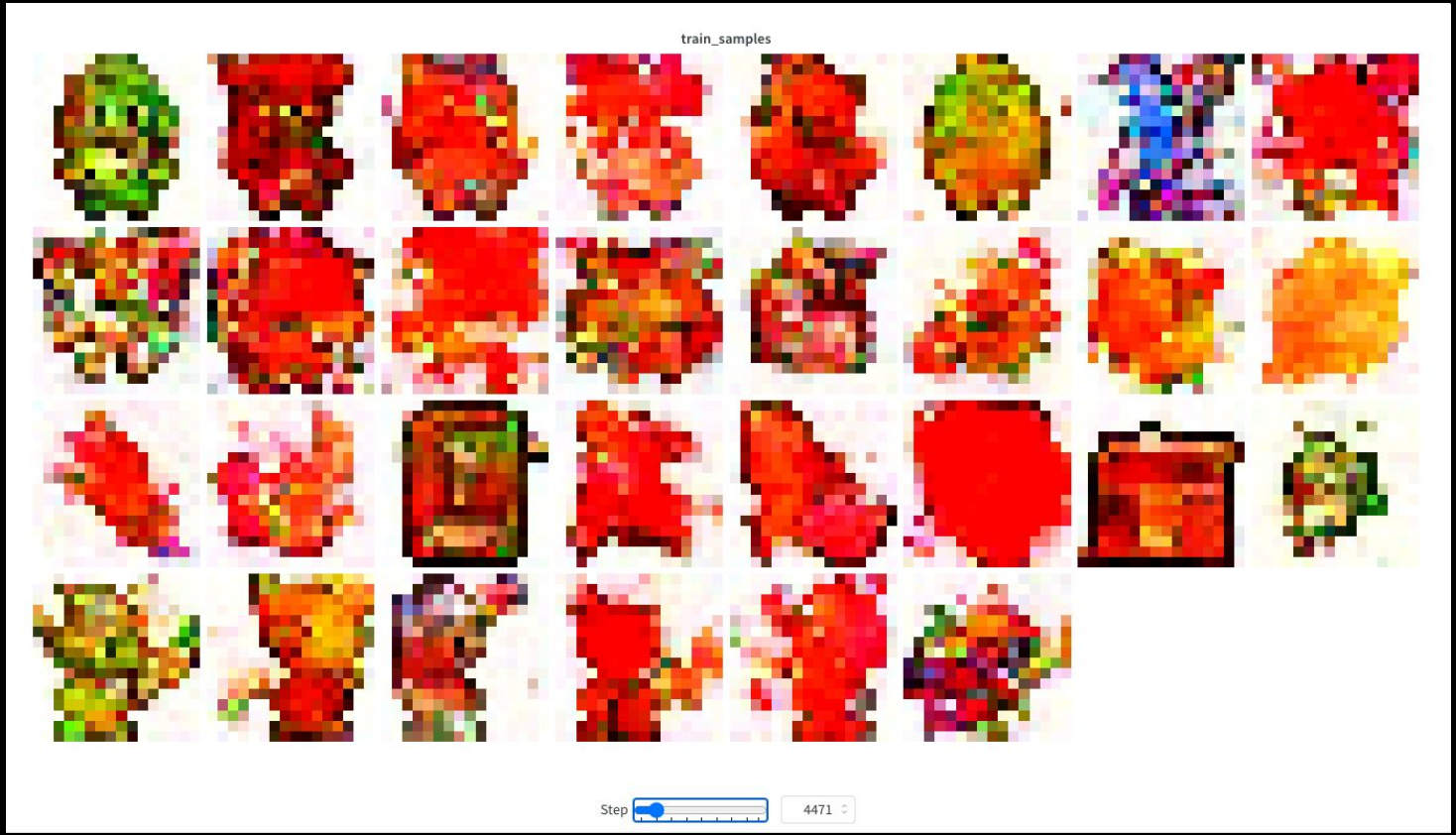




Which checkpoint would you choose?

