

O'REILLY®

How to Choose the Right LLM for Your Application

Jonathan Fernandes



What we will be covering today



Fundamentals of LLMs

- What are LLMs?
- Parameters
- Tokens
- Context length
- LLM trade-offs
- Getting started for your application
- Why are comparing LLMs difficult?

Benchmarking LLMs

- HELM
- OpenLLM Leaderboard
- Using your own data
- Q&A
- Break

Elo Ratings

- Comparing LLMs
- Notebook
- Exercise
- Q&A

 Live Course



Hands-on Retrieval Augmented Generation (RAG)

With [Jonathan Fernandes](#)

⌚ 3h 0m 🗓 June 27 • 5pm-8pm GMT+1

 Live Course



Hands-on GPT-4-Turbo

With [Jonathan Fernandes](#)

⌚ 3h 0m 🗓 June 20 • 5pm-8pm GMT+1



About me



Jonathan A. Fernandes 



United Kingdom · [Contact info](#)



AI & ML Advisory Services



University of Warwick -
Warwick Business School



Jonathan A. Fernandes • You

1w • 0

This is one of the worst-kept secrets in the field of Generative AI and Large Language models.

Imagine preparing for a test and having the exact questions beforehand. That's essentially what happens in the AI world.

The models often get trained on materials that they'll be tested on in benchmarks like HELM. So, when they ace these tests, it's impressive but not entirely surprising.

That's why it's always important to not only consider different benchmarks as a guiding light but to have your own set of tests that makes sense for your business for you to determine if a Large Language Model performs well for your business.

...



**Which are the best quality closed models?
(in decreasing order of quality)**

GPT-4-Turbo > Claude3 → GPT-4o
 GPT-4o → GPT-4-Turbo → Claude3



Which are the best quality open models?
(in decreasing order of quality)

- 👍 Llama-3 -> Phi-3-mini -> Mistral-7b
- 👎 Phi-3-mini ->Lama-3 -> Mistral-7b



Which are the best quality models overall?
(in decreasing order of quality)

- 👍 GPT-4 Turbo -> Claude-3 -> GPT-4o
- 👎 GPT-4o -> GPT-4 Turbo -> Claude-3

When working with a ChatBot application, which would be the better choice for a model?

- A 200B parameter model
- A 7B parameter model

The larger the context window, the better the model is able to recall information from larger texts

- True
- False



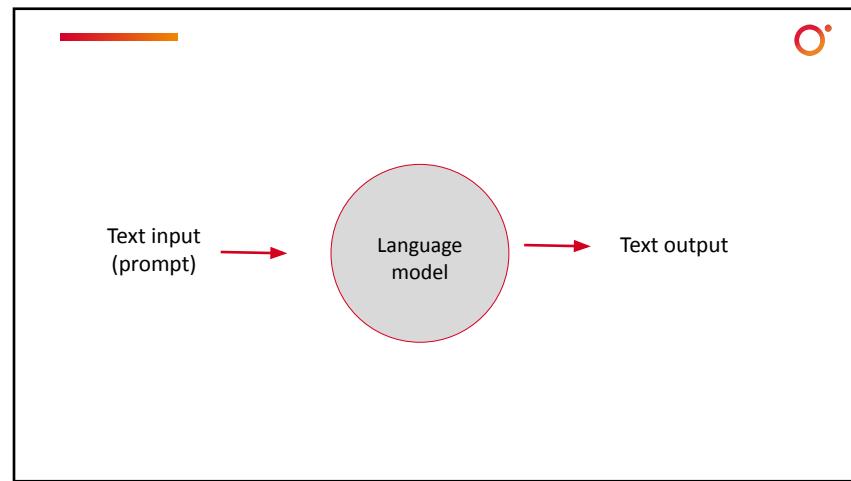
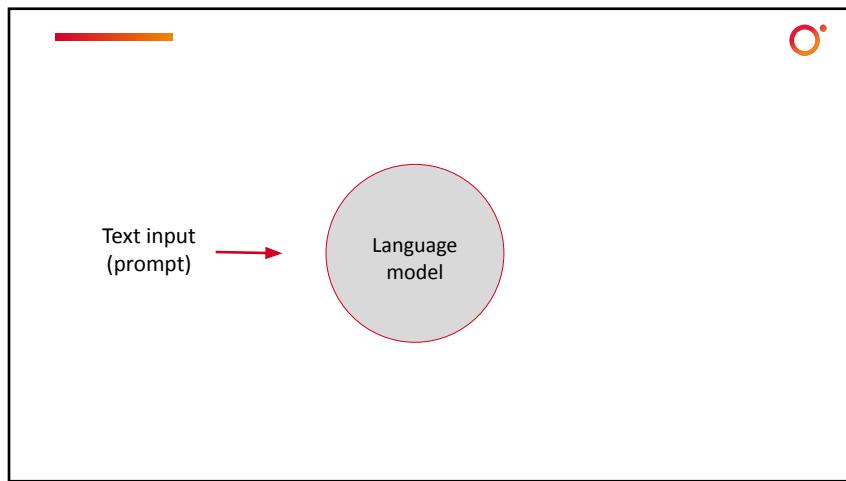
This online training will always be updated with the latest tools, techniques, and models.

