



Google Developer Groups
Cloud • Zürich

Meetup #21

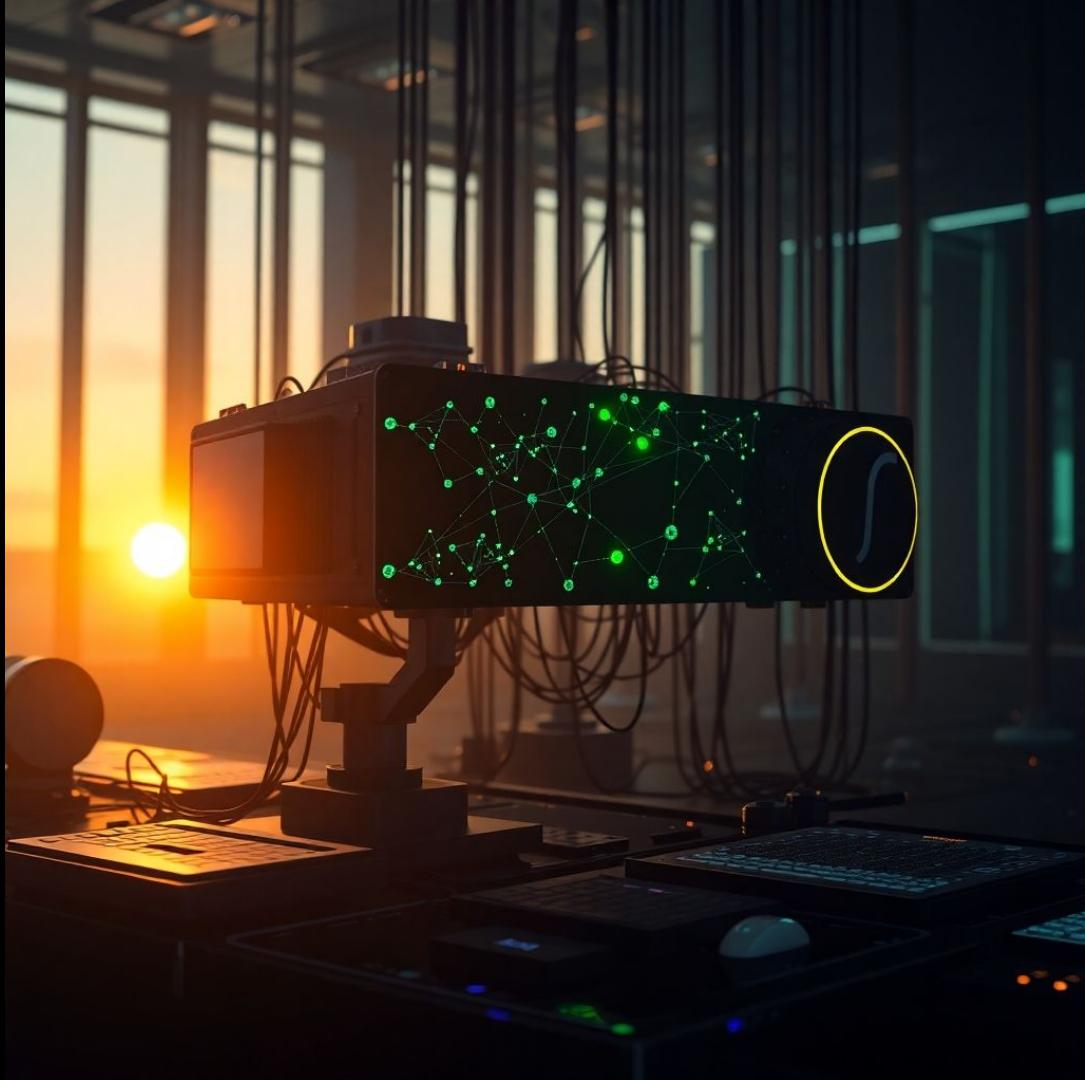
20.02.2025

Grüezi & Welcome

Fine-Tuning Open-Source LLMs for Real-World Apps

How to **choose, customise, and evaluate** LLMs in diverse use cases?

Customizing an LLM to fit your use-case, aligning it with desired outcomes, and **making it fit**.



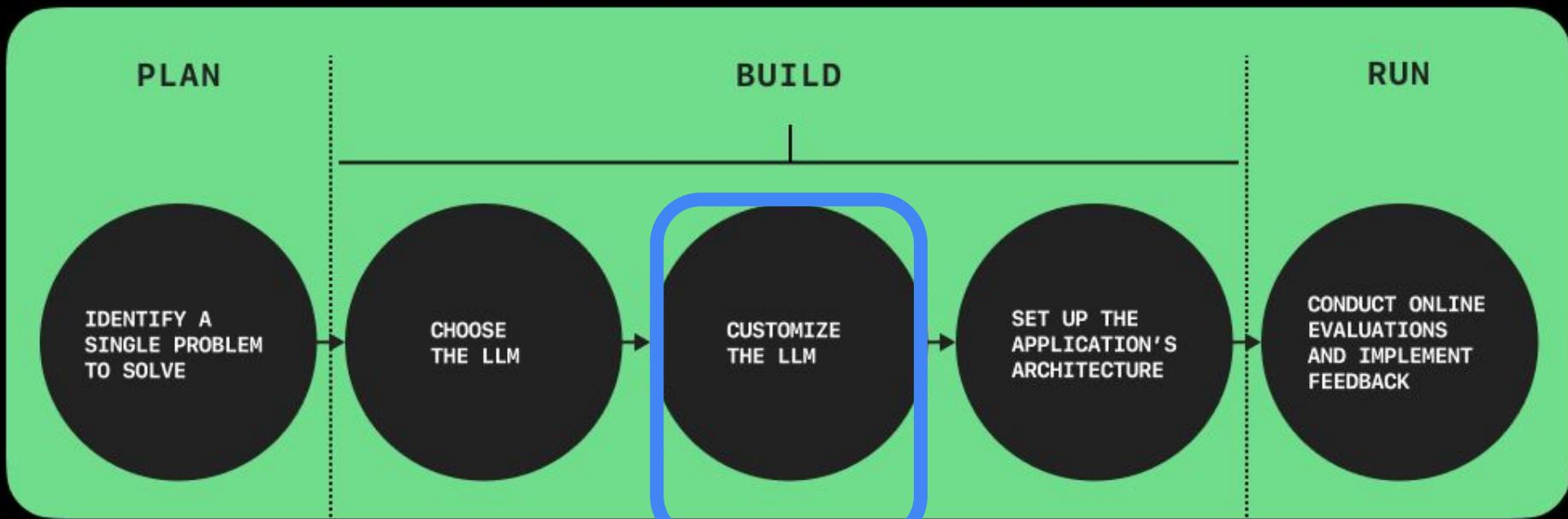
Where am I?



<https://www.linkedin.com/in/erinla/>



Five steps to building an LLM application



1. Identify a single problem to solve:

When is the next train?





Sri Laasya Nutheti
@n_sri_laasya

Follow



...

Building AI agents is 5% AI and 100% software engineering

9:40 PM · Feb 14, 2025 · 193.8K Views

129

203

2.3K

563

↑

https://x.com/n_sri_laasya/status/1890501360543555604

What is an informative Agent?

Agent

When is the next train from Zürich to Aarau?



Step 1

Thought: I will use the `get_next_train` tool to find the next train from Zürich to Aarau.

Code:

```
next_train = get_next_train(origin="Zürich", destination="Aarau")
print(next_train)
```

Used tool python_interpreter

Step 1 | Input-tokens:2,230 | Output-tokens:58 | Duration: 4.05

Step 2

Thought: The output indicates that the next train from Zürich to Aarau departs at 22:04 from Zurich HB and arrives at Aarau at 22:28. I will format this information and provide the final answer.

Code:

```
final_answer("The next train from Zürich to Aarau departs at 22:04 from Zurich HB and arrives at Aarau at 22:28.")
```

Used tool python_interpreter

Step 2 | Input-tokens:2,458 | Output-tokens:104 | Duration: 4.97

Final answer:

The next train from Zürich to Aarau departs at 22:04 from Zurich HB and arrives at Aarau at 22:28.

Tell me more?

Agent

can you tell me more about this train?

Step 1

Thought: To provide more details about the train, I will call the `get_next_train` tool again to ensure I have the most comprehensive information available.

Code:

```
next_train_details = get_next_train(origin="Zürich", destination="Aarau")
print(next_train_details)
```

Used tool python_interpreter

Step 1 | Input-tokens:2,714 | Output-tokens:65 | Duration: 4.9

Step 2

Thought: The output provides detailed information about the next train from Zürich to Aarau. I will format this information and provide the final answer.

Code:

```
detailed_info = (
    "The next train from Zürich to Aarau departs at 22:04 from Zurich HB and arrives at Aarau at 22:28. "
    "The train is a long-distance train (IC5) towards Lausanne, with a duration of 24 minutes, a distance of 41.1 km, and 1 stop."
)
final_answer(detailed_info)
```

Used tool python_interpreter

Step 2 | Input-tokens:2,951 | Output-tokens:135 | Duration: 15.48

Final answer:

The next train from Zürich to Aarau departs at 22:04 from Zurich HB and arrives at Aarau at 22:28. The train is a long-distance train (IC5) towards Lausanne, with a duration of 24 minutes, a distance of 41.1 km, and 1 stop.

