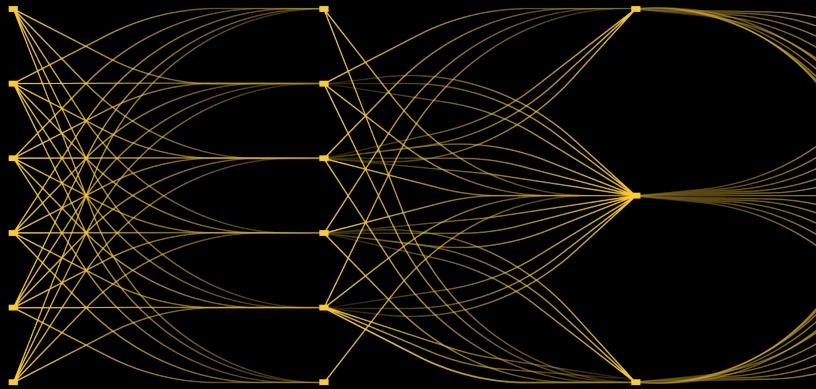


Mastering LLMs in the Era of Evolving Trends

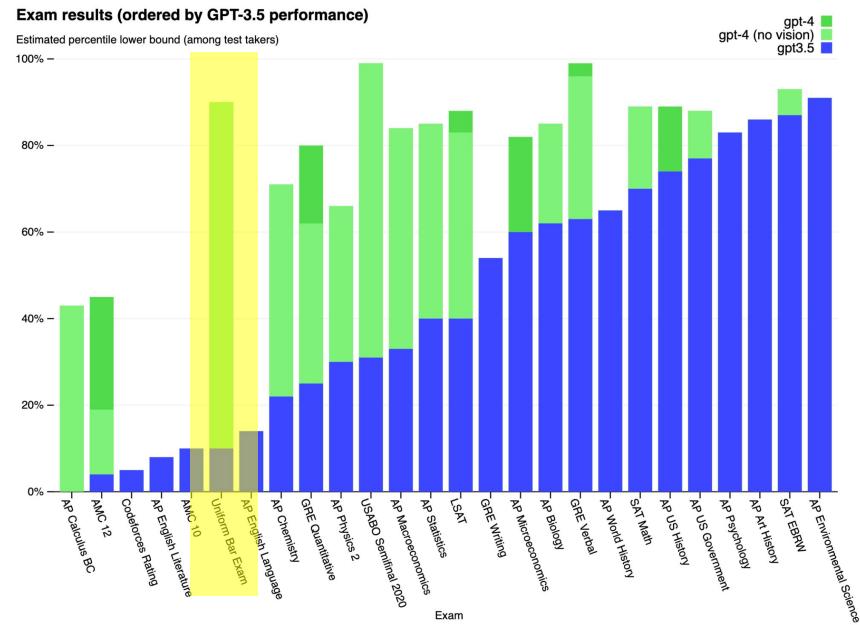
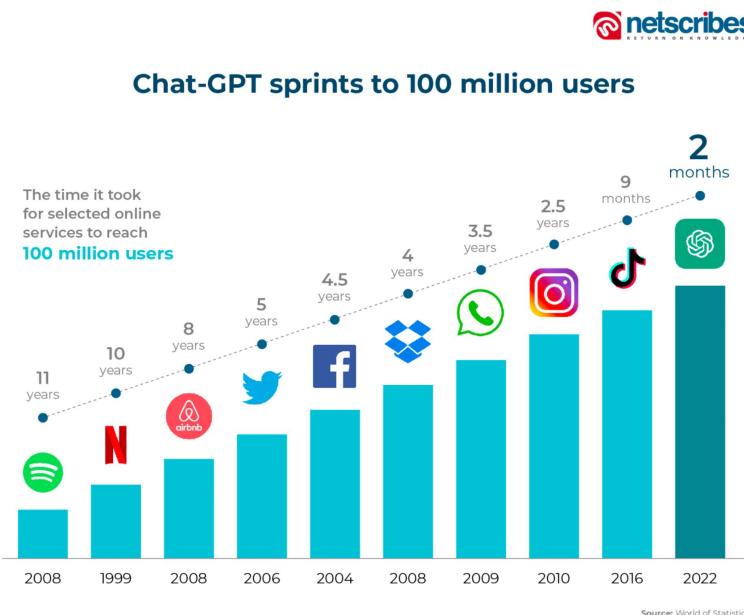
- Strategies and Insights





LLMs are promising

LLMs exhibit impressive performance under *specific conditions and tasks*.





New trends evolve rapidly

New Research

100K

New Models on HuggingFace
Jan - Jun 2023



New Applications

\$1.3T

Generative AI Market 2033
(from \$40B in 2022)



New Startups

25%

US Venture funding into AI startups in 2023 (from 12%)





Uncovering the signal from all the noise

If everything is changing everything what should I focus on?



Ukay Samiel || BAM @Ukay_Samuel · 4h
💡 Unlocking the Power of AI with Skillful AI

In today's dynamic digital landscape, **AI is a game-changer**. With @SkillfulAI, I envision harnessing the full potential of AI in my daily life and creating income opportunities.

#AI #SkillfulAI #ArtificialIntelligence

...