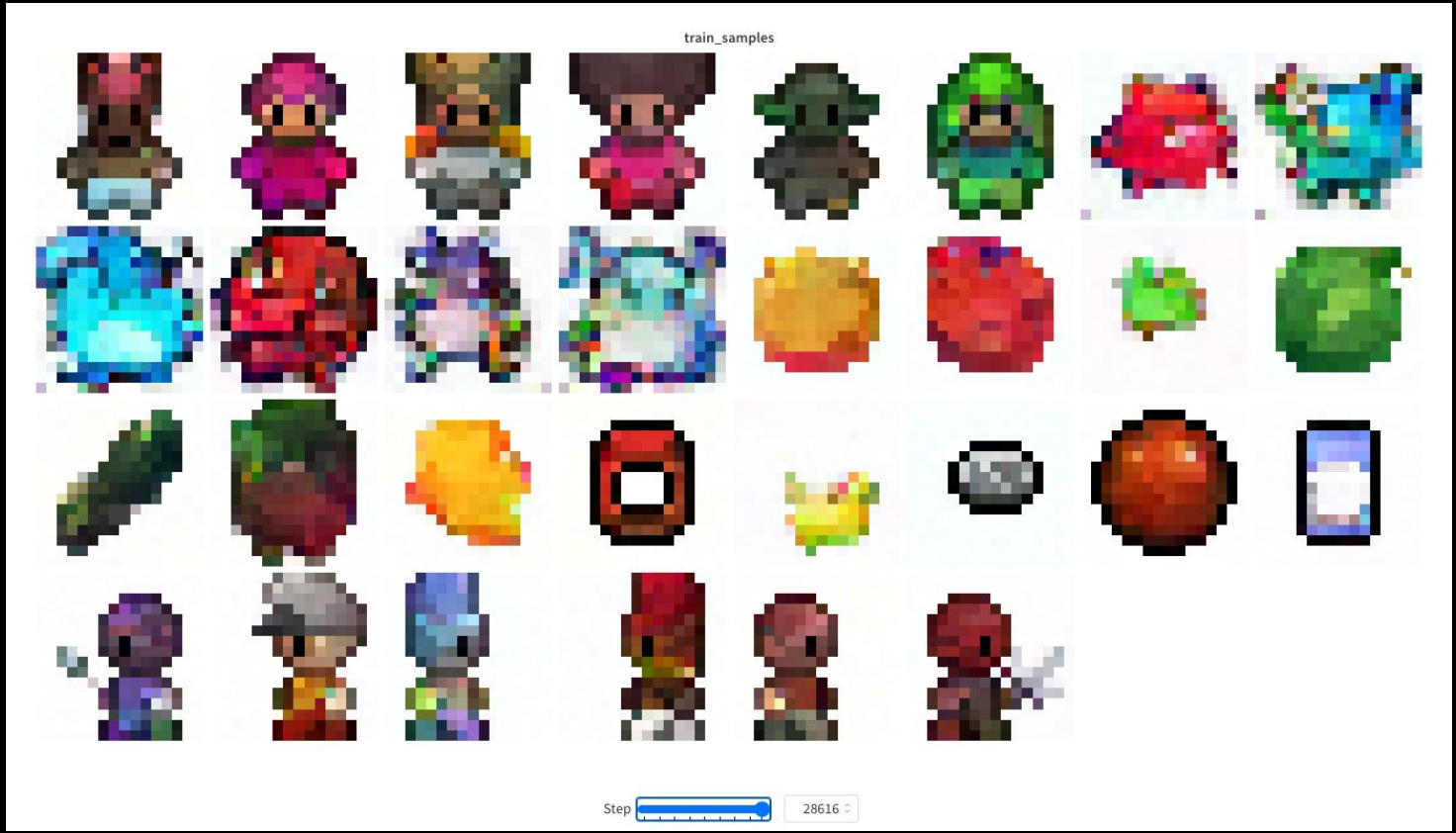


Which checkpoint would you choose?



Which checkpoint would you choose?

Which checkpoint would you choose?



Which sampler will you choose?

Tables 1

Add panel

```
runs.summary["samplers_table"]
```

	input_noise	ddpm	ddim	class
1				hero
2				hero
3				hero
4				hero
5				hero
6				hero
7				non-hero

1 - 7 of 30

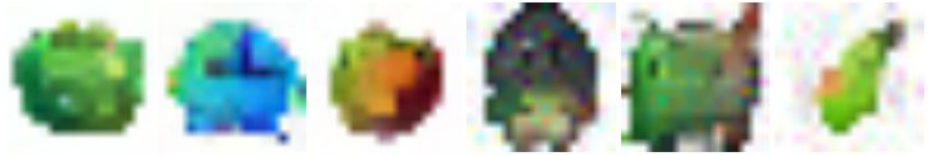
Export as CSV Columns... Reset Table

Which sampler will you choose?