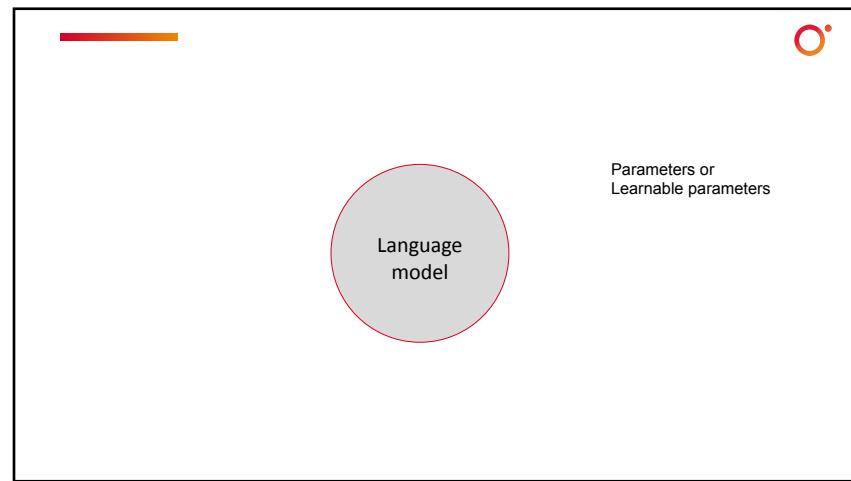
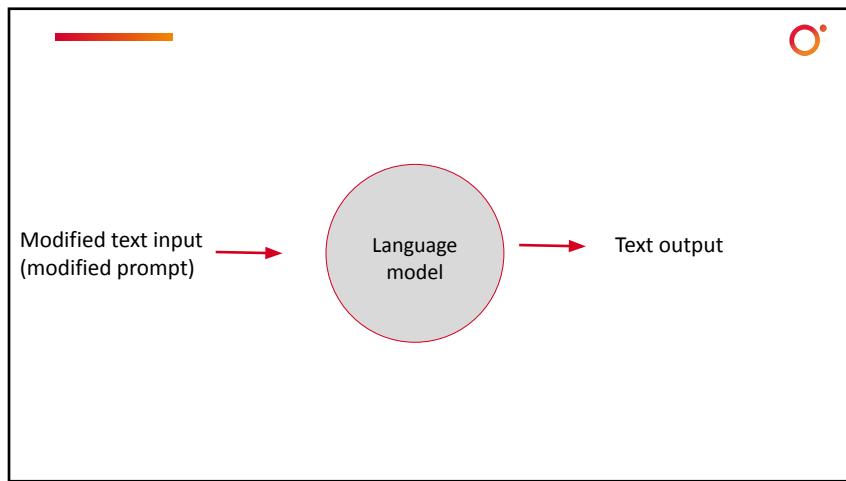
Today we will also cover:

GPT-4o (released on Monday)
Gemini updates (released Tuesday)
MMLU-Pro (released yesterday)

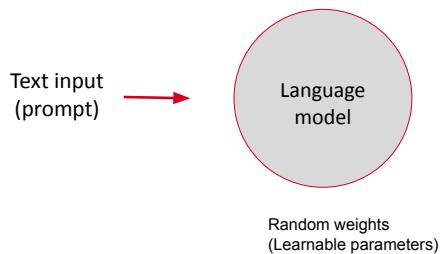


What Are Large Language Models and GPT?

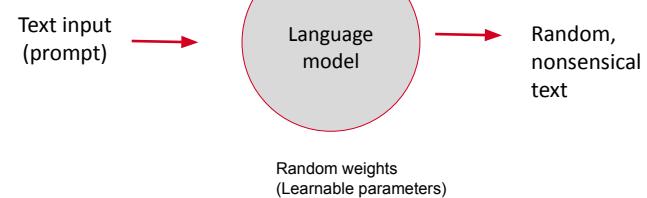




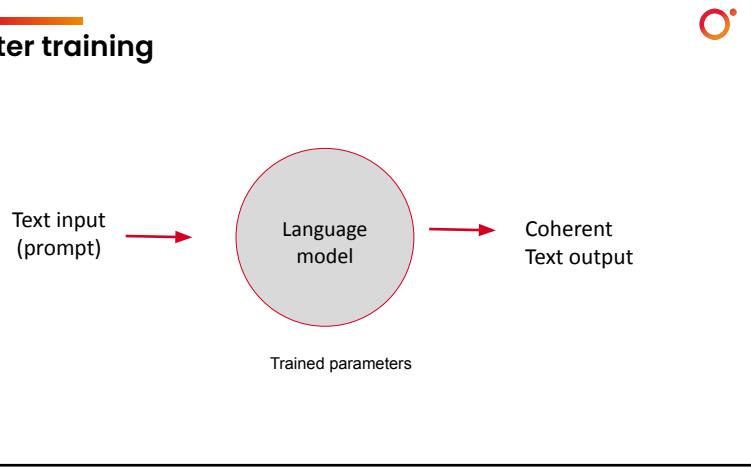
Initial model (Before training Language model)