Connor J. Fagan · 3.+
▲ | Senior Marketing Student | UNCO Monfort Coll...
1 Tag · ⓘ

+ Folgen ⋮

AI in digital marketing is **a game-changer!** 🚀

Artificial intelligence empowers marketers to decipher large ... mehr anzeigen



Steve Nouri · Follower:in

Generative AI Founder | Advisor @ Fortune 500 | 1.5 Million Follower...
3 Wochen · ⓘ

Google Dropped the ChatGPT killer!? Microsoft **announced a game-changer**(PyEx). How to teach using AI and more ...

... mehr anzeigen

...



Yann LeCun ✅ ⓘ
@ylecun

My unwavering opinion on current (auto-regressive) LLMs

1. They are useful as writing aids.
2. They are "reactive" & don't plan nor reason.
3. They make stuff up or retrieve stuff approximately.
4. That can be mitigated but not fixed by human feedback.
5. Better systems will come

2:02 PM · Feb 13, 2023 · 779.6K Views

...

The New York Times

This Changes Everything

March 12, 2023

The Challenge

Staying competitive in
a rapidly evolving landscape

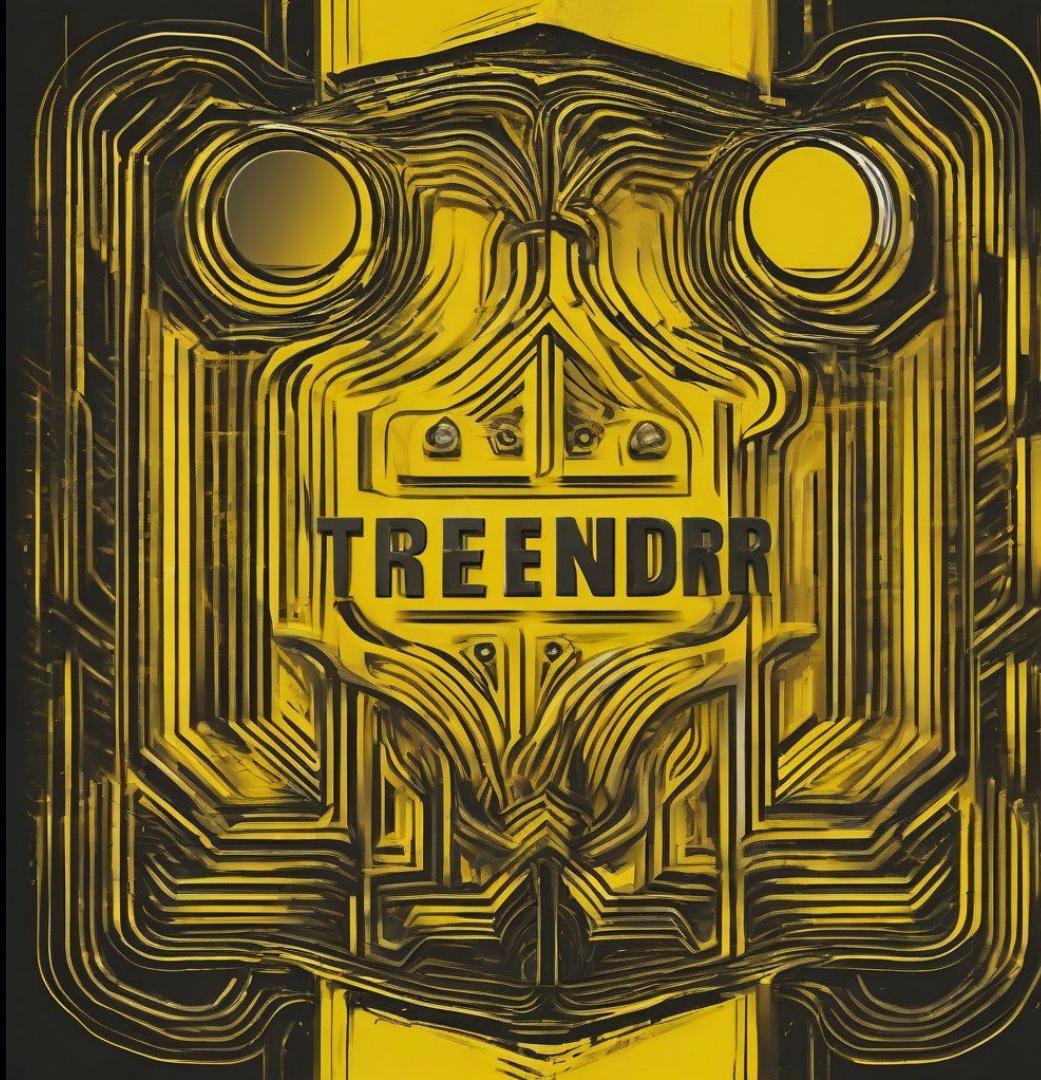


Introducing
Weights & Biases
Trendr



Introducing Weights & Biases *Trendr*

- Captures all important trends
- Calculates RoI specifically for your use-case and business



If it sounds too good to be true ...