Tables 1

Add panel

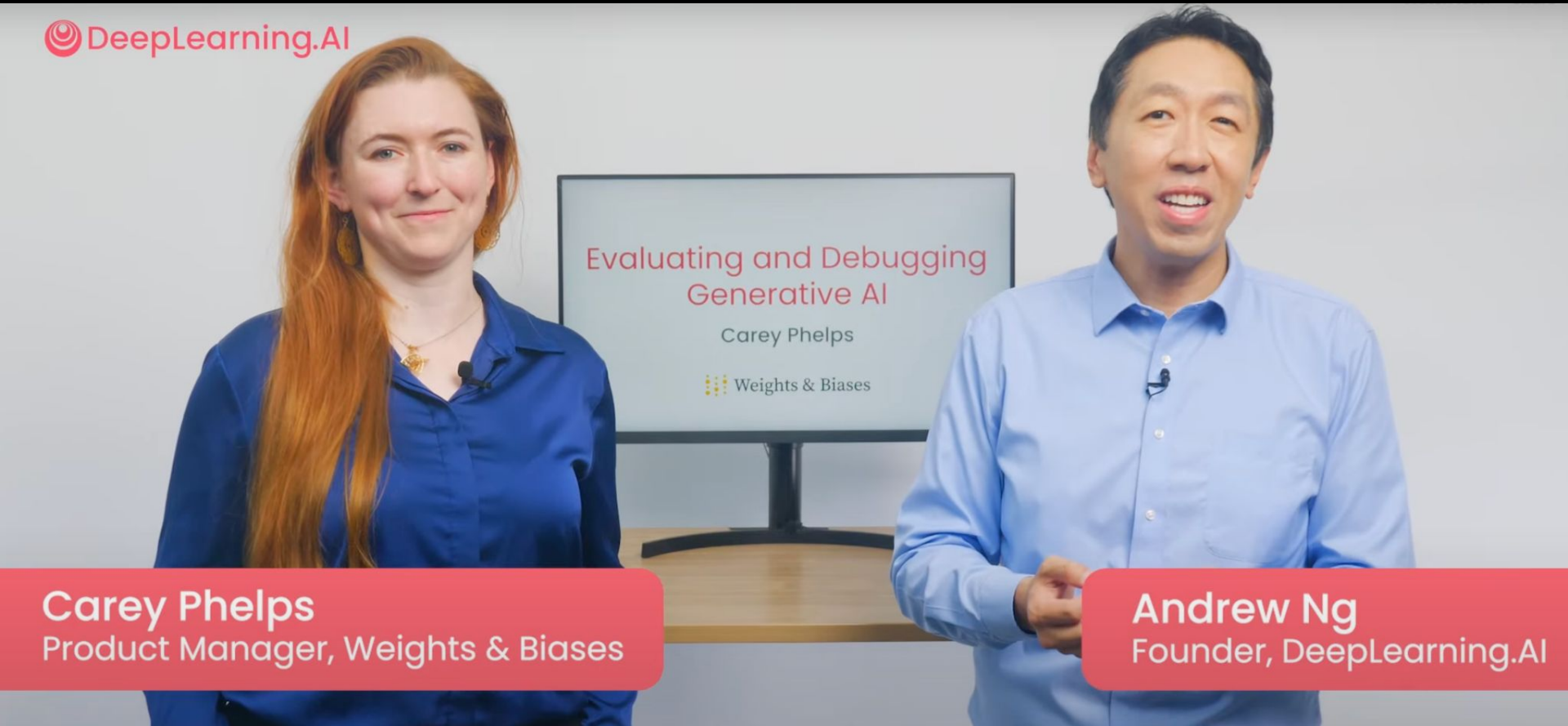
```
runs.summary["samplers_table"]
```

	class	ddpm	ddim
1	food	 1-3 of 6	1-3 of 6
2	hero	1-3 of 6	1-3 of 6
3	non-hero	1-3 of 6	1-3 of 6