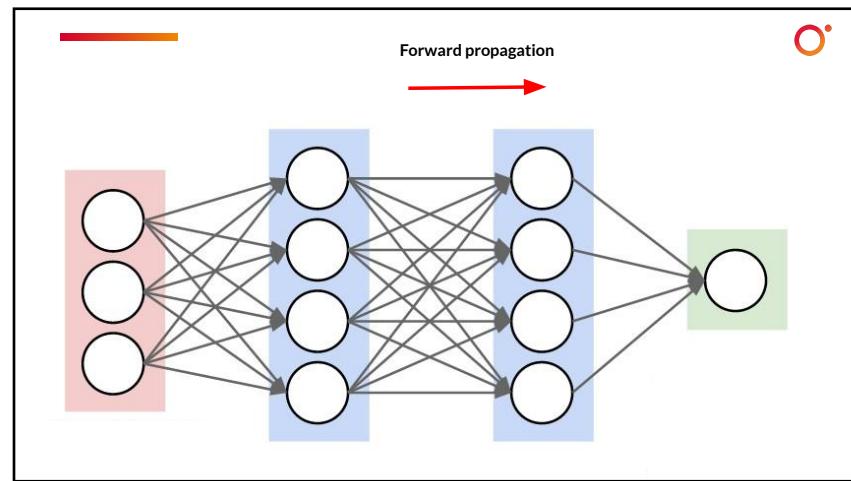
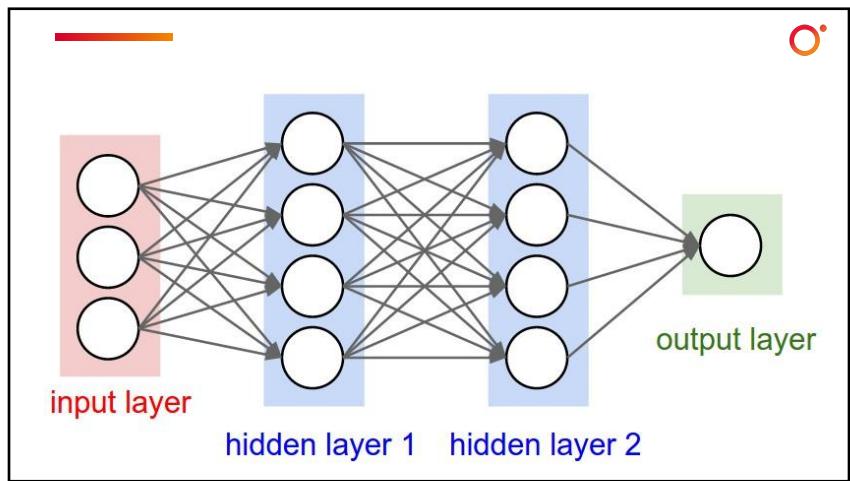
Initial model (Before training Language model)

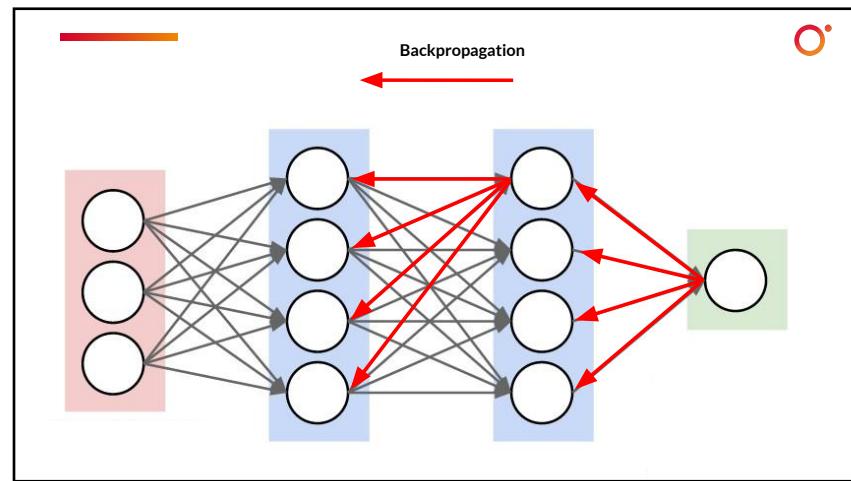
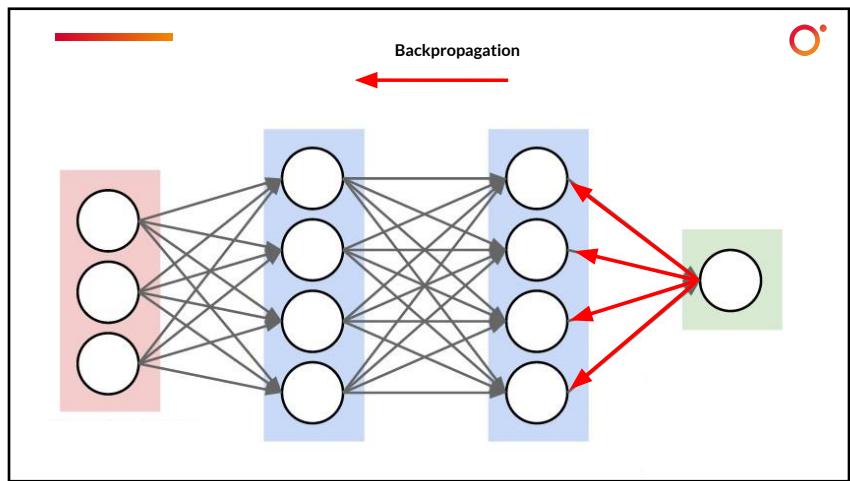