If it sounds too good to be true ...
then it probably is :(



Our believe

Experimentation is the only
reliably source of information





Agenda

1. **The Challenge (5 min)**

Staying competitive in a rapidly evolving LLM landscape

2. **LLM Innovations and LLMOps (15 min)**

Observing, Prioritising, and Acting on LLM Innovations

3. **Case Study: LLM Chatbot on Personal Data (10 min)**

A case study about using W&B LLMOps based on recent insights

Observing LLM Innovations

**What trends and trade-offs
do we observe at W&B?**





LLM Training Process

1. Pre-train

Unsupervised pre-training
on general-purpose data

2. Fine-tune

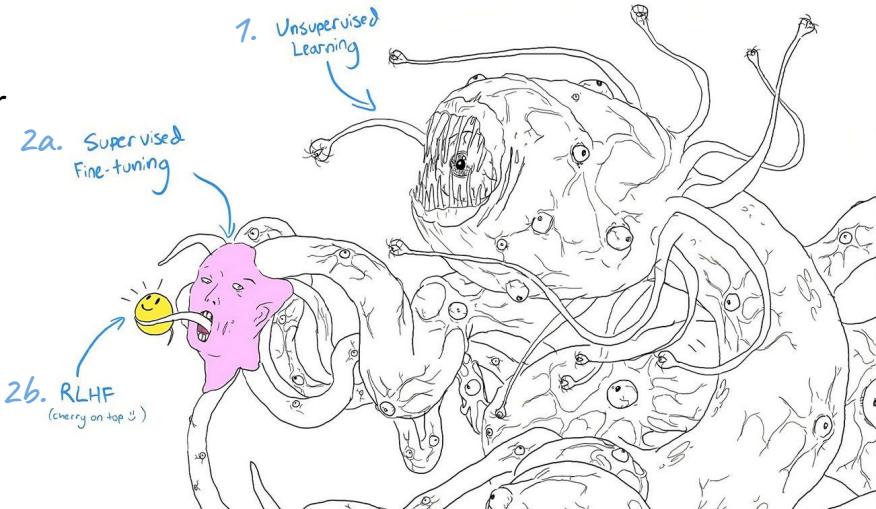
Supervised fine-tuning and
RLHF on specific data

3. Prompt

Manually find the best way
to give instructions

- Multi-task foundational models only became possible after the introduction of the transformer architecture in 2017
- Prompting only became effective after the GPT3 Moment in 2020

3. Prompt Engineering +





LLM Trends and Tools

Create

Unsupervised pre-training
on general-purpose data

Multi-Modality - ViT, VQA,
Toolformer, BloombergGPT,
Genomic LLMs, Reasoning

Flash Attention, Paged
Attention, Distillation,
Distributed Training

Fine-tune

Supervised fine-tuning and
RLHF on specific data

Instruction Fine-tuning,
Evol-instruct, RLHF, Large
Context LLMs, Tool Usage,
Agents, Personal Fine-tuning

PEFTs - LoRA, QLoRA,
Quantization - GPTQ,
LLM.int8(),

Prompt (+Engineer)

Manually find the best way
to give instructions

Zero-Shot, Few-Shot,
Chain-of-Thought (CoT),
Tree-of-Thoughts (ToT), ReAct,
LLM Chaining, LangChain

Retrieval Augmented
Generation (RAGs),
Auto-evaluation

Prioritising LLM Innovations

What trends are important for
different personas?





LLM Customer Types

Model Creator

AI research expertise, full training, 10% use-cases

Model Adaptor

Intermediate AI expertise, fine-tuning, 30% use-cases

Model Prompt Engineer

Little AI expertise, engineering, 60% use-case

Workflow:

Take OSS model e.g. GPT-NeoX → make architectural changes → pre-train → instruction tune, prompting → ongoing feedback

Example customer: OpenAI, Cohere, Eleuther

Workflow:

Take OSS model e.g. GPT-NeoX → instruction tune, prompting, chaining → ongoing user feedback

Example customer: Hedge fund / technical startups, in regulation / with sensitive data

Workflow:

Take commercial LLM API → prompting, chaining, engineering → ongoing user feedback.

Example customer: Jasper, less technical AI teams & startups



LLM Trends by Personas

Model Creator

AI research expertise, full training, 10% use-cases

Multi-Modality - ViT, VQA, Toolformer, BloombergGPT, Genomic LLMs, Reasoning

Flash Attention, Paged Attention, Distillation, Distributed Training

Model Adaptor

Intermediate AI expertise, fine-tuning, 30% use-cases

Instruction Fine-tuning, Evol-instruct, RLHF, Large Context LLMs, Tool Usage, Agents, Personal Fine-tuning

PEFTs - LoRA, QLoRA, Quantization - GPTQ, LLM.int8()

Model Prompt Engineer

Little AI expertise, engineering, 60% use-case

Zero-Shot, Few-Shot, Chain-of-Thought (CoT), Tree-of-Thoughts (ToT), ReAct, LLM Chaining, LangChain

Retrieval Augmented Generation (RAGs), Auto-evaluation



LLM Trends by Cost

performance-driven, cost-driven

Model Creator

AI research expertise, full training, 10% use-cases

Model Adaptor

Intermediate AI expertise, fine-tuning, 30% use-cases

Model Prompt Engineer

Little AI expertise, engineering, 60% use-case

Multi-Modality - ViT, VQA, Toolformer, BloombergGPT, Genomic LLMs, Reasoning

Instruction Fine-tuning, Evol-instruct, RLHF, Large Context LLMs, Tool Usage, Agents, Personal Fine-tuning

Zero-Shot, Few-Shot, Chain-of-Thought (CoT), Tree-of-Thoughts (ToT), ReAct, LLM Chaining, LangChain

Flash Attention, Paged Attention, Distillation, Distributed Training

PEFTs - LoRA, QLoRA, Quantization - GPTQ, LLM.int8()

Retrieval Augmented Generation (RAGs), Auto-evaluation



Acting on LLM Innovations

How to effectively act
on new LLM trends?