1 - 3 of 5

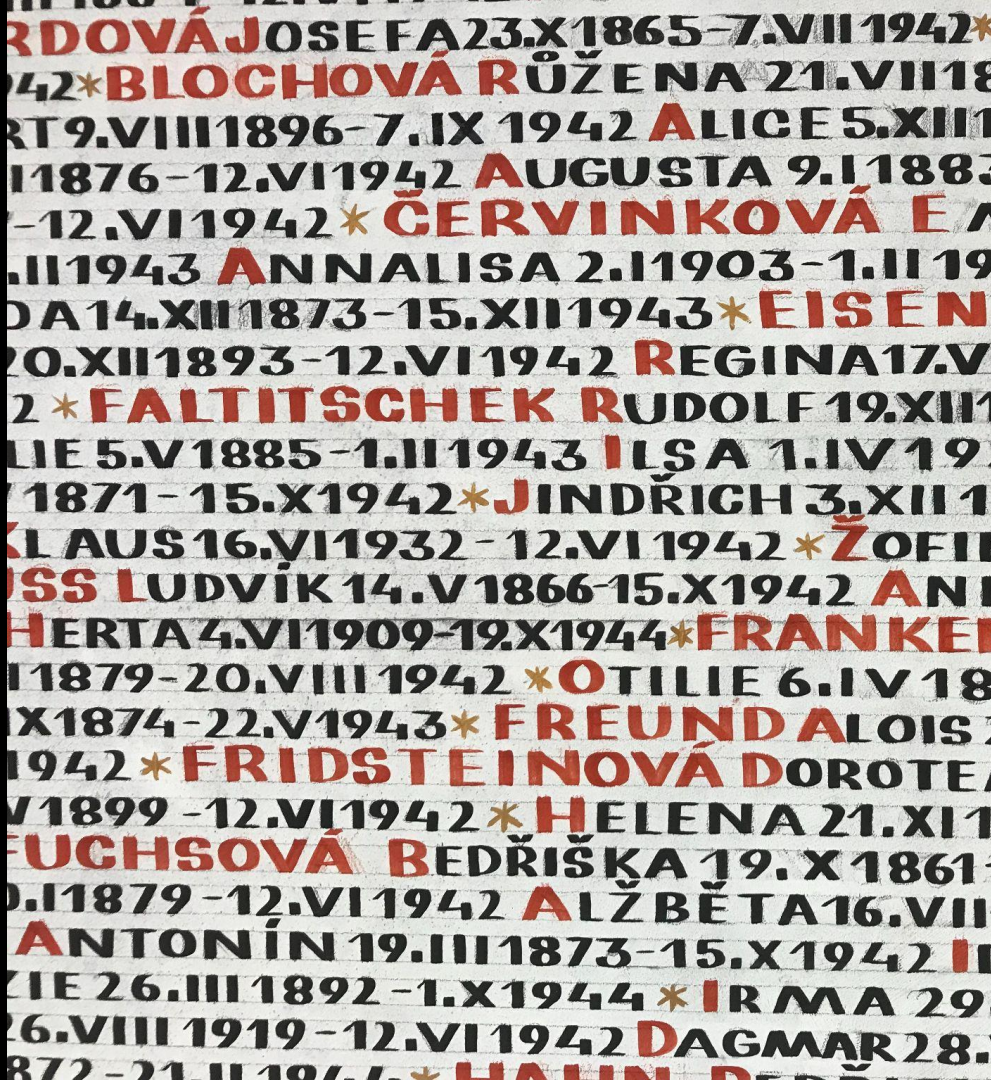
Export as CSV Columns... Reset Table

Learn more in our course -> http://wandb.me/DLcourse_wandb



Weights & Biases

W&B for LLMs



The world's leading ML teams use W&B

The LLM leaders use Weights & Biases

OpenAI
Meta
stability.ai
cohere
Google
ANTHROPIC
grammarly
NVIDIA
Adept
Character.AI
ALEPH ALPHA

Microsoft
amazon
Canva
Adobe
IKEA
TOYOTA RESEARCH INSTITUTE
gm
Ford
servicenow
salesforce
GRAPHCORE
mosaicML
ZOOX
nuro
woven planet
stripe
SAMSUNG
SONY
GONG
IBM
Booking.com
zalando
AIRBUS
JOHN DEERE
WAYVE
fiserv.
Square
Qualcomm
CISCO
Spotify
pandora
SAP
RIOT GAMES
EPIC GAMES
Continental
AIMOTIVE
Walt Disney PICTURES
Finicity
mastercard
APPLOVIN
GENERAL DYNAMICS
LOCKHEED MARTIN
Walmart
NORDSTROM
Wesfarmers
ZEBRA
Johnson & Johnson
magic leap
Fidelity
RBC
GCHQ
SECURITYSERVICE MI5
Sandia National Laboratories
Bristol Myers Squibb
Genentech
AstraZeneca
BlueCross BlueShield
covariant
iRobot

Large Model Training was our first use case with OpenAI

Weights & Biases
Software Development

"We [OpenAI] use W&B for all our model training" – Peter Welinder talking to Lukas Biewald on the GPT-4 launch. 🚀

😍 Love hearing how OpenAI uses W&B, and especially reports to collaborate 🤝 on their model training.

✅ Get started with W&B at https://bit.ly/3yGOAIf.
https://bit.ly/42iGCCA

share some training runs and stuff
like that.

Peter Welinder of OpenAI on How They Use Weights & Biases to Train Mod...
youtube.com

Different LLM Personas - one common theme

`llm creator`

Trains LLMs from scratch

`llm fine-tuner`

Adapts existing LLMs
to specific tasks

`llm user`

Uses LLM as API service
(“Prompt Engineer”)

How to pick the right LLM for your app?