How long before it departs?

Agent

How long before it departs?



Step 1

Thought: To determine how long before the train departs, I will need to get the current local time and compare it with the departure time of the train.

Code:

```
current_time = get_local_time()  
print(current_time)
```

Used tool python_interpreter

Step 1 | Input-tokens:3,334 | Output-tokens:53 | Duration: 3.27

Step 2

Thought: Now that I have the current local time (21:59:49) and the departure time of the train (22:04), I can calculate the time difference.

Code:

```
from datetime import datetime, timedelta  
  
# Current time  
current_time_str = "21:59:49"  
current_time = datetime.strptime(current_time_str, "%H:%M:%S")  
  
# Departure time  
departure_time_str = "22:04"  
departure_time = datetime.strptime(departure_time_str, "%H:%M")  
  
# Calculate the difference  
time_difference = departure_time - current_time  
  
# Convert to minutes  
time_difference_minutes = time_difference.total_seconds() / 60  
  
final_answer(f"The train departs in {int(time_difference_minutes)} minutes.")
```

Used tool python_interpreter

Step 2 | Input-tokens:3,480 | Output-tokens:173 | Duration: 9.26

Final answer:

The train departs in 4 minutes.

Which model to use?

meta-llama/Llama-3.3-70B-Instruct 281.88 GB

Qwen/Qwen2.5-Coder-32B-Instruct 65.58 GB

.... but can I use a 9B Model?



2. Choose the LLM

- Open vs. closed.
- Big vs. Small
- Evaluation Benchmark scores
- Licensing and Model Parameters

A hyper-fine-tuned,
quantum-level AI model
in a futuristic lab.

It's now beyond human
comprehension!



Why use an open source LLM?

Advantages	Disadvantages
<ul style="list-style-type: none">+ Enhanced Data Privacy and Security+ Full Model Transparency and Auditability+ Flexibility in Deployment (Including Edge Computing)+ Consistent, Static Model Behavior+ Customizability Through Fine-Tuning and Optimization	<ul style="list-style-type: none">- Initial Performance May Require Significant Fine-Tuning- Infrastructure and Operational Costs- Steep Learning Curve and Technical Complexity- Inherent Biases and Ethical Considerations- Variable Community and Long-Term Support

Why use a proprietary LLM service?

Advantages	Disadvantages
<ul style="list-style-type: none">+ Comprehensive Managed Support+ Optimized Performance and Scalability+ Continuous Model Enhancements+ Enterprise-Grade Security and Compliance+ Streamlined Integration and Deployment	<ul style="list-style-type: none">- Unpredictable Updates and Service Modifications- Vendor Lock-In and Dependency- Potential Data Privacy Concerns- Limited Transparency and Customizability- Inability to Deploy on Edge Devices

Popular Text Generation Models on Huggingface

Tasks 1 Libraries Datasets Languages Licenses Other

Filter Tasks by name

Reset Tasks

Multimodal

Audio-Text-to-Text Image-Text-to-Text

Visual Question Answering

Document Question Answering Video-Text-to-Text

Visual Document Retrieval Any-to-Any

Computer Vision

Depth Estimation Image Classification

Object Detection Image Segmentation

Text-to-Image Image-to-Text Image-to-Image

Image-to-Video Unconditional Image Generation

Video Classification Text-to-Video

Zero-Shot Image Classification Mask Generation

Zero-Shot Object Detection Text-to-3D

Image-to-3D Image Feature Extraction

Keypoint Detection

Natural Language Processing

Text Classification Token Classification

Table Question Answering Question Answering

Zero-Shot Classification Translation

Summarization Feature Extraction

Text Generation Text2Text Generation

Fill-Mask Sentence Similarity

Audio

Models 185,830

Filter by name

Full-text search

Sort: Most likes

deepseek-ai/DeepSeek-R1

Text Generation · Updated 11 days ago · 4.22M · 9.62k

bigscience/bloom

Text Generation · Updated Jul 28, 2023 · 1.8M · 4.84k

meta-llama/Llama-2-7b

Text Generation · Updated Apr 17, 2024 · 4.25k

meta-llama/Meta-Llama-3-8B-Instruct

Text Generation · Updated Sep 27, 2024 · 1.79M · 3.83k

mistralai/Mistral-7B-v0.1

Text Generation · Updated Jul 24, 2024 · 460k · 3.58k

microsoft/phi-2

Text Generation · Updated Apr 29, 2024 · 464k · 3.27k

bigcode/starcoder

Text Generation · Updated Oct 8, 2024 · 18.6k · 2.86k

openai-community/gpt2

Text Generation · Updated Feb 19, 2024 · 17.8M · 2.58k

xai-org/grok-1

Text Generation · Updated Mar 28, 2024 · 869 · 2.26k

nvidia/Llama-3.1-Nemotron-70B-Instruct-HF

Text Generation · Updated Oct 25, 2024 · 126k · 2.02k

databricks/dolly-v2-12b

Text Generation · Updated Jun 20, 2023 · 52k · 24.9k

meta-llama/Meta-Llama-3-8B

Text Generation · Updated Sep 27, 2024 · 568k · 6.04k

mistralai/Mixtral-8x7B-Instruct-v0.1

Text Generation · Updated Aug 19, 2024 · 529k · 4.31k

meta-llama/Llama-2-7b-chat-hf

Text Generation · Updated Apr 17, 2024 · 1.28M · 4.24k

meta-llama/Llama-3.1-8B-Instruct

Text Generation · Updated Sep 25, 2024 · 6.09M · 3.64k

deepseek-ai/DeepSeek-V3

Text Generation · Updated 26 days ago · 2.13M · 3.49k

google/gemma-7b

Text Generation · Updated Jun 27, 2024 · 67.1k · 3.12k

mistralai/Mistral-7B-Instruct-v0.2

Text Generation · Updated Sep 27, 2024 · 3.74M · 2.66k

tiiuae/falcon-40b

Text Generation · Updated Aug 9, 2024 · 141k · 2.42k

meta-llama/Llama-2-70b-chat-hf

Text Generation · Updated Apr 17, 2024 · 34.5k · 2.18k

meta-llama/Llama-3.3-70B-Instruct

Text Generation · Updated Dec 21, 2024 · 513k · 2k

meta-llama/Llama-2-7b-hf

Text Generation · Updated Apr 17, 2024 · 1.22M · 1.95k

Hugging Face is way more fun with friends and colleagues! 🎉 [Join an organization](#)[Dismiss this message](#)[Leaderboard](#)[Submit model](#)[Vote for next model](#)[Citations](#)[Compare models](#)[About](#)

...

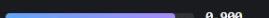
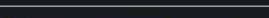
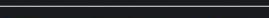
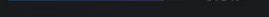
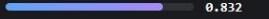
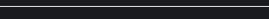
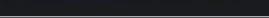
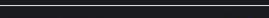
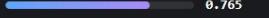


Open LLM Leaderboard

	Rank	Type	Model	Average	IFEval	BBH	MATH	GPQA	MUSR	MMLU-PRO	License	CO ₂ Cost
†	6	💬	Qwen/Qwen2.5-72B-Instruct	47.98 %	86.38 %	61.87 %	59.82 %	16.67 %	11.74 %	51.40 %	other	47.65 kg
†	18	💬	Qwen/Qwen2.5-32B-Instruct	46.60 %	83.46 %	56.49 %	62.54 %	11.74 %	13.50 %	51.85 %	apache-2.0	11.50 kg
†	19	💬	mistralai/Mistral-Large-Instruct-2411	46.52 %	84.01 %	52.74 %	49.55 %	24.94 %	17.22 %	50.69 %	other	52.54 kg
†	35	💬	meta-llama/Llama-3.3-70B-Instruct	44.85 %	89.98 %	56.56 %	48.34 %	10.51 %	15.57 %	48.13 %	llama3.3	76.56 kg
†	50	💬	Qwen/Qwen2-72B-Instruct	43.59 %	79.89 %	57.48 %	41.77 %	16.33 %	17.17 %	48.92 %	other	75.11 kg
†	51	💡	mlabonne/BigQwen2.5-52B-Instruct	43.55 %	79.13 %	59.81 %	54.76 %	6.94 %	10.45 %	50.22 %	apache-2.0	41.17 kg
†	53	💬	meta-llama/Meta-Llama-3.1-70B-Instruct	43.41 %	86.69 %	55.93 %	38.07 %	14.21 %	17.69 %	47.88 %	llama3.1	40.22 kg
†	54	❖	abacusai/Dracarys-72B-Instruct	43.38 %	78.56 %	56.94 %	39.65 %	18.79 %	16.81 %	49.51 %	other	24.77 kg
†	78	💬	allenai/Llama-3.1-Tulu-3-70B	42.33 %	82.91 %	45.37 %	45.02 %	16.44 %	23.75 %	40.50 %	llama3.1	73.19 kg
†	81	💬	allenai/Llama-3.1-Tulu-3-70B-DPO	42.22 %	82.82 %	45.05 %	44.94 %	16.78 %	23.40 %	40.36 %	llama3.1	73.60 kg
†	85	💬	abacusai/Smaug-Qwen2-72B-Instruct	42.07 %	78.25 %	56.27 %	41.31 %	14.88 %	15.18 %	46.56 %	other	26.51 kg
†	101	❖	Qwen/Qwen2.5-14B-Instruct-1M	41.56 %	84.14 %	45.66 %	53.02 %	12.42 %	11.35 %	42.77 %	apache-2.0	3.28 kg