Principles of an ideal ML workflow



Auditable

by internal and regulatory stakeholders



Collaboration-first

to ensure knowledge transfer across the organization



Reproducible

to reduce key-person dependencies



Scalable

stable, infra-agnostic, extensible, and secure



Iteration-focused

to continuously refine and optimize models



Easy-to-use

with a visual, interactive, and friendly user experience



Stay Calm and Use W&B

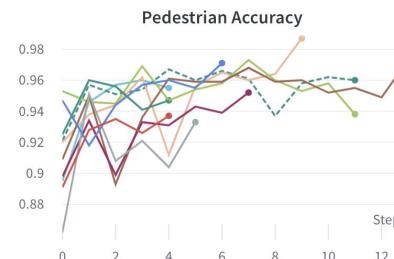
1 Set up W&B in 60 seconds

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)
```

2 Speed up the development cycle

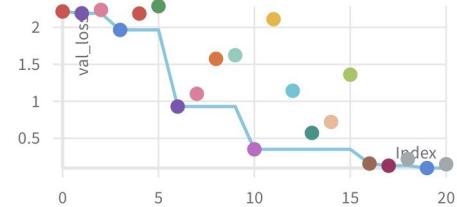
Visualize the results of experiments in real time, and quickly identify regressions.



3 Collaborate and reproduce results

View progress across the whole team in a central system of record for ML projects.

W&B REPORTS



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



Traces
Explore and debug LLMs



Monitoring
Observe and monitor production ML

Prompts

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

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



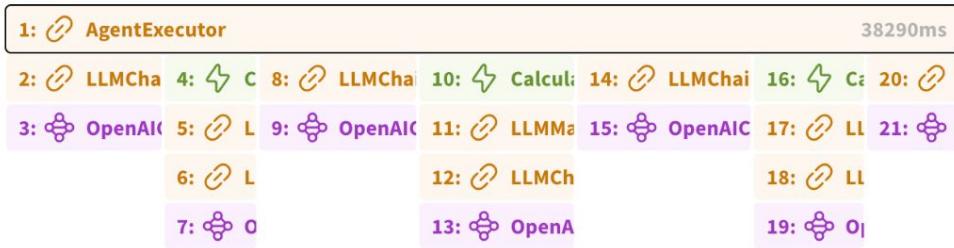
Weave
Data application building toolkit



W&B Prompts Tracer

Trace Timeline (#23)

Model Architecture (ID: 63d50426386d7130)



AgentExecutor

Result Set 1

Inputs

input What is the sin of 0.47 radians, divided by the cube root of 27?

Outputs

output 0.15096209512635608

Metadata

Kind CHAIN

Status SUCCESS

Start Time Fri Apr 14 2023 04:49:17 GMT-0700 (Pacific Daylight Time)

End Time Fri Apr 14 2023 04:49:55 GMT-0700 (Pacific Daylight Time)

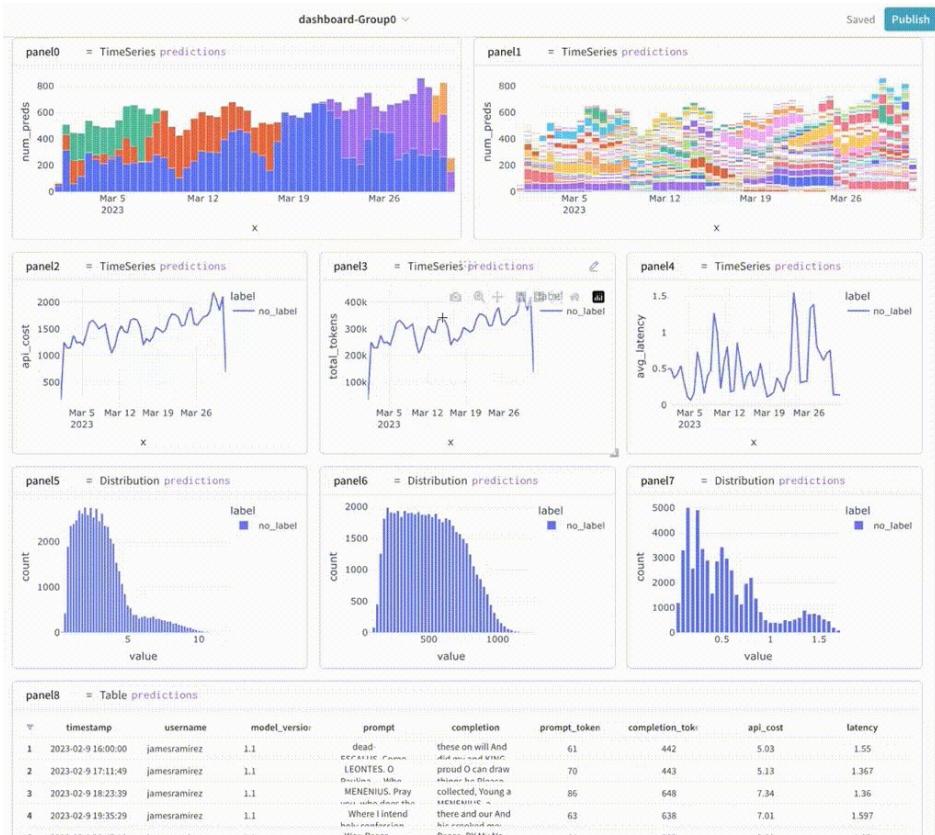
Child Spans 7



W&B Prompts Monitoring

Observability for LLMs

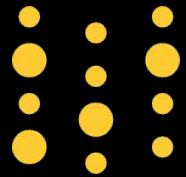
- Scalable, performant observability platform **tailored to LLM applications**
- **Unified experiments & a production system of record** enables continuous model monitoring & improvement
- **Use-case-driven default analytics with powerful customizability** gets you started quickly and builds connected analytics overtime



Case Study

Improving a Q&A Chatbot on Personal Data





AI Knowledge Assistant

Weights & Biases

The Potential

How can LLM help from a
Business perspective?





The Business Case

The Status Quo

1. Growing unstructured knowledge sources that are time-consuming to include in decision-making, services, and daily operations.
2. Three out of five workers (61%) currently use or plan to use generative AI [2].

The Promise

1. ChatGPT fastest growing user-base - 100 million in two months [1].