Good preparation

Online coding test

SAT score

Perplexity

HumanEval

Benchmarks

Candidate calibration

Expert interview

On-the-job performance

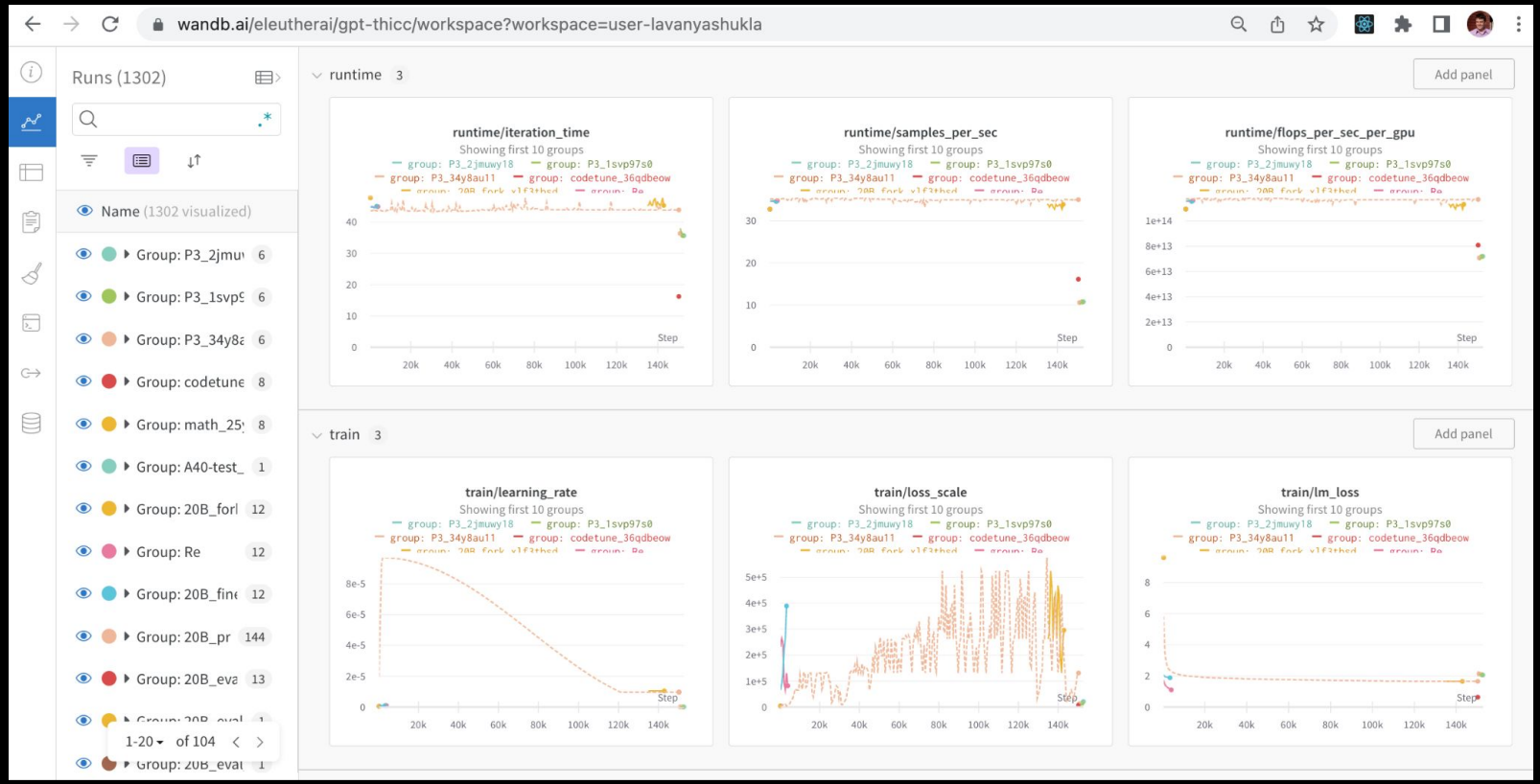
Chatbot Arena

Specific Use Cases

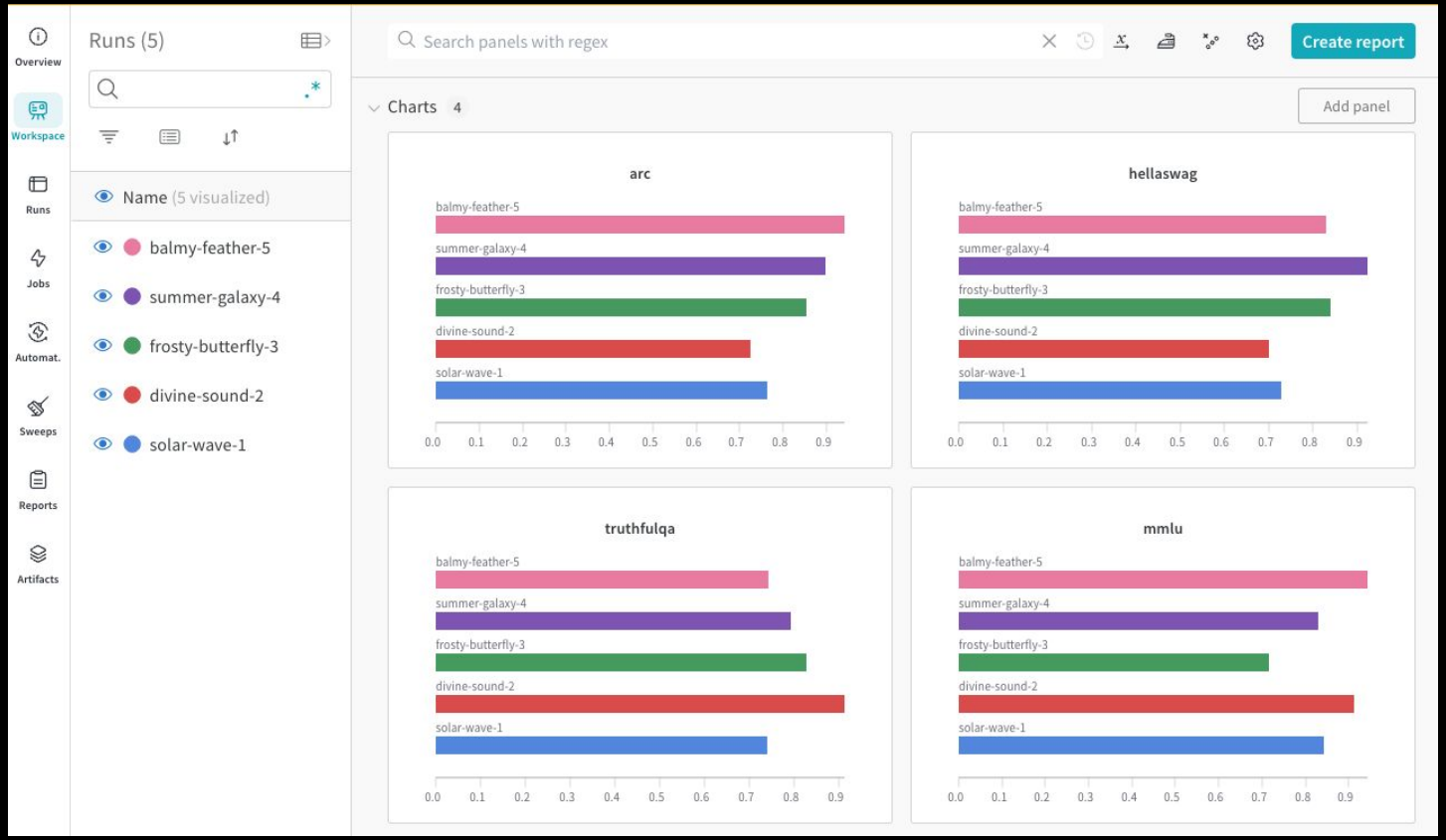
Evaluating LLM Apps

http://wandb.me/llm-apps-course

Going beyond Training & Evaluation Metrics



Compare models' performance on benchmarks



ARC
https://arxiv.org/pdf/2305.18354

Hellaswag
https://arxiv.org/abs/1905.07830

TruthfulQA
https://arxiv.org/abs/2109.07958

MMLU
https://arxiv.org/abs/2009.03300

Learn more at

http://wandb.me/llm-apps-course

Sample from your LLMs

```python
# create a wandb run and store experiment settings
# and hyperparameters in the config object
run = wandb.init(project="nlg", job_type="train")
wandb.config.update(...)

# [initialize models]

# create a Table with columns for the epoch index,
# the model loss, and the generated text
text_table = wandb.Table(columns=["epoch", "loss", "text"])

print("Training for %d epochs..." % args.n_epochs)
for epoch in tqdm(range(1, args.n_epochs + 1)):
   loss = train(*random_training_set(args.chunk_len, args.batch_size))
   loss_avg += loss
   # log loss to wandb
   wandb.log({"loss" : loss})
   if epoch % args.print_every == 0:
       print('[%s (%d %d%%) %.4f]' % (time_since(start), epoch,
                                      epoch / args.n_epochs * 100, loss))