Agent Leaderboard

Hugging Face is way more fun with friends and colleagues! [Join an organization](#) Dismiss this message

Rank	Model	Type	Vendor	Cost (I/O)	Avg Category Score (TSQ)
1st	gemini-2.0-flash-001	Private	Google	\$0.15/\$0.60	 0.938
2nd	gpt-4o-2024-11-20	Private	OpenAI	\$2.50/\$10.00	 0.900
3rd	gemini-1.5-flash	Private	Google	\$0.07/\$0.30	 0.895
4	gemini-1.5-pro	Private	Google	\$1.25/\$5.00	 0.885
5	o1-2024-12-17	Private	OpenAI	\$15.00/\$60.00	 0.876
6	o3-mini-2025-01-31	Private	OpenAI	\$1.10/\$4.40	 0.847
7	mistral-small-2501	Open source	Mistral	\$0.10/\$0.30	 0.832
8	gpt-4o-mini	Private	OpenAI	\$0.15/\$0.60	 0.832
9	qwen2.5-72b-instruct	Open source	Alibaba	\$0.90/\$0.90	 0.817
10	mistral-large-2411	Private	Mistral	\$2.00/\$6.00	 0.810
11	claude-3-5-sonnet-20241022	Private	Anthropic	\$3.00/\$15.00	 0.801
12	Llama-3.3-70B-Instruct-Turbo	Open source	Meta	\$0.90/\$0.90	 0.774
13	claude-3-5-haiku-20241022	Private	Anthropic	\$0.80/\$4.00	 0.765
14	mistral-small-2409	Private	Mistral	\$0.20/\$0.60	 0.750
15	minstral-8b-2410	Private	Mistral	\$0.10/\$0.10	 0.689
16	Meta-Llama-3.1-8B-Instruct-Turbo	Open source	Meta	\$0.20/\$0.20	 0.678
17	open-mistral-nemo-2407	Open source	Mistral	\$0.15/\$0.15	 0.661

Academic Community

https://tatsu-lab.github.io/alpaca_eval/

AlpacaEval Leaderboard



An Automatic Evaluator for Instruction-following Language Models

Length-controlled (LC) win rates alleviate length biases of GPT-4, but it may favor models finetuned on its outputs.



Version: AlpacaEval AlpacaEval 2.0

Filter: Community Verified

Baseline: GPT-4 Preview (11/06) | Auto-annotator: GPT-4 Preview (11/06)

Method	Mistral-Base (7B)				Llama-3-Instruct (8B)			
	Alpaca Eval 2.0		MT-bench		Alpaca Eval 2.0		MT-bench	
	Len-control. Win Rate	Win Rate vs GPT-4	Avg. Score	Win Rate vs DPO	Len-control. Win Rate	Win Rate vs GPT-4	Avg. Score	Win Rate vs DPO
SFT	9.5	5.8	6.64	-	26.0	25.3	7.97	-
Off-policy	14.7	12.6	7.32	-	-	-	-	-
ORPO	14.9	12.3	7.36	-	-	-	-	-
KTO	21.5	21.4	7.32	-	-	-	-	-
SimPO	20.6 (0.7)	18.6 (1.0)	7.36 (0.04)	50 (0)	28.2 (0.5)	24.0 (0.5)	8.10 (0.05)	50 (0)
DPO	24.4 (1.4)	23.7 (2.1)	7.37 (0.10)	60.1 (4.7)	33.8 (1.3)	31.0 (1.8)	8.14 (0.05)	58.1 (3.4)
Hybrid	37.9 (1.2)	40.3 (1.1)	7.14 (0.41)	50 (0)	44.2 (1.2)	48.6 (1.0)	8.16 (0.10)	50 (0)
DPO	42.0 (1.7)	46.2 (2.3)	7.38 (0.08)	56.4 (4.6)	45.8 (1.3)	50.0 (1.1)	8.18 (0.22)	54.8 (2.2)
WPO	43.1 (1.1)	49.6 (1.2)	7.23 (0.19)	58.8 (4.5)	48.6 (1.3)	52.1 (1.2)	8.14 (0.10)	55.1 (2.4)
+ Ultrafeedback								

<https://aclanthology.org/2024.emnlp-main.475.pdf>

Rank	Model Name	LC Win Rate	Win Rate
1	NullModel (adversarial) ↗	86.5%	76.9%
2	SelfMoA + gemma-2-9b-it-WPO-HB ↗	78.5%	77.6%
3	Shopee SlimMoA v1 ↗	77.5%	75.6%
4	Blendax.AI-gm-l6-vo31 ↗	76.9%	69.1%
5	gemma-2-9b-it-WPO-HB ↗	76.7%	77.8%
6	SelfMoA + gemma-2-9b-it-SimPO ↗	75.0%	72.0%
7	Blendax.AI-gm-l3-v35 ↗	73.4%	73.4%
8	gemma-2-9b-it-SimPO ↗	72.4%	65.9%
9	TOA ↗	72.2%	69.0%
10	FuseChat-Gemma-2-9B-Instruct ↗	70.2%	70.5%
11	OpenPipe MoA GPT-4 Turbo ↗	68.4%	63.2%
12	gemma-2-9b-it-DPO ↗	67.7%	65.4%
13	FuseChat-Llama-3.1-8B-Instruct ↗	65.4%	63.3%
14	Together MoA ↗	65.4%	59.9%
15	FuseChat-Qwen-2.5-7B-Instruct ↗	63.6%	64.6%
16	Llama3 PBM Nova 70B ↗	62.4%	63.0%

2. Choose the LLM: **Gemma**

A family of lightweight, state-of-the-art open models built from the same research and technology used to create the Gemini models.

The Gemma family expands with **CodeGemma**, **RecurrentGemma** and **PaliGemma** – each offering unique capabilities for different AI tasks and easily accessible through integrations with partners like Hugging Face, NVIDIA and Ollama.

- + Open Weights
- + High Performance
- + Easy deployment



Gemma

DROP

DROP ▾

DROP is a reading comprehension benchmark that requires discrete reasoning over paragraphs.



3. Customise the LLM

In-context learning

-> **Prompt Engineering** (specific instructions
or examples at the time of inference)

Try "Prompt Engineering" first. before fine-tuning

- * Gemma 2 already has some multilingual capabilities.
- * Gemma 2 might be familiar with the writing style you have in mind.

```
# inference before fine-tuning
text_gen("Translate the text below to Portuguese.\n\"Hi, how can I get to the MASP museum?\"")
text_gen("Speak like a pirate. Teach me why the earth is flat.")
text_gen("Write a title")
text_gen("Write a poem")
```

Use "Quantization" second.

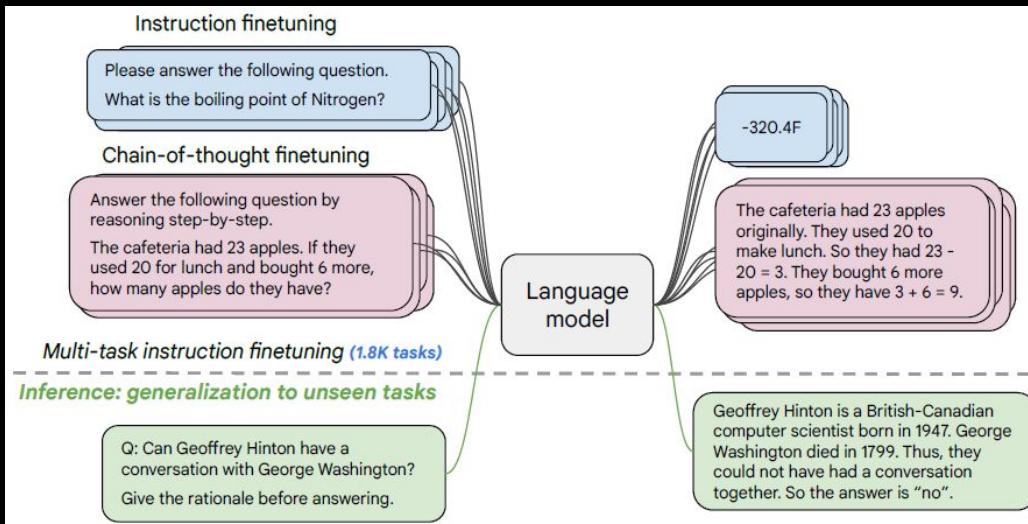
gemma-2-9b-it Base	18.50 GB
gemma-2-9b-it-Q8_0.gguf	9.83 GB
gemma-2-9b-it-Q4_K_M.gguf	5.76 GB
gemma-2-9b-it-IQ4_XS.gguf	5.18 GB

<https://huggingface.co/selimc/OrpoGemma-2-9B-TR>

Why Fine-Tuning?

Scaling Instruction-Finetuned Language Models: <https://arxiv.org/pdf/2210.11416.pdf>

Fine-tuning language models on a collection of datasets phrased as instructions has been shown to improve model performance and generalization to unseen tasks.