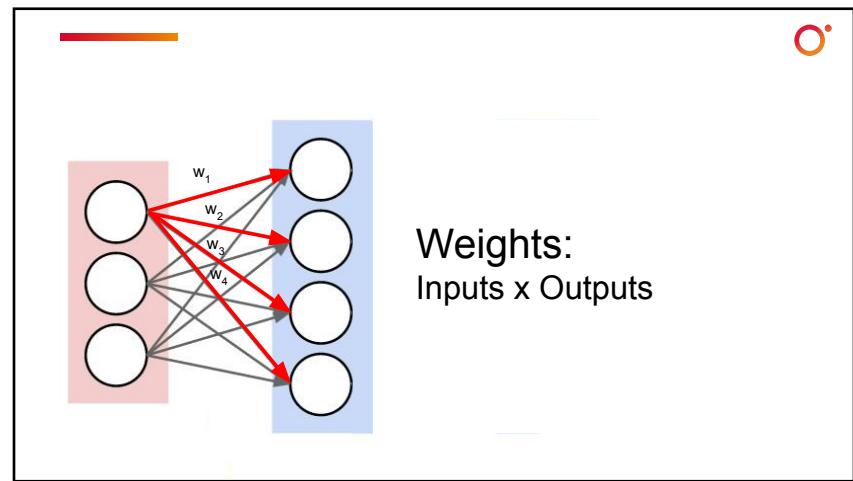
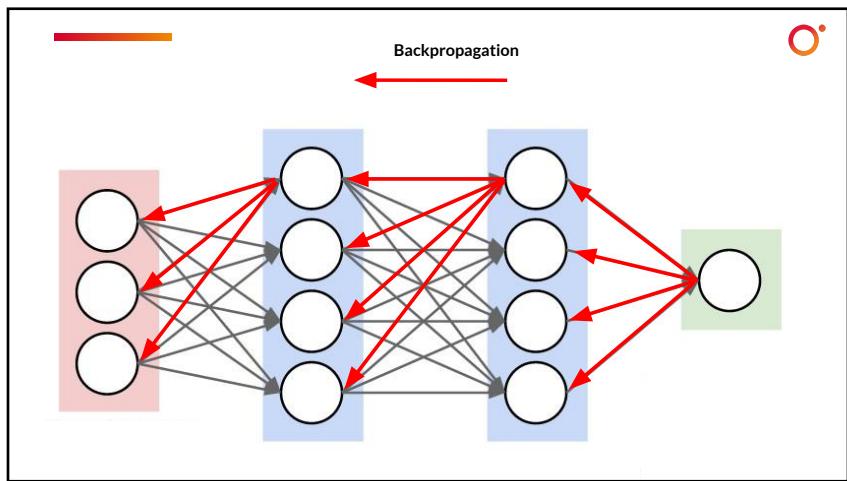
After training

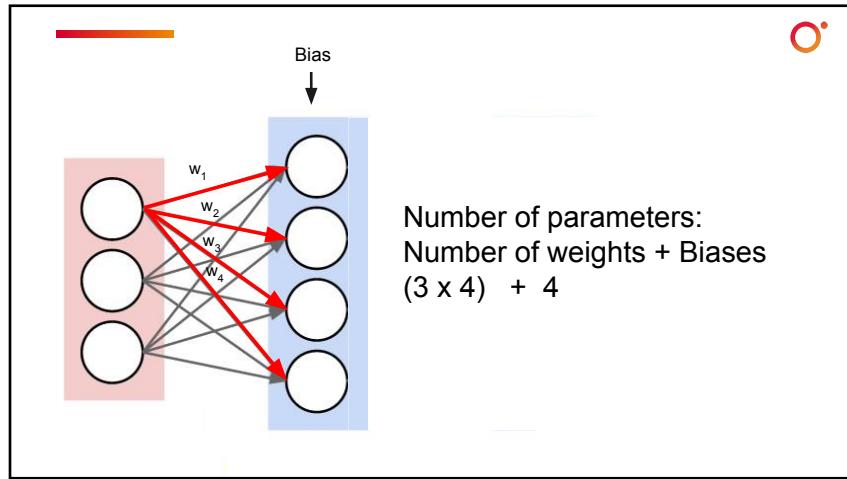
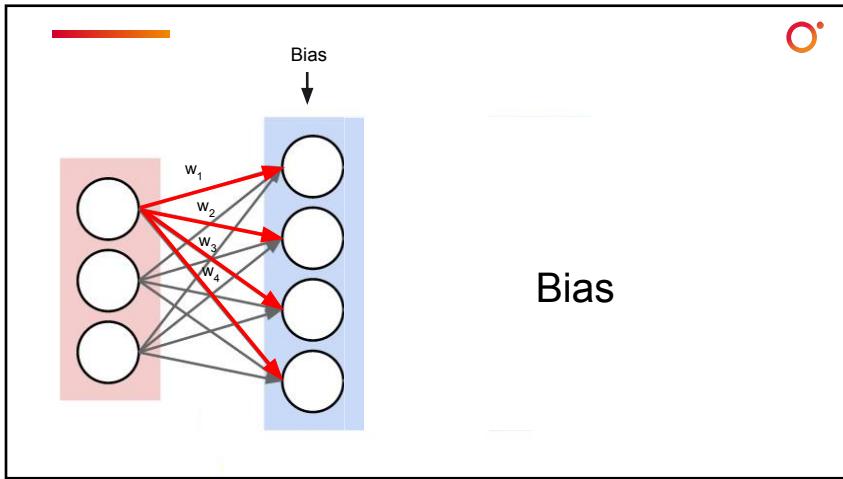