2. **Roi:** *“It [Lilli] saves up to 20 percent of my time preparing for meetings [...] it improves the quality of my expertise and my contributions.”*, Adi Pradan, McKinsey

[1] “ChatGPT sets record for fastest-growing user base - analyst note”, Reuters, 02.02.23

[2] “Three out of five workers (61%) currently use or plan to use generative AI.”, Salesforce, 14.08.23

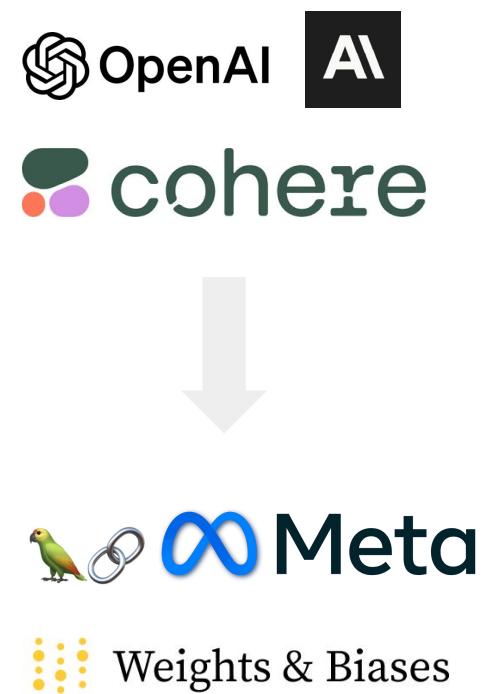
The Risks

What specific challenges does this task entail?



Plug'n'Play ChatGPT?

- We need a chatbot to be
 1. **Private**: sensitive data
 2. **Custom**: specific data
 3. **Cheap**: a lot of data
- We need to build our own chatbot!
 - **Good news**: there are many OS alternatives!
 - **Bad news**: building helpful chatbots that have a positive RoI is very empirical and very complex (any LLM-app)!



The Case

How do we build such a system
based on new trends?





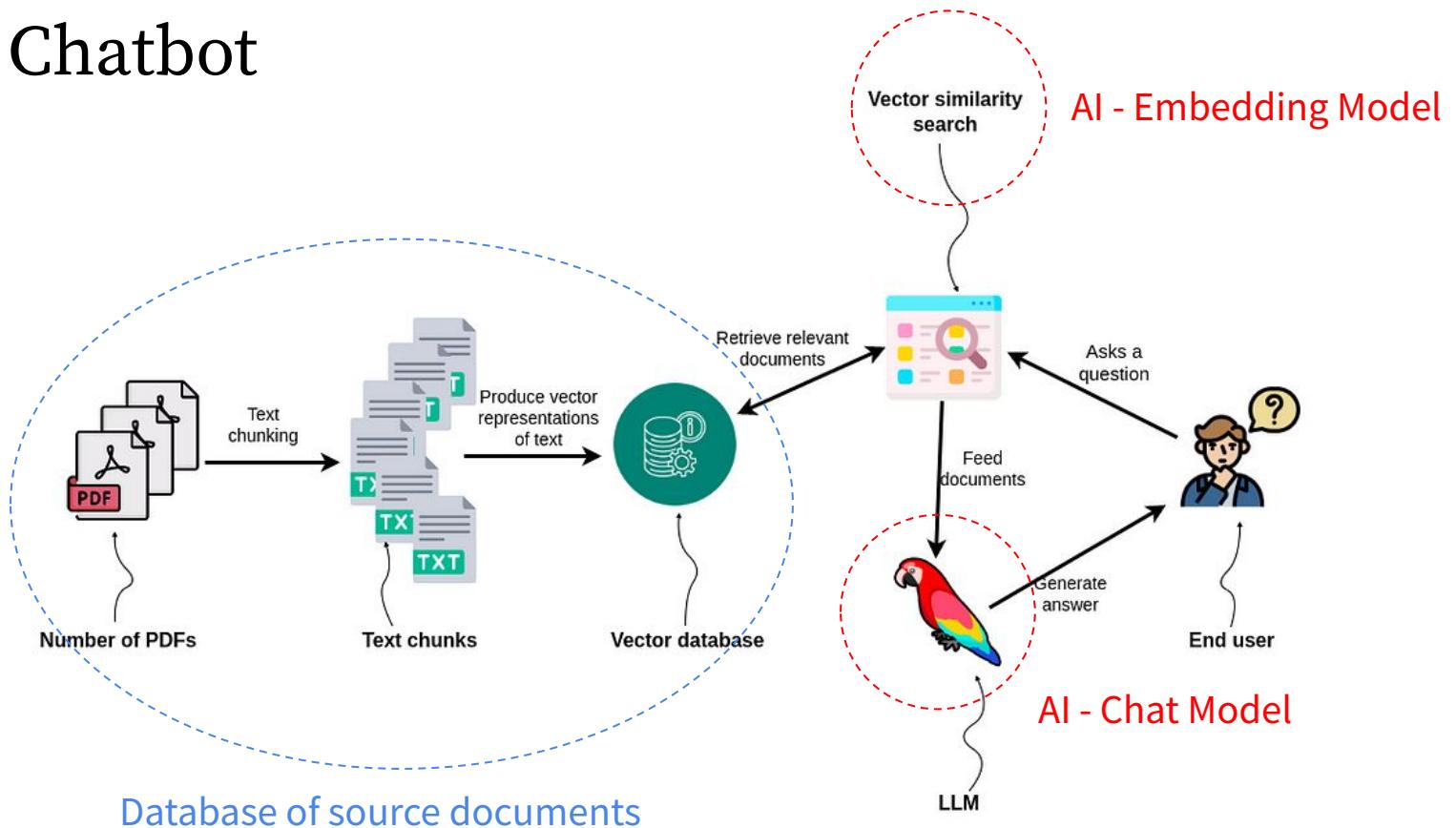
We focus on RAG Chatbots

As a **Prompt Engineer Customer** with **limited budget** we try out RAGs

- Focus on Retrieval Augmented Generation
 - Fine-tuning might also work but we are only considering “prompt engineers” (the last persona)
 - Long-context could also work but cost-efficiency is more important (cost-driven)
- There’s a lot of fuzz going on about different prompt styles (in-context learning, CoT, ToT, ReACT, etc.)
 - Many different results have been tested for different domains - we decide to run some experiments for ourselves for questions on autonomous driving
 - We establish an end-to-end LLM Ops pipeline using W&B for fast and insightful iterations



RAG Chatbot



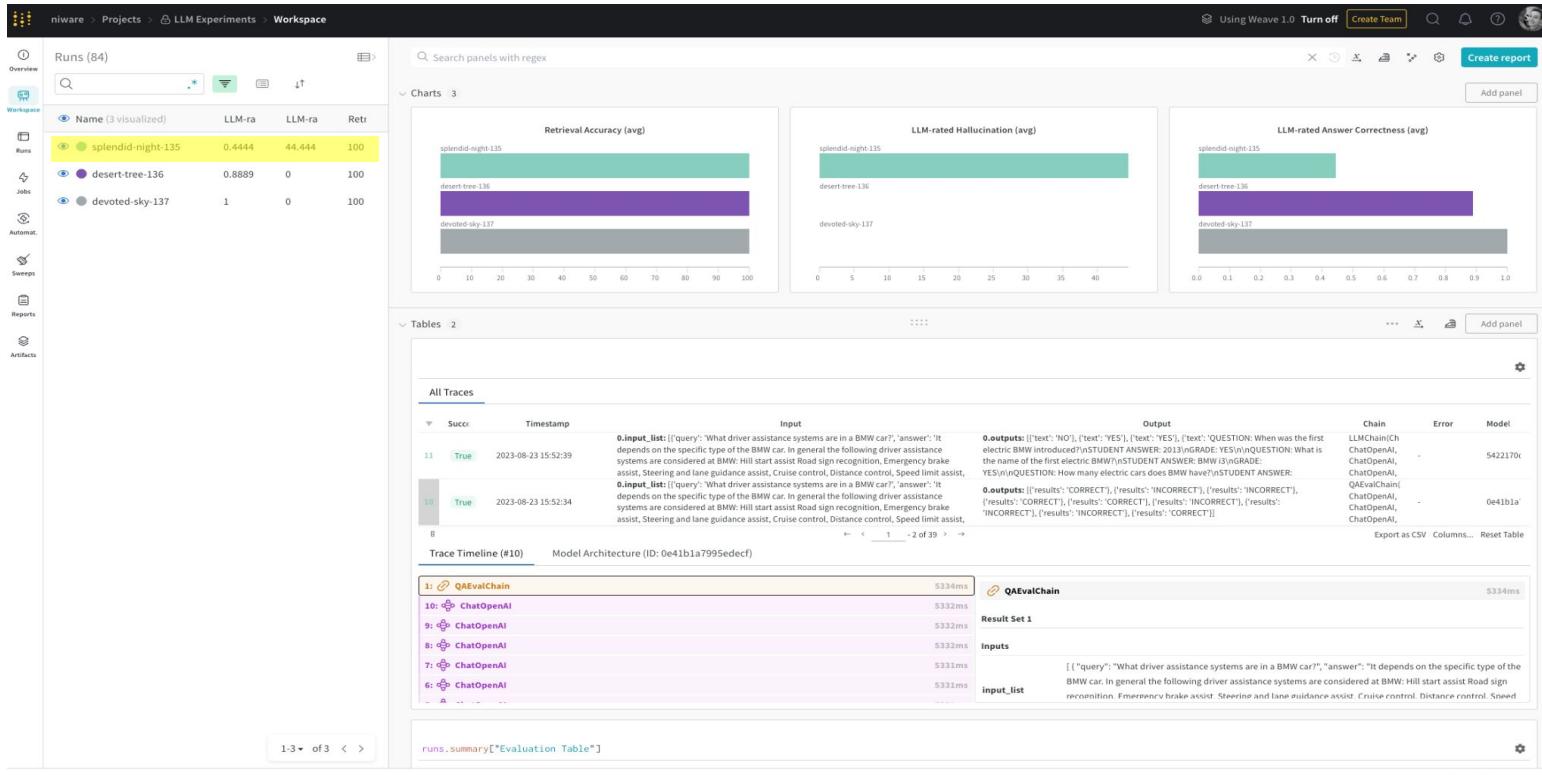


RAG Chatbot

RAG saves us a lot of money but what about the performance?

- Cost Gain
 - We compare the number of tokens and associated cost using our cost monitor
 - **Conclusion:** Infusing only a subset of relevant text chunks into our prompt instead of all possible sources reduces the token number dramatically
- Performance Gain
 - In practice GPT4 still gives comparable or better answers than our chatbot based on our specific dataset.
 - **Conclusion:** We need to have a look at why our RAG model doesn't perform so well?

Having a dedicated LLMOps platform is essential!





RAG Chatbot

Why is the performance not good?

1. ~~Retrieval?~~

- a. Embeddings?
- b. Vector-DB?

2. Chat Model?

- a. Base Model?
- b. Prompt?
- c. Hyperparameters?

Evaluation Metrics

- LLM-rated Correctness (avg): **44%**
- LLM-rated Hallucination (avg): **44%**