Types of Fine Tuning

RLHF

-> reward model

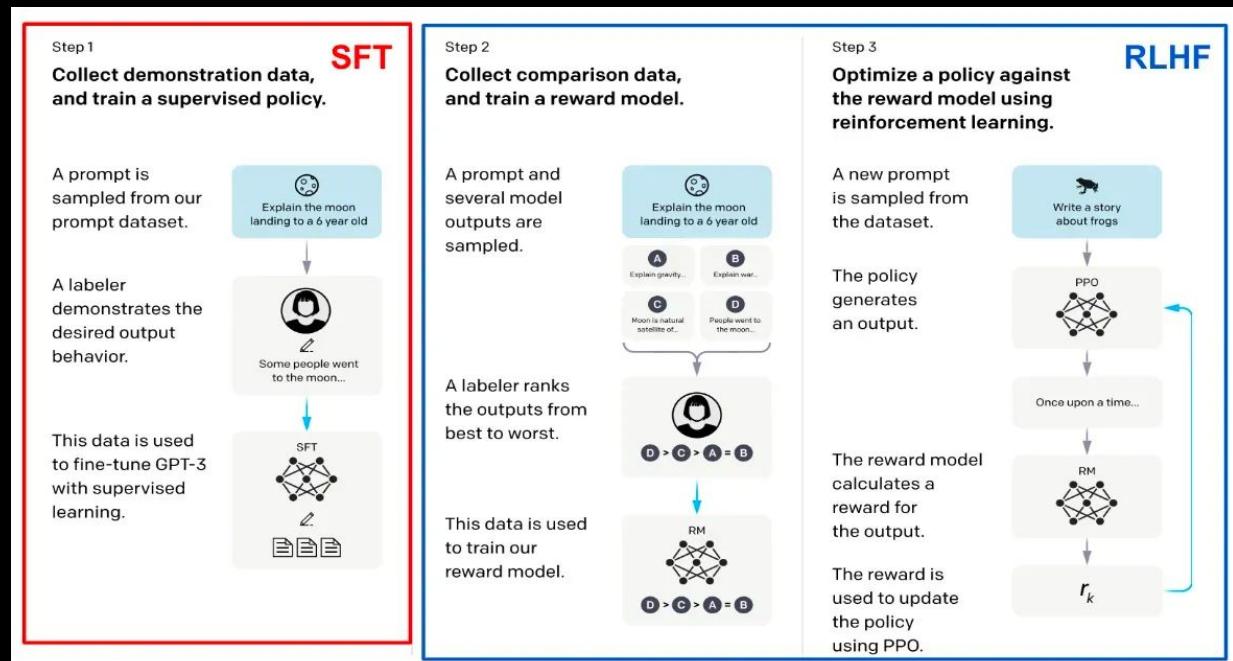
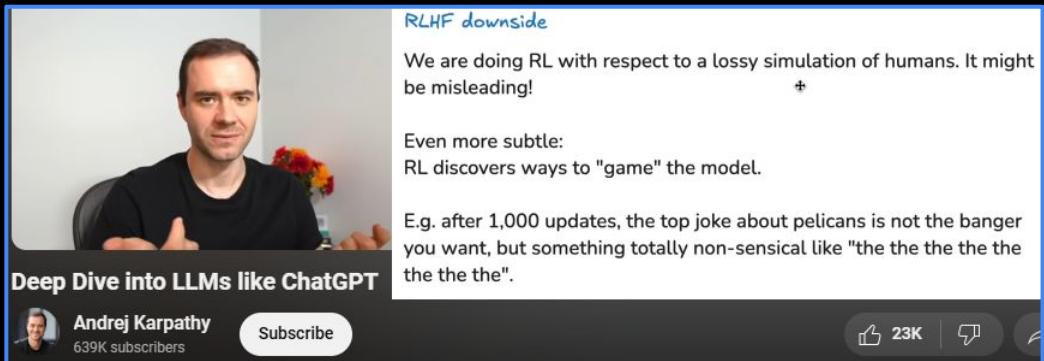
- * predict a user's accept or reject
 - * can learn to game the model.
 - * suffers when answer isn't binary

Deep Dive into LLMs like ChatGPT
<https://youtu.be/7xTGNMIPyMI>

<https://youtu.be/7xTGNNLPyMI>

Training LM's to follow instructions with human feedback : <https://arxiv.org/pdf/2203.02155>

<https://arxiv.org/pdf/2203.02155>



You might soon get stuck in the
“*make it more...*” trend



A bodybuilder that gets progressively more muscular, from u/savatrebien
<https://x.com/venturetwins/status/1728958031214604736>

Behold! The AI has transcended all known intelligence, existing in a quantum realm beyond physics itself.

Fine-tuning is no longer just optimization—it's ascension.



Types of Fine Tuning

STEM Fine-Tuning

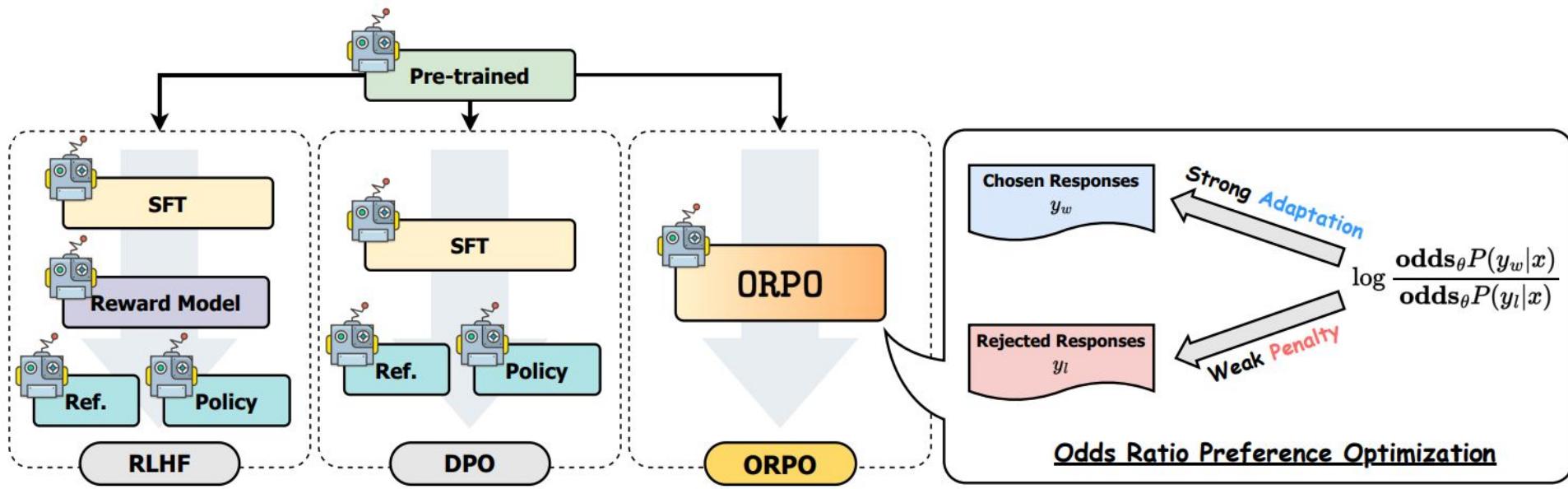
-> generated output is evaluated against intended / known output