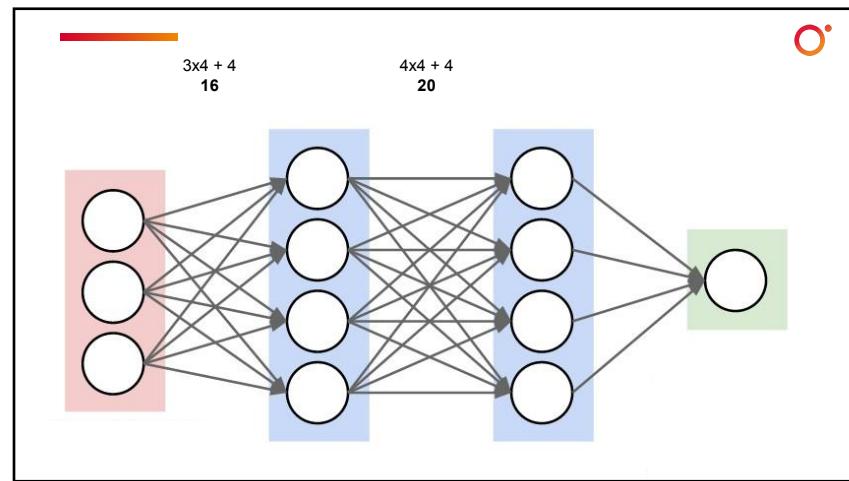
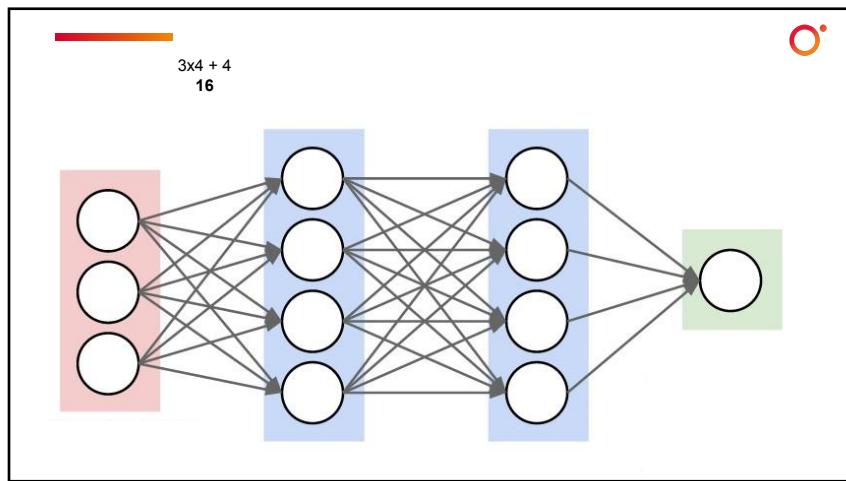
What are these (learnable) parameters?

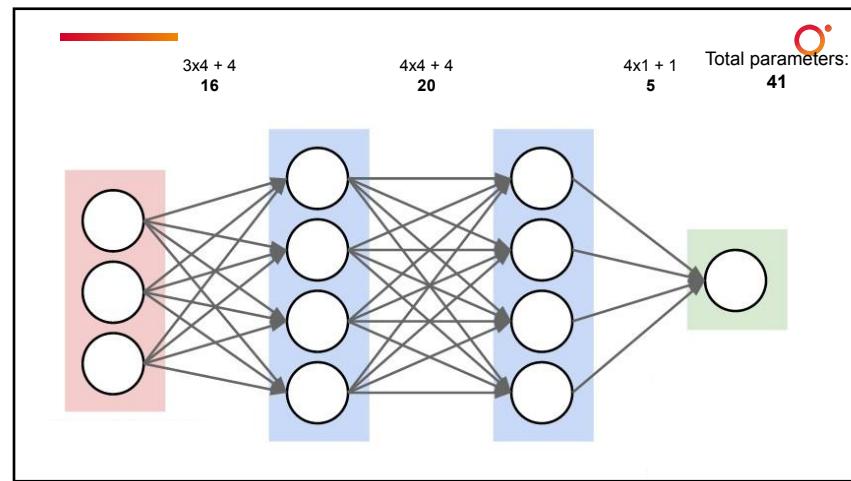
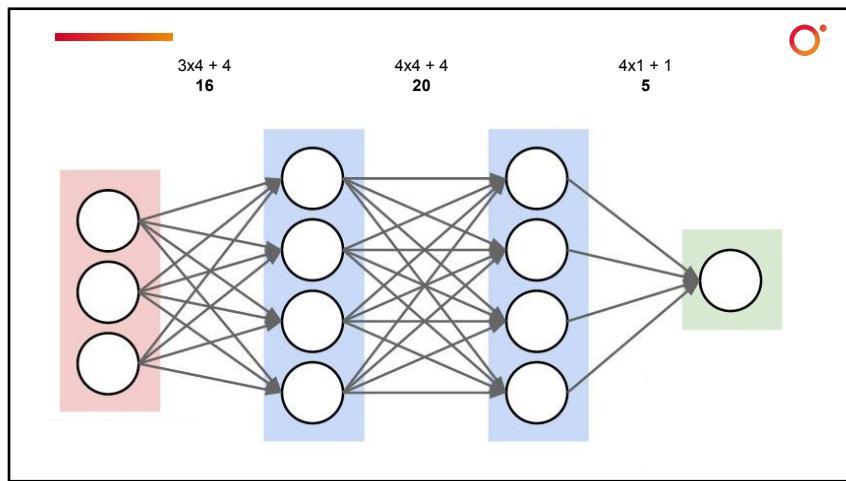












What are tokens?



What are tokens?

GPT-4o (coming soon) GPT-3.5 & GPT-4 GPT-3 (Legacy)

Tokenization involves decomposing words into smaller chunks or tokens.
These tokens are then converted into token ids

[Clear](#) [Show example](#)

Tokens
20

Characters
117

Tokenization involves decomposing words into smaller chunks or tokens.
These tokens are then converted into token ids

[Text](#) [Token IDs](#)