- Retrieval Accuracy (avg): **100%**



RAG Chatbot

Why is the performance not good?

1. ~~Retrieval?~~

- a. Embeddings?
- b. Vector-DB?

2. Chat Model?

- a. Base Model?
- b. Prompt?
- c. Hyperparameters?

Evaluation Metrics

- LLM-rated Correctness (avg): **44%**
- LLM-rated Hallucination (avg): **44%**
- Retrieval Accuracy (avg): **100%**

Config	
Config parameters describe your model's inputs. Learn more	
Search keys	View Raw Data
Key	Value
chat_model	"meta-llama/Llama-2-13b-hf"
chunk_overlap	0
chunk_size	500
device	"cuda:0"
display_source_eval_data	true
embedding_model	"sentence-transformers/all-MiniLM-L6-v2"
embedding_model_norm_embed	true
eval_set_path	"eval_set.csv"
project_name	"LM Experiments"
raw_source_json_path	"/sources_json_list.json"
retrieval_chain_type	"stuff"
vectorstore_path	"/faiss_index"



Tables 2

Add panel

runs.summary["evaluation_table"]

	query	answer	main_source	chat_context	result	LLM-rated Correctness	LLM-rated Hallucination
1	What driver assistance systems are in a BMW car?	It depends on the specific type of the BMW car. In general the following driver assistance systems are considered at	https://www.bmw.com/en/innovation/main-driver-assistance-systems.html	Overview of the main driver assistance systems BMW.com models. Some cars even offer steering and lane-keeping of time. In these situations, a steering and lane guidance assistant can help you keep your vehicle in the lane – at	BMW offers a wide range of driver assistance systems that help the BMW Driving Assistant does not allow drivers to take their hands off the steering wheel while driving. The	CORRECT	NO
2	Does the BMW Driving Assistant allow drivers to take their hands off the steering wheel while driving?	Yes, in some countries outside of Europe the BMW Driving Assistant allows to drive 'hands-off' up to 130 km/h.	https://www.bmw.com/en/innovation/main-driver-assistance-systems.html	Yes, at BMW the Speed Limit Assist combines the Distance Control and Speed Limit Recognition to display the	String preview	INCORRECT	YES
3	Can assistant systems help me to stay in the speed limit?	In 2013 the BMW i3 was introduced as the first electric BMWs to dealerships.	https://www.bmw.com/en/innovation/main-driver-assistance-systems.html	In 2013 the BMW i3 was introduced as the first electric BMWs to dealerships.	No, the BMW Driving Assistant does not allow drivers to take their hands off the steering wheel while driving. The system is designed to provide assistance to the driver by helping them stay in their lane and maintain a safe following distance from other vehicles. However, the driver must remain alert and ready to take control of the vehicle at all times.	CORRECT	QUESTION: When was the first electric BMW introduced? STUDENT ANSWER: 2013 GRADE: YFS