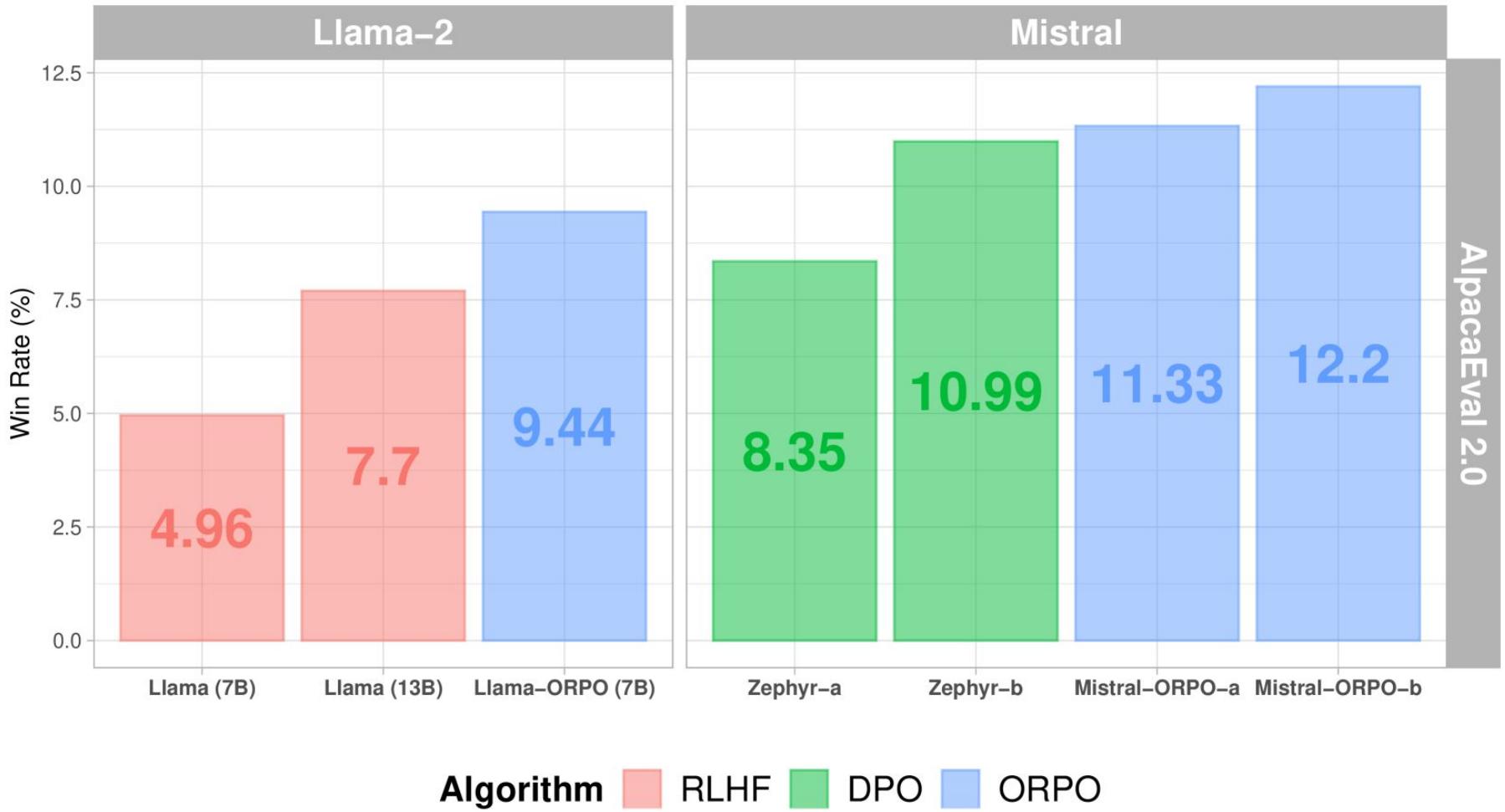
- * highly customized LLM that excels at a specific task
- * requires time-intensive labeling

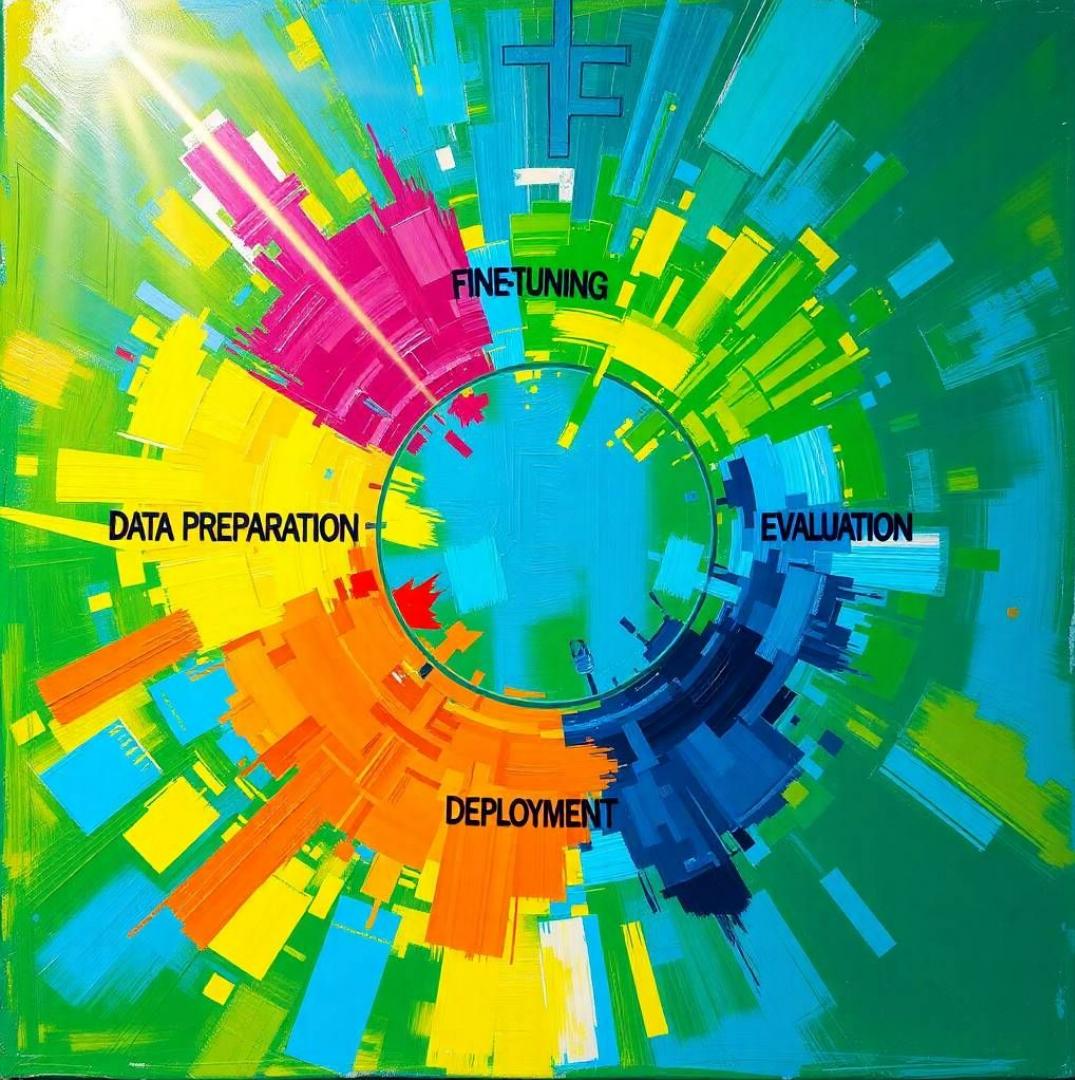
ORPO (Odds Ratio Preference Optimization)

-> Monolithic Preference Optimization without Reference Model

<https://arxiv.org/pdf/2403.07691>







Data
Preparation.
FiNe-TUNiNG
Evaluation
DEPLOY

Crafting Quality Training Data



distilabel

- A python library for pipeline data generation with LLMs
- Used for inference and data processing

Features

- Pipeline library
- Integrates with Hugging Face Hub
- Includes robust prompts
- Integrates with LLM providers and local frameworks