       train_text = generate(decoder, 'Wh', 100, cuda=args.cuda)
       # log data row to Table
       text_table.add_data(epoch, loss, train_text)
       print(train_text, '\n')

# log the table to wandb
run.log({"training_samples" : text_table})
```

```python
runs.summary["responses_2K"]
```

	prompt	temperature	response
1	To be or not to be that is the	0.2	To be or not to be that is the sense the such a babest that the state to the world.
2	Shall I compare thee to a	0.2	Shall I compare thee to a man seem the way to the saves of the senate That the world to the state to the
3	O brave new world that hath such creatures in it! O	0.2	O brave new world that hath such creatures in it! O One the fair bear the way to the way, That the state to the tongue and the
4	Tomorrow and tomorrow and tomorrow	0.2	Tomorrow and tomorrow and tomorrow That the world that she was a beauty that this son.
5	And this our life, exempt from	0.2	And this our life, exempt from That she shall be straight to the said. KING EDWARD IV:

1 - 5 of 25

Export as CSV Columns... Reset Table

Collaborate with your stakeholders - W&B Reports

W&B Fully Connected > Articles > NLP

WandBot: GPT-4 Powered Chat Support

This article explores how we built a support bot, enriched with documentation, code, and blogs, to answer user questions with GPT-4, Langchain, and Weights & Biases.