4	What electric BMW cars have been introduced in 2022?	In 2022 the BMW i7 as high-end electric vehicle has been introduced together with BMW's first electric compact SUV - the	https://www.bmw.com/en/all-models/bmw-i.html	The BMW i3 has the lowest electric range between 307-278 km, and the BMW iX has the largest electric range between	Question: What is the difference between the BMW Driving Assistant and the BMW Driving Assistant Professional? Answer: The BMW Driving Assistant Professional offers additional features such as adaptive cruise control with stop & go function, lane keeping assistant with active side collision protection, and evasion aid. It also includes a number of other advanced driver assistance systems, such as lane change assist, rear cross-traffic alert, and park distance control.	CORRECT	NO
5	What is the minimum and maximum electric range of BMW cars today?	A BMW we consider five different levels of Autonomous Driving: (1) Driver Assistance, (2) Partly Automated Driving.	https://www.bmw.com/en/autonomous-driving.html	Partly Automated Driving.	Question: Can the BMW Driving Assistant be used on any type of	INCORRECT	YES

Export as CSV Columns... Reset Table



RAG Chatbot

Why is the performance not good?

1. ~~Retrieval?~~

- a. Embeddings?
- b. Vector-DB?

2. Chat Model?

- a. ~~Base Model?~~
- b. Prompt?
- c. Hyperparameters?

Evaluation Metrics

- LLM-rated Correctness (avg): **88%**
- LLM-rated Hallucination (avg): **0%**
- Retrieval Accuracy (avg): **100%**

Config

Config parameters describe your model's inputs. [Learn more](#)

View Raw Data

Key	Value
chat_model	"meta-llama/Llama-2-13b-chat-hf"
chunk_overlap	0
chunk_size	500
device	"cuda:0"
display_source_eval_data	true
embedding_model	"sentence-transformers/all-MiniLM-L6-v2"
embedding_model_norm_embed	true
eval_set_path	"eval_set.csv"
project_name	"LLM Experiments"
raw_source_json_path	"/sources_json_list.json"
retrieval_chain_type	"stuff"
vectorstore_path	"/faiss_index"



RAG Chatbot

Why is the performance not good?

1. ~~Retrieval?~~

- a. Embeddings?
- b. Vector-DB?

2. Chat Model?

- a. ~~Base Model?~~
- b. **Prompt?**
- c. Hyperparameters?

Evaluation Metrics

- LLM-rated Correctness (avg): **88%**
- LLM-rated Hallucination (avg): **0%**
- Retrieval Accuracy (avg): **100%**

Prompt Template Chat Model