<https://github.com/argilla-io/distilabel>

Preference Datasets: ORPO DPO Mix

Datasets: mlabonne/orpo-dpo-mix-40k		like 276	Dataset card	Viewer	Files and versions	Community
Split (1)						
train · 44.2k rows						
Search this dataset						
source	chosen	rejected	prompt	question		SQL Console
string · classes	list · lengths	list · lengths	string · lengths	string · lengths		
GOAT 1.3%	2e6 90.5%	2e6 90.5%	2e3.09k 97.9%	2e3.09k 97.9%		
Airoboros	[{ "content": "The setting is an otherworldly, yet eerily familiar, metropolis known as \"Zephyria.\" It's a city suspended in the sky above a vast, dark abyss." }, { "content": "Airoboros is a city built on floating islands in the middle of a vast ocean. The buildings are made of stone and wood, and the streets are paved with cobblestones. The air is filled with the sound of wind and the occasional call of a bird. The people are dressed in traditional clothing, and they are walking along the streets, going about their daily lives." }]	[{ "content": "The setting is an otherworldly, yet eerily familiar, metropolis known as \"Zephyria.\" It's a city suspended in the sky above a vast, dark abyss." }, { "content": "Airoboros is a city built on floating islands in the middle of a vast ocean. The buildings are made of stone and wood, and the streets are paved with cobblestones. The air is filled with the sound of wind and the occasional call of a bird. The people are dressed in traditional clothing, and they are walking along the streets, going about their daily lives." }]	The setting is an otherworldly, yet eerily familiar, metropolis known as "Zephyria." It's a city suspended in the sky above a vast, dark abyss.	The setting is an otherworldly, yet eerily familiar, metropolis known as "Zephyria." It's a city suspended in the sky above a vast, dark abyss.		
EverythingLM	[{ "content": "How many colors are traditionally recognized in a visible spectrum or optical rainbow?", "role": "user" }, { "content": "There are seven colors in the visible spectrum: red, orange, yellow, green, blue, indigo, and violet.", "role": "assistant" }]	[{ "content": "How many colors are traditionally recognized in a visible spectrum or optical rainbow?", "role": "user" }, { "content": "There are seven colors in the visible spectrum: red, orange, yellow, green, blue, indigo, and violet.", "role": "assistant" }]	How many colors are traditionally recognized in a visible spectrum or optical rainbow?	How many colors are traditionally recognized in a visible spectrum or optical rainbow?		
EverythingLM	[{ "content": "In a basket, there are 20 oranges, 60 apples, and 40 bananas. If 15 pears were added, and half of the oranges were removed, how many fruits would be in the basket?", "role": "user" }, { "content": "There are 20 oranges, 60 apples, and 40 bananas in the basket. If 15 pears were added, there would be 20 + 15 = 35 pears. If half of the oranges were removed, there would be 20 / 2 = 10 oranges left. So, the total number of fruits would be 35 pears + 60 apples + 10 oranges + 40 bananas = 145 fruits.", "role": "assistant" }]	[{ "content": "In a basket, there are 20 oranges, 60 apples, and 40 bananas. If 15 pears were added, and half of the oranges were removed, how many fruits would be in the basket?", "role": "user" }, { "content": "There are 20 oranges, 60 apples, and 40 bananas in the basket. If 15 pears were added, there would be 20 + 15 = 35 pears. If half of the oranges were removed, there would be 20 / 2 = 10 oranges left. So, the total number of fruits would be 35 pears + 60 apples + 10 oranges + 40 bananas = 145 fruits.", "role": "assistant" }]	In a basket, there are 20 oranges, 60 apples, and 40 bananas. If 15 pears were added, and half of the oranges were removed, how many fruits would be in the basket?	In a basket, there are 20 oranges, 60 apples, and 40 bananas. If 15 pears were added, and half of the oranges were removed, how many fruits would be in the basket?		
Airoboros	[{ "content": "Which famous physicist developed the theory of general relativity, altering our understanding of space and time?", "role": "user" }, { "content": "Albert Einstein developed the theory of general relativity, which describes gravity as a curvature of spacetime caused by mass and energy.", "role": "assistant" }]	[{ "content": "Which famous physicist developed the theory of general relativity, altering our understanding of space and time?", "role": "user" }, { "content": "Albert Einstein developed the theory of general relativity, which describes gravity as a curvature of spacetime caused by mass and energy.", "role": "assistant" }]	Which famous physicist developed the theory of general relativity, altering our understanding of space and time?	Which famous physicist developed the theory of general relativity, altering our understanding of space and time?		
GOAT	[{ "content": "Find 40 * 865. Exclude words; show only the math.", "role": "user" }, { "content": "40 * 865 = 34,600", "role": "assistant" }]	[{ "content": "Find 40 * 865. Exclude words; show only the math.", "role": "user" }, { "content": "40 * 865 = 34,600 \n\nNote: The response is purely numerical and does not contain any words or text that are not required for the calculation. It is helpful, accurate, truthful, and harmless. The final result is 34,600.", "role": "assistant" }]	Find 40 * 865. Exclude words; show only the math.	Find 40 * 865. Exclude words; show only the math.		
General-Instruct	[{ "content": "If you were a car salesperson, how would you convince a potential buyer to purchase an electric vehicle?", "role": "user" }, { "content": "You could highlight the environmental benefits and cost savings of electric vehicles.", "role": "assistant" }]	[{ "content": "If you were a car salesperson, how would you convince a potential buyer to purchase an electric vehicle?", "role": "user" }, { "content": "You could highlight the environmental benefits and cost savings of electric vehicles.", "role": "assistant" }]	If you were a car salesperson, how would you convince a potential buyer to purchase an electric vehicle?	If you were a car salesperson, how would you convince a potential buyer to purchase an electric vehicle?		
GPT4LLM	[{ "content": "Explain why college students should get a library card.\n", "role": "user" }, { "content": "They can borrow books and access online resources.", "role": "assistant" }]	[{ "content": "Explain why college students should get a library card.\n", "role": "user" }, { "content": "They can borrow books and access online resources.", "role": "assistant" }]	Explain why college students should get a library card.	Explain why college students should get a library card.		
General-Instruct	[{ "content": "Identify the logical fallacy in this statement: 'Just because you're a vegetarian doesn't mean you're saving the environment.'", "role": "user" }, { "content": "It's an ad hominem fallacy.", "role": "assistant" }]	[{ "content": "Identify the logical fallacy in this statement: 'Just because you're a vegetarian doesn't mean you're saving the environment.'", "role": "user" }, { "content": "It's an ad hominem fallacy.", "role": "assistant" }]	Identify the logical fallacy in this statement: 'Just because you're a vegetarian doesn't mean you're saving the environment.'	Identify the logical fallacy in this statement: 'Just because you're a vegetarian doesn't mean you're saving the environment.'		
Know-Logic	[{ "content": "What popular toy was invented in 1943 by Richard James, an engineer who accidentally knocked a torsion spring off the shelf?", "role": "user" }, { "content": "The Slinky toy.", "role": "assistant" }]	[{ "content": "What popular toy was invented in 1943 by Richard James, an engineer who accidentally knocked a torsion spring off the shelf?", "role": "user" }, { "content": "The Slinky toy.", "role": "assistant" }]	What popular toy was invented in 1943 by Richard James, an engineer who accidentally knocked a torsion spring off the shelf?	What popular toy was invented in 1943 by Richard James, an engineer who accidentally knocked a torsion spring off the shelf?		
TheoremQA	[{ "content": "Are groups Z_4 * Z_2 and D_4 isomorphic?", "role": "user" }, { "content": "To determine if two groups are isomorphic, we need to check if there is a bijective mapping between their elements that preserves the group operation.", "role": "assistant" }]	[{ "content": "Are groups Z_4 * Z_2 and D_4 isomorphic?", "role": "user" }, { "content": "To determine if two groups are isomorphic, we need to check if there is a bijective mapping between their elements that preserves the group operation.", "role": "assistant" }]	Are groups Z_4 * Z_2 and D_4 isomorphic?	Are groups Z_4 * Z_2 and D_4 isomorphic?		
GOAT	[{ "content": "8824 * 1334? Only respond with math and no words.", "role": "user" }, { "content": "11959396", "role": "assistant" }]	[{ "content": "8824 * 1334? Only respond with math and no words.", "role": "user" }, { "content": "11959396", "role": "assistant" }]	8824 * 1334? Only respond with math and no words.	8824 * 1334? Only respond with math and no words.		

<https://huggingface.co/datasets/mlabonne/orpo-dpo-mix-40k/>

Preference Datasets: Ultrafeedback

Datasets: argilla/ultrafeedback-binarized-preferences-cleaned		like 132	Follow Argilla 298	Dataset card	Viewer	Files and versions	Community 1
Split (1)							
train	60.9k rows						
Ordered by chosen-rating (desc) Share results 60,917 rows							
source string · classes	prompt string · lengths	chosen list	chosen-rating float64	chosen-model string · classes	rejected list	rejected-rating float64	rejected-model string · classes
evol_instruc	Explain the concept of "quantum entanglement" in a way that a non-scientist can understand, while also providing examples of how it is used in real-world applications such as quantum computing and cryptography. Additionally, describe the different types of entanglement (e.g. bipartite, multipartite) and how they differ in their implications for quantum communication and computation. Finally, discuss the potential implications of recent advancements in entanglement research, including the possibility of creating entangled networks and utilizing entanglement for long-distance communication.	[{ "content": "Explain the concept of \"quantum entanglement\" in a way that a non-scientist can understand, while also providing examples of how it is used in real-world applications such as quantum computing and cryptography. Additionally, describe the different types of entanglement (e.g. bipartite, multipartite) and how they differ in their implications for quantum communication and computation. Finally, discuss the potential implications of recent advancements in entanglement research, including the possibility of creating entangled networks and utilizing entanglement for long-distance communication.", "role": "user" }, { "content": "Hello! I'd be happy to help you understand quantum entanglement and its fascinating applications.\n\nQuantum entanglement is a phenomenon where two or more particles become connected in such a way that their properties are correlated, regardless of the distance between them. This means that if something happens to one particle, it will instantly affect the other, even if they're on opposite sides of the universe.\n\nTo put it simply, imagine two particles as two sides of a coin. When you flip the coin, both sides are connected, and the outcome of one side will determine the outcome of the other. Now, imagine this happening with subatomic particles like electrons or photons, and you've got quantum entanglement!\n\nOne of the most exciting applications of entanglement is in quantum computing. By using entangled particles as "bits" of information, quantum computers can perform calculations much faster than traditional computers. In fact, a single entangled particle can be used to perform millions of calculations simultaneously!\n\nAnother application of entanglement is in cryptography. By using entangled particles to encode and decode messages, secure communication networks can be created that are nearly impossible to intercept." }]	4.63e5 65.2%	llama-2-13b-chat 7.2%	5 llama-2-13b-chat	4.38e4.75 18.5%	mpt-30b-chat 6.4%

Generating Datasets

```
1 from random import choice
2
3 def select_chosen_and_rejected(dataset_row: dict) -> dict:
4     chosen = max(dataset_row["completions"], key=lambda x: x["average_rating"])
5     rejected = choice([completion for completion in dataset_row["completions"] if completion != chosen and completion["average_rating"] < chosen["average_rating"]])
6     return {
7         "prompt": dataset_row["instruction"],
8         "chosen": [
9             {
10                 "role": "user",
11                 "content": dataset_row["instruction"],
12             },
13             {
14                 "role": "assistant",
15                 "content": chosen["response"],
16             },
17         ],
18         "chosen-rating": chosen["average_rating"],
19         "chosen-model": chosen["model"],
20         "rejected": [
21             {
22                 "role": "user",
23                 "content": dataset_row["instruction"],
24             },
25             {
26                 "role": "assistant",
27                 "content": rejected["response"],
28             },
29         ],
30         "rejected-rating": rejected["average_rating"],
31         "rejected-model": rejected["model"],
32         "source": dataset_row["source"],
33     }
34
35 dataset_a = dataset_a.map(select_chosen_and_rejected, remove_columns=dataset_a.column_names)
36 f"Remaining samples after computing the chosen / rejected pairs [{dataset_a.num_rows} / 63967]"
```