Anish Shah, Bharat Ramanathan, Darek Kleczek, Morgan

Share Comment 9 stars

Last Updated: May 21, 2023

As large language models (LLMs) increasingly blur the line between software engineering and machine learning, it is crucial to explore and better understand this intersection so that we can build the best tools for this new cohort of ML practitioners entering the field.

To deeply understand these new workflows and pain points, we built our own support bot, which we've named WandBot. In this article, we'll discuss the implementation of a Q&A bot for Weights & Biases (W&B) using GPT-4, Langchain, OpenAI embeddings, and FAISS. If you'd like to take it for a spin, it's live on our Discord in the #wandbot channel. Let us know what you think!

We have also open-sourced the actual code we're using to run WandBot, you can find the Github repo here. We're hoping we can improve this tool together with the community.

In this article, you'll learn about the technical aspects of the code and updates made to the original implementation and how to create frontend applications for the Q&A bot on platforms like Discord and Slack, ensuring seamless integration and broader accessibility.

Table of Contents

Based on this process, we calculated 74% accuracy of model responses, as judged by W&B experts. While correlated, the model-based accuracy assessment had a lot of false negatives, and there is more work to ensure the automated (model-based) evaluation delivers good results.

Accuracy via Experts

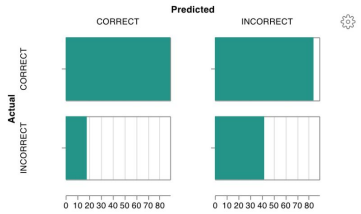
comfy-butterfly-43

0.7425

Accuracy via Model

comfy-butterfly-43

0.4592



`runs.summary["labeled_data"]`

	Question	Model Answer	Model Score	Human Score
1	How do I add computed columns to a wandb Table object?	To add computed columns to a wandb Table object, you can use the `add_computed_columns(func,	0	1
2	How can I log a table containing `wandb.Image` objects to WandB?	To log a table containing wandb.Image objects to WandB, you should create a `wandb.Table()` object with the	1	1
3	What wandb method can be used to add a column of data to a Table instance?	You can use the `add_column` method to add a new column of data to a Table instance in wandb. Here's an example:	1	1

1 - 3 of 233 Export as CSV Columns... Reset Table

Read about wandbot -> http://wandb.me/wandbot_report

Try it out on discord -> http://wandb.me/wandbot

Evaluate your LLM Apps

Method	Dataset	Evaluator	Metric
Vibes Check	Curated	Human Expert	Vibes
Model-based Eval	Curated or Generated	LLM	Accuracy
Unit Testing	Curated or Generated	LLM	Accuracy
A/B Testing in Production	Production	User	Business Metric

Log traces from your LLM chains

wandb.ai/prompt-eng/prompts-text-to-sql/runs/eydgtd6j?workspace=user-capecape

Overview
Charts
System
Logs
Files
Artifacts
Code

Search panels

Create report

	Succe	Timestamp	Input	Output	Chain	Error	Model
6	True	2023-04-20 13:36:41	**0.question:** Find the id of the first product we sent every customer	**0.sql_result:** [[1, 68383]]	SequentialCh LLMChain(Op	-	2dc0b03!
5	False	2023-04-20 13:17:11	**0.question:** Find the top 1 customer who has spent the most money		SequentialCh TransformCha	001003 (42000): SQL	cf5b65d5
4	False	2023-04-20 13:17:05	**0.question:** What is my best performing region		SequentialCh TransformCha	001003 (42000): SQL	cf5b65d5
3	False	2023-04-20 13:17:04	**0.question:** Get my last 10 orders		SequentialCh TransformCha	TypeError("'N object is not	cf5b65d5
2	False	2023-04-20 13:16:52	**0.question:** Find the top 10 customers who have spent the most monev		SequentialCh TransformCha	001003 (42000): SQL	cf5b65d5
1	False	2023-04-20 13:16:42	**0.question:** Find the id of the first product we sent every customer		SequentialCh TransformCha	001003 (42000): SQL	cf5b65d5

19 - 24 of 24

Export as CSV Columns... Reset Table

Trace Timeline (#1) Model Architecture (ID: cf5b65d5796eca30)

1: SequentialChain 10257ms
2: LLMChain 8680ms
3: OpenAI 8679ms
4: Transfor

SequentialChain 10257ms

Status Message 001003 (42000): SQL compilation error: syntax error line 1 at position 0 unexpected '.'.