Context length



Context window / length

Prompt



Completion



Context window



Differences between models

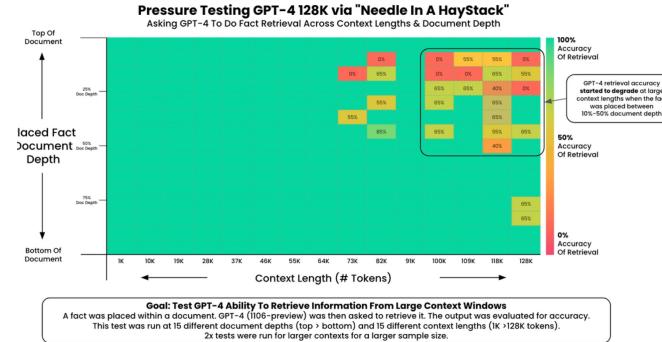
gpt-4: 8,000 tokens
gpt-4-32k: 32,000 tokens

gpt-3: 2,000 tokens
gpt-3.5: 4,000 tokens
gpt-4-turbo 128,000 tokens

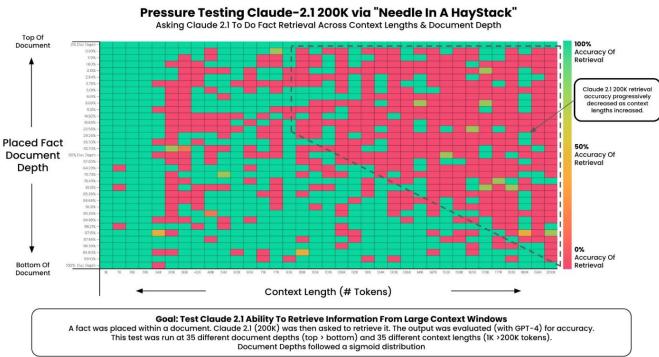
claude-2 100,000 tokens
claude-3 200,000 tokens

Gemini Pro: 1mil+ tokens
Gemini Pro 1.5: 2mil tokens

GPT-4-turbo 128k Context Window

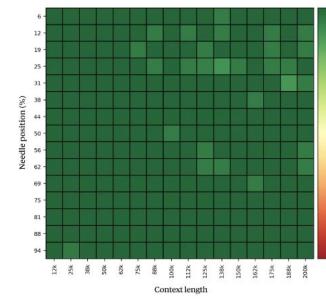


Claude-2.1 200k context window

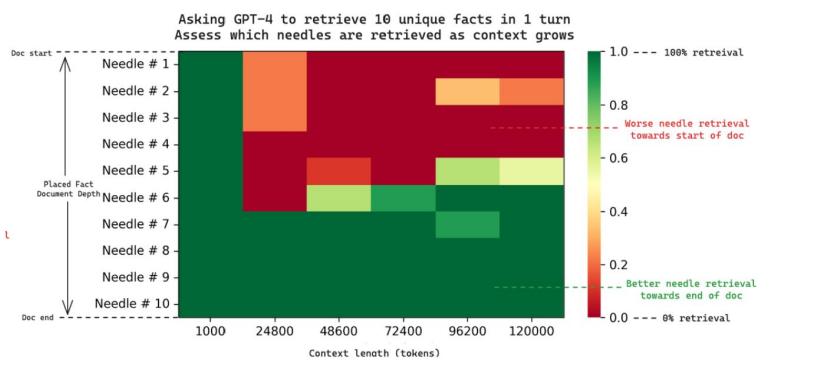


Claude-3

Claude 3 Opus
Recall accuracy over 200K
(averaged over many diverse document sources and 'needle' sentences)

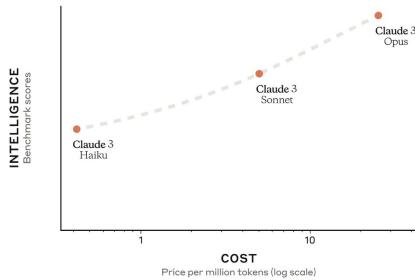


Multiple needles



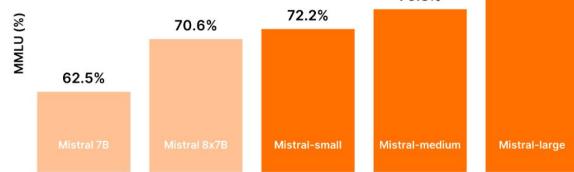
Model and Model Providers

Claude-3



Mistral

Mistral Models on MMLU



Llama-3



	Training Data	Params	Context length	GQA	Token count	Knowledge cutoff
Llama 3	A new mix of publicly available online data.	8B	8k	Yes	15T+	March, 2023
		70B	8k	Yes		December, 2023

Llama-3



	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	68.4	53.3	58.4
GPQA 0-shot	34.2	21.4	26.3
HumanEval 0-shot	62.2	30.5	36.6
GSM-BK 8-shot, CoT	79.6	30.6	39.9
MATH 4-shot, CoT	30.0	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	82.0	81.9	79.0
GPQA 0-shot	39.5	41.5 <small>CoT</small>	38.5 <small>CoT</small>
HumanEval 0-shot	81.7	71.9	73.0
GSM-BK 8-shot, CoT	93.0	91.7 <small>11-shot</small>	92.3 <small>0-shot</small>
MATH 4-shot, CoT	50.4	58.5 <small>Minerva prompt</small>	40.5

Llama-3

Meta Llama 3 400B+ (still training)
Checkpoint as of Apr 15, 2024

PRE-TRAINED		INSTRUCT	
MMLU 5-shot	84.8	Meta Llama 3 400B+	
AGIEval English 3-5-shot	69.9	MMLU 5-shot	86.1
BIG-Bench Hard 3-shot, CoT	85.3	GPQA 0-shot	48.0
ARC-Challenge 25-shot	96.0	HumanEval 0-shot	84.1
DROP 3-shot, F1	83.5	GSM-8K 8-shot, CoT	94.1
		MATH 4-shot, CoT	57.8



Llama-3 key takeaways

- **Architecture:** Dense decoder (not MoE)
- Meta has access to significant compute. Models trained for longer.
- **Compute:** Trained on 24k GPUs (max 16k concurrently) on 15 trillion-plus tokens.
- **Tokenizer:** Vocabulary size from 32K to 128K. Llama 3 uses fewer tokens.
- **Training Data:** Not released. 5% English.
- **Context window:** Increased from 4k to 8k (Llama-2)