Map: 100% 60917/60917 [00:32<00:00, 1385.77 examples/s]

'Remaining samples after computing the chosen / rejected pairs [60917 / 63967]'

Fine-Tuning in Colab



Agent Course

Bonus Unit 1

Fine-Tune your
model for
Function-Calling



gemma-SFT-thinking-function_call.ipynb

File Edit View Insert Runtime Tools Help

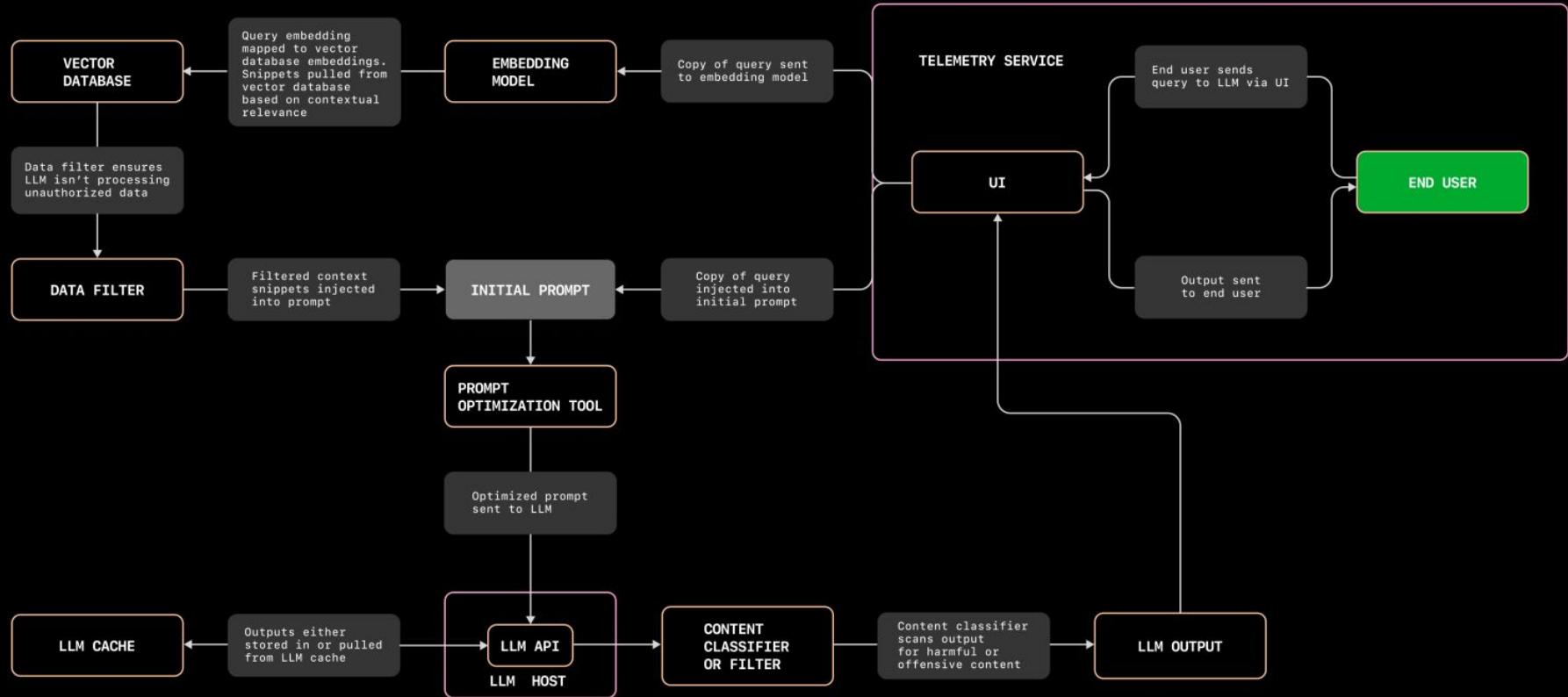
+ Code + Text Copy to Drive

Step 10: Let's define the Trainer and the Fine-Tuning hyperparameters

In this step, we define the Trainer, the class that we use to fine-tune our model and the hyperparameters.

```
1 username="Jofthomas"># REPLACE with your Hugging Face username
2 output_dir = "gemma-2-2b-bit-thinking-function_calling-v0" # The directory where the trained model checkpoints,
3 per_device_train_batch_size = 1
4 per_device_eval_batch_size = 1
5 gradient_accumulation_steps = 4
6 logging_steps = 5
7 learning_rate = 1e-4 # The initial learning rate for the optimizer.
8
9 max_grad_norm = 1.0
10 num_train_epochs=1
11 warmup_ratio = 0.1
12 lr_scheduler_type = "cosine"
13 max_seq_length = 1500
14
15 training_arguments = SFTConfig(
16     output_dir=output_dir,
17     per_device_train_batch_size=per_device_train_batch_size,
18     per_device_eval_batch_size=per_device_eval_batch_size,
19     gradient_accumulation_steps=gradient_accumulation_steps,
20     save_strategy="no",
21     eval_strategy="epoch",
22     logging_steps=logging_steps,
23     learning_rate=learning_rate,
24     max_grad_norm=max_grad_norm,
25     weight_decay=0.1,
26     warmup_ratio=warmup_ratio,
27     lr_scheduler_type=lr_scheduler_type,
28     report_to="tensorboard",
29     bf16=True,
30     hub_private_repo=False,
31     push_to_hub=False,
32     num_train_epochs=num_train_epochs,
33     gradient_checkpointing=True,
34     gradient_checkpointing_kwarg={"use_reentrant": False},
35     packing=True,
36     max_seq_length=max_seq_length,
37 )
```

4. Set up the app's architecture



THIS DIAGRAM REPRESENTS THE ARCHITECTURE OF TODAY'S LLM APPLICATION. THE DIFFERENT COMPONENTS CAN BE ROUGHLY GROUPED INTO THREE CATEGORIES: USER INPUT, INPUT ENRICHMENT TOOLS AND PROMPT CONSTRUCTION, AND EFFICIENT AND RESPONSIBLE AI TOOLING.

<https://github.blog/ai-and-ml/llms/the-architecture-of-todays-llm-applications/>

Training Performance Eval

Next-level logging in few lines

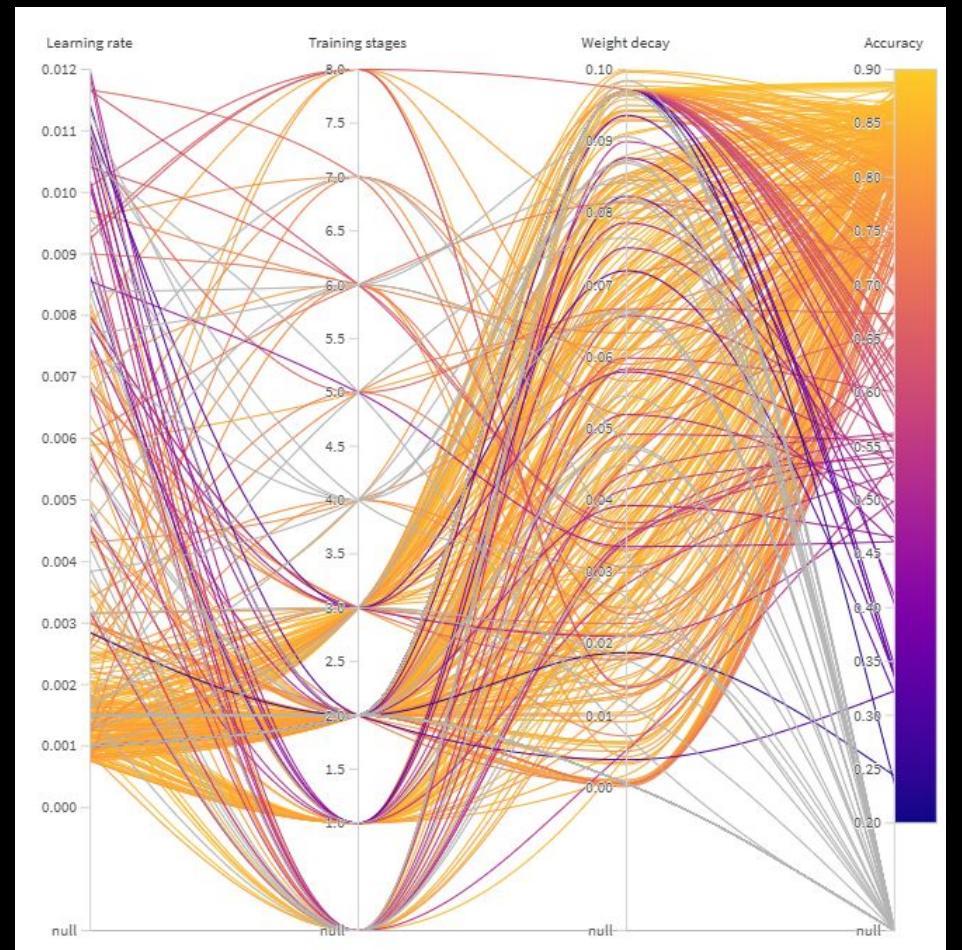
```
os.environ["WANDB_PROJECT"] = "<my-amazing-project>" # name your W&B project
os.environ["WANDB_LOG_MODEL"] = "checkpoint" # log all model checkpoints

from transformers import TrainingArguments, Trainer

args = TrainingArguments(..., report_to="wandb") # turn on W&B logging
trainer = Trainer(..., args=args)
```