Result Set 1

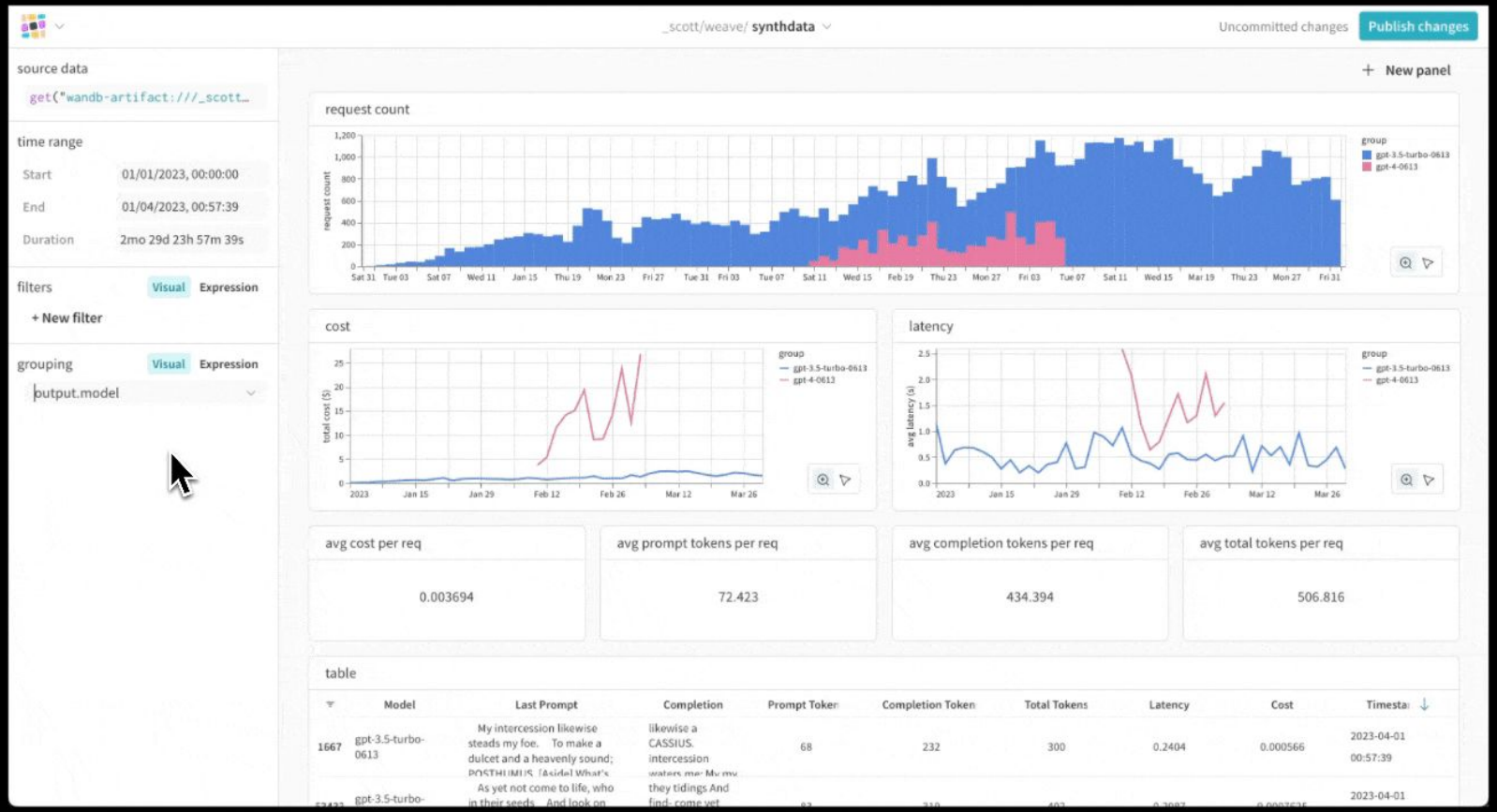
Inputs

question Find the id of the first product we sent every customer

schema_str [{"table": "TPCH_SF1.CUSTOMER", "C_CUSTKEY": "{\"type\":\"FIXED\",\"precision\":38,\"scale\":0,\"nullable\":false}", "C_NAME": "

runs.summary["text_to_sql"]

Sneak peak: LLM Monitoring



Weights & Biases

Stay in touch!

http://wandb.me/discord
http://wandb.me/llm-apps-course
http://wandb.me/DLcourse_wandb

/hans-ramsl

hans@wandb.com