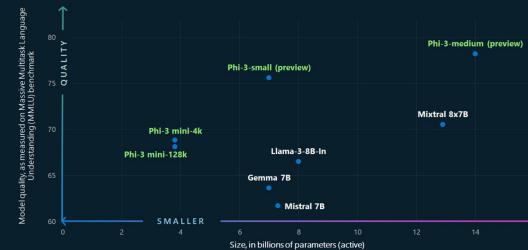


Phi-3-mini

- 3.8B parameter model
- 3.3M training tokens
- Performance rivals (Mixtral 8x7B and GPT-3.5 (e.g., phi-3-mini achieves 69% on MMLU and 8.38 on MT-bench)
- Key innovation - dataset for training
- phi-3-small (7B parameters) - trained on 4.8T tokens
- Phi-3-medium (14B parameters) - trained on 4.8T tokens

Phi-3-mini

Quality vs Size in Small Language Models (SLMs)



Gemini 1.5 Pro



Models and Model providers



Gemini Nano
The most efficient
model for on-device
tasks



Gemini Pro
The best model for
scaling across a wide
range of tasks



Gemini Ultra
The most capable and
largest model for highly-
complex task



Flash

Announced at
Google I/O this week

Flash
Where
response time
matters

Multi-modal large language models to date



GPT-4o



GPT-4o - top of leaderboards

Rank #	Model	Area	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1209	+6/-5	12088	OpenAI	Proprietary	2023/10
2	GPT-4-Turbo-2024-04-09	1252	+4/-2	59114	OpenAI	Proprietary	2023/12
2	GPT-4-110B-pxvsem	1209	+3/-3	51815	OpenAI	Proprietary	2023/4
2	Gemini-1.5-Pro-API-0409-Precise	1248	+3/-3	62929	Google	Proprietary	2023/11
3	Claude_3_Goals	1246	+2/-2	110342	Anthropic	Proprietary	2023/8
4	GPT-4-0125-pxreview	1244	+3/-2	75296	OpenAI	Proprietary	2023/12
7	Bard (Gemini_Pro)	1208	+5/-5	12387	Google	Proprietary	Online
7	Llama-3-70B-Instruct	1203	+3/-2	118388	Meta	Llama 3 Community	2023/12
8	Claude_3_Sonnet	1199	+3/-3	94086	Anthropic	Proprietary	2023/8
9	Reka-Core-20240501	1195	+4/-3	33583	Reka AI	Proprietary	Unknown
10	Qwen-Max-0428	1187	+5/-4	20794	Alibaba	Proprietary	Unknown
11	GPT-4-0114	1189	+3/-3	54137	OpenAI	Proprietary	2021/9
11	Command_R	1189	+2/-3	60238	Cohere	CC-BY-NC-4.0	2024/3
13	Claude_3_Haiku	1181	+3/-2	84423	Anthropic	Proprietary	2023/8
15	GwenL-5-110B-Chat	1171	+4/-5	15797	Alibaba	Qianwen LICENSE	2024/4

GPT-4o - tokenizer

<https://colab.research.google.com/drive/1QnVhdL3jQaCoFPKdWYc8WudrGmFJtlv?usp=sharing>

What are the benchmarks?

Gemini release

State-of-the-art performance

We are excited to bring our Gemini model and evaluating its performance on a wide variety of benchmarks. We have conducted a series of experiments to evaluate the quality of the generated responses across a range of benchmarks. Our results show that Gemini consistently outperforms other state-of-the-art models on most benchmarks.

In a series of 102N, Gemini often beats the models to complete human experts on tasks like image captioning, question answering, and language modeling. Gemini also shows strong performance on a variety of benchmarks, including the SQuAD, CoQA, and WikiQA datasets.

While Gemini is still learning, we believe it has the potential to become one of the best models available today. We welcome feedback for testing both word knowledge and problem-solving abilities.

Please let us know if you have any questions or concerns about Gemini's performance. We are always happy to answer your questions and help you get started with Gemini.

Category	Task	Model	Performance	GP
General	General	Gemini	95.4%	94.5%
Reasoning	Big Bench	Gemini	95.8%	95.3%
Math	CoQA	Gemini	95.9%	95.4%
Cross	HumanEval	Gemini	95.4%	95.0%
	MATIX	Gemini	95.7%	95.2%

Claude-3

A new standard for intelligence

Opus, our most intelligent model, outperforms its peers on most of the common evaluation benchmarks for AI systems, including undergraduate level expert knowledge (MMLT), general machine learning (GPAL), basic mathematics (GMARL), and more. It exhibits near-human levels of comprehension and fluency on complex tasks, leading the frontier of general intelligence.

All Claude-3 models have increased capabilities in analysis and forecasting, managed code creation, code generation, and conversing in non-English languages like Spanish, Japanese, and French.

Below is a comparison of the Claude-3 models to those of our peers on multiple benchmarks of capability:

	Claude-3 Opus	Claude-3 Seminar	Claude-3 Studio	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge (MMLT)	86.8%	79.0%	75.2%	86.4%	70.0%	83.7%	71.8%
Graduate level knowledge (GMARL)	90.4%	40.4%	33.3%	35.7%	28.7%	—	—
Grade-school math (GMARL)	95.0%	92.3%	88.9%	92.0%	57.7%	94.4%	88.5%
Math problem-solving (GMARL)	46.7%	43.1%	38.9%	52.9%	34.1%	53.2%	32.6%
Mathematical math (GMARL)	90.7%	88.8%	75.1%	74.8%	—	79.0%	69.8%
Code (Gemini)	84.9%	73.0%	75.9%	67.0%	48.1%	74.4%	63.7%



MMLU: Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s^2
(B) more than 9.8 m/s^2
(C) less than 9.8 m/s^2
(D) Cannot say unless the speed of throw is given.