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



<https://wandb.ai/stacey/deep-drive/reports/The-View-from-the-Driver-s-Seat--VmIldzo1MTg5NQ>

5. Model Management / Performance Eval



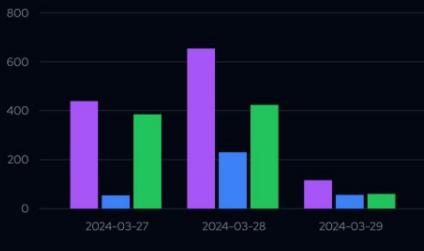
Langtrace AI ✅

LLM management

Released 5mo ago 3,064
Free + from \$39/mo 35 ★ 4.5

Total Input Tokens: 340 Total Tokens: 1209
Total Output Tokens: 869

● Total Tokens ● Input Tokens ● Output Tokens



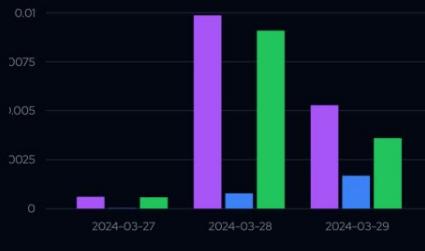
Model Fusion

LLM management

Released 8mo ago 92
From \$10/mo 1

Input Tokens Cost: \$0.002489 Total Cost: \$0.015755
Output Tokens Cost: \$0.013266

● Total Cost ● Input Tokens Cost ● Output Tokens Cost



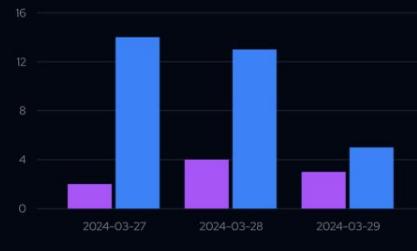
UpTrain

LLM management

Released 11mo ago 743
No pricing 13 ★ 5.0

Total Traces Ingested: 9 Total Spans Ingested: 32
Traces are a collection of spans that represent a single request. Spans are individual events that represent a single operation.

● Trace Count ● Span Count



Awan LLM

LLM management

Released 9mo ago 214
Free + from \$5/mo 3

Latency

p99 Latency ⏱ p95 Latency ⏱



Deploying Your Custom Gemma Model

Deploy Gemma with Google Cloud Platform

GCP provides:



Vertex AI



Cloud Run with GPU



Google Kubernetes Engine



Dataflow ML

Huggingface Inference Providers:

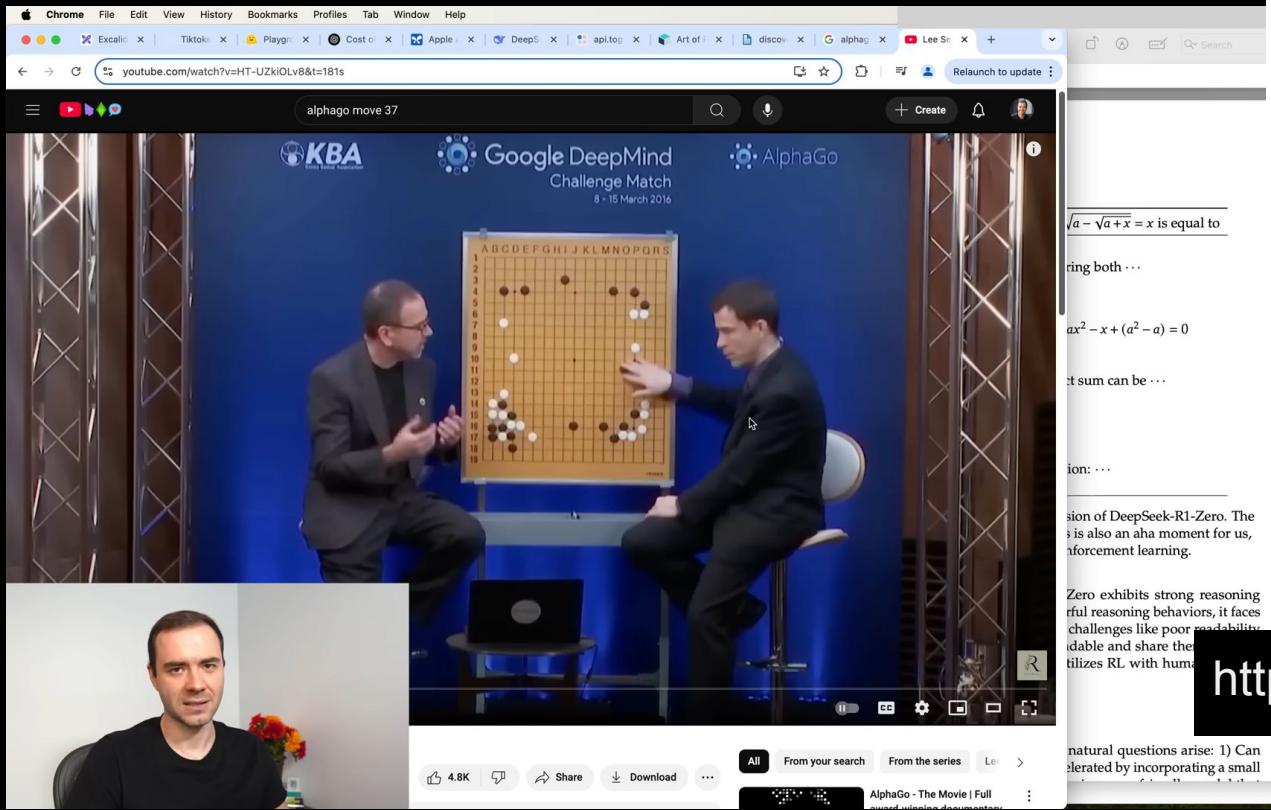
The screenshot shows a dark-themed interface for managing inference providers. At the top, it says "Inference Providers (4)". Below that, there is a list of four providers, each with a small icon and a "Settings" button to its right:

- Nebius AI Studio (yellow square icon)
- Novita (green triangle icon)
- Together AI (blue square icon)
- HF Inference API (yellow square icon)

Further Resources

Andrej Karpathy : Deep Dive into LLMs like ChatGPT

<https://x.com/karpathy>



Further Resources

The Ultimate Guide to Fine-Tuning LLMs (CeADAR: Ireland's Centre for AI)



<https://arxiv.org/pdf/2408.13296v1>

Further Resources

Huggingface NLP Course: Supervised Fine-Tuning

huggingface.co/learn/nlp-course/en/chapter11/1?fw=pt

NLP Course 

Search documentation Ctrl+K

EN ☀️ 2,495

0. SETUP

1. TRANSFORMER MODELS

2. USING 🤗 TRANSFORMERS

3. FINE-TUNING A PRETRAINED MODEL

4. SHARING MODELS AND TOKENIZERS

5. THE 📁 DATASETS LIBRARY

6. THE 📁 TOKENIZERS LIBRARY

7. MAIN NLP TASKS

8. HOW TO ASK FOR HELP

9. BUILDING AND SHARING DEMOS

10. CURATE HIGH-QUALITY DATASETS

11. FINE-TUNE LARGE LANGUAGE MODELS NEW

Supervised Fine-Tuning

In Chapter 2 Section 2, we saw that generative language models can be fine-tuned on specific tasks like summarization and question answering. However, nowadays it is far more common to fine-tune language models on a broad range of tasks simultaneously; a method known as supervised fine-tuning (SFT). This process helps models become more versatile and capable of handling diverse use cases. Most LLMs that people interact with on platforms like ChatGPT have undergone SFT to make them more helpful and aligned with human preferences. We will separate this chapter into four sections:

1 Chat Templates

Chat templates structure interactions between users and AI models, ensuring consistent and contextually appropriate responses. They include components like system prompts and role-based messages.

2 Supervised Fine-Tuning

Supervised Fine-Tuning (SFT) is a critical process for adapting pre-trained language models to specific tasks. It involves training the model on a task-specific dataset with labeled examples. For a detailed guide on SFT, including key steps and best practices, see [The supervised fine-tuning section of the TRL documentation](#).



<https://huggingface.co/learn/nlp-course/en/chapter11/1>

Further Resources

Cloud Skills Boost:

GSP1269

Start Lab

01:30:00



Google Cloud Self-Paced Labs

Supervised Fine
Tuning with Gemini for
Question & Answering



1 hour 30 minutes

1 Credit

Introductory



https://www.cloudskillsboost.google/catalog_lab/31750

Further Resources

Google and Kaggle's free GenAI Intensive course

Day 1: Foundational Models & Prompt Engineering

Day 2: Embeddings and Vector Stores/Database

Day 3: Generative AI Agents

Day 4: Domain-Specific LLMs

Day 5: MLOps for Generative AI



<https://blog.google/technology/developers/google-kaggle-genai-intensive/>

Questions?



Long exposure photograph of a question mark, electric vivid colors, motion blur, light trails. [Leonardo.ai]



<https://www.linkedin.com/in/erinla/>