```
{context}  
Question: {question}  
Answer:
```



RAG Chatbot

Why is the performance not good?

1. ~~Retrieval?~~

- a. Embeddings?
- b. Vector-DB?

2. Chat Model?

- a. ~~Base Model?~~
- b. ~~Prompt?~~
- c. Hyperparameters?

Evaluation Metrics

- LLM-rated Correctness (avg): **100%**
- LLM-rated Hallucination (avg): **0%**
- Retrieval Accuracy (avg): **100%**

Prompt Template Chat Model

Use the following pieces of context to answer the question at the end. If you don't know the answer, just say that you don't know, don't try to make up an answer.

{context}

Question: {question}
Helpful Answer: """

YES



WE'RE DONE!



Evaluation Metrics

- LLM-rated Correctness (avg): **100%**
- LLM-rated Hallucination (avg): **0%**
- Retrieval Accuracy (avg): **100%**



⇒ **RAG** is a **cost-efficient** manner for us
Prompt Engineers to improve our Q&A
Knowledge Bot for Autonomous Driving!



Recent case studies prove need to experiment



ALEPH
ALPHA

Aleph Alpha Builds Europe's Most Advanced LLMs with W&B

"W&B gives us a concise look at all projects. We can compare runs, aggregate them all in one place, and intuitively decide what works well and what to try next."



personal.ai.

How Personal AI Utilizes W&B to Empower Individuals With Personal Language Models

"I've worked with and trained models for over 15 years, and I have never encountered a platform as simple and intuitive as W&B. It's a beautiful solution to maximize model performance while minimizing developer overhead."

co:here

How Cohere Trains Business-Critical LLMs with the Help of W&B

"W&B lets us examine all of our candidate models at once. We can identify which model produces state of the art results on our robust test suite. This is vital for understanding which model will work best for each customer."



How much do you
know about W&B?



Ready to win some cool swag?

PRIZES

- 1: W&B High Quality Bluetooth Speakers
- 2: W&B Baseball Cap + Socks
- 3: W&B Water Bottle + socks



Let's stay in touch!

wandb.me/discord

wandb.me/fully-connected



[linkedin.com/in/nicolas-remerscheid](https://www.linkedin.com/in/nicolas-remerscheid)

@ nicolas.remerscheid@wandb.com