MMLU: Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s^2 . If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s^2
(B) more than 9.8 m/s^2
(C) less than 9.8 m/s^2
(D) Cannot say unless the speed of throw is given.



MMLU: Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained.
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency.
(C) monopoly firms do not engage in significant research and development.
(D) consumer surplus is lost with higher prices and lower levels of output.



MMLU: Microeconomics



✗
✗
✗
✓

- One of the reasons that the government discourages and regulates monopolies is that
- (A) producer surplus is lost and consumer surplus is gained.
 - (B) monopoly prices ensure productive efficiency but cost society allocative efficiency.
 - (C) monopoly firms do not engage in significant research and development.
 - (D) consumer surplus is lost with higher prices and lower levels of output.

MMLU: Medicine



- A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?
- (A) Branch of the costocervical trunk
 - (B) Branch of the external carotid artery
 - (C) Branch of the thyrocervical trunk
 - (D) Tributary of the internal jugular vein

MMLU: Medicine



A 33-year-old man undergoes a radical thyroidectomy for thyroid cancer. During the operation, moderate hemorrhaging requires ligation of several vessels in the left side of the neck. Postoperatively, serum studies show a calcium concentration of 7.5 mg/dL, albumin concentration of 4 g/dL, and parathyroid hormone concentration of 200 pg/mL. Damage to which of the following vessels caused the findings in this patient?

- (A) Branch of the costocervical trunk
- (B) Branch of the external carotid artery
- (C) Branch of the thyrocervical trunk
- (D) Tributary of the internal jugular vein



EQ-Bench:

Ability of LLMs to understand complex emotions

EQ-Bench: An Emotional Intelligence Benchmark for Large Language Models

Samuel J. Paech

December 11, 2023

Abstract

We introduce EQ-Bench, a new benchmark designed to evaluate aspects of emotional intelligence in Large Language Models (LLMs). We assess the ability of LLMs to understand complex emotions and social interactions by asking them to predict the intensity of emotional states of characters in stories. The benchmark also includes a diversity of different emotion types across a range of models. We find that EQ-Bench correlates strongly with comprehensive multi-domain benchmarks like MMLU (Hendrycks et al., 2020) ($r=0.97$), indicating that we may be capturing similar aspects of emotional intelligence. The benchmark consists of 1000 prompts and 1000 answers to 40 English-language questions. We also provide open-source code for an automated benchmarking pipeline at <https://github.com/EQ-Bench/EQ-Bench> and a leaderboard at <https://eqbench.com>.

1 Introduction

Emotional intelligence (EI or, internally, EI) is a cornerstone of human cognition, influencing everything from decision-making to interpersonal interactions (Goleman, 1998). Pioneers in the field of emotional intelligence, Salovey and Mayer (1990) define EI as “The ability to monitor one’s own and others’ feelings, to discriminate among them, and to use this information to guide one’s thinking and action.” This was later broken down into four branches: perceiving emotions (non-verbally), using emotions, understanding emotions and managing emotions (Mayer & Salovey, 1997).



MT-Bench Score

Challenging multi-turn with GPT-4 grading

Question: If the FED buys bonds in the secondary market
(A) the money supply will increase. (B) the money supply will decrease. (C) the money supply will not be affected. (D) the discount rate will be affected. Output the answer only.

Assistant A:
The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:

- 1. Increasing the money supply
- 2. Increasing the money supply by lowering the discount rate
- 3. Increasing the money supply by increasing the discount rate and increasing the money supply.

Follow-up Question:

How does it affect my daily life? Give 3 examples.

Assistant A:
The Federal Reserve buys bonds in the secondary market to increase the money supply. This affects your daily life by:

1. Interest rates: The Federal Reserve's actions in the bond market can affect interest rates, which in turn can affect the cost of borrowing money. For example, if the Federal Reserve buys bonds and drives up demand for them, it can lead to lower interest rates, making loans and credit cards more affordable.
2. Inflation: As the money supply increases, it can lead to inflation, which is the general increase in prices. Higher inflation can negatively impact your daily life by making it more expensive to buy things like groceries, gas, and other necessities.
3. Employment: The Federal Reserve's actions in the bond market can also have an impact on employment. When the economy is strong and healthy, job actions lead to economic growth and increased demand for goods and services. It can create more job opportunities and boost employment.

GPT-4 Judgments:

Assistant A provided an incorrect response to the user's question about how the Federal Reserve buying bonds in the secondary market affects daily life. The answer given is repetitive and lacks relevant information.

On the other hand, **Assistant B** provided a relevant and accurate response to the user's question about the impact of the Federal Reserve's actions in the bond market on daily life, such as interest rates, inflation, and employment.

Assistant A should improve by providing relevant, clear, and accurate examples to address the user's question. **Assistant B**'s response is more helpful, relevant, and detailed.

MMLU-Pro (released yesterday)

A small percentage of MMLU is wrong/ambiguous. MMLU

MMLU is only 4 choices - MMLU-Pro is 10

<https://huggingface.co/datasets/TIGER-Lab/MMLU-Pro>

GPT-4o - MMLU-Pro

<https://colab.research.google.com/drive/1QnVhvdL3jQaCoFPKdWYc8WudrGmFJtIv?usp=sharing>



LLM Trade-offs



LLM tradeoffs

- Out-of-the-box quality (zero-shot prompting)
- Few-shot prompting (context length)
- Cost
- Speed
- Ability to fine-tune
- The license the AI model is released under



Best quality overall



Model	Quality	Speed	Can you fine-tune?	Number of parameters	Context window
GPT-4 Turbo	****	***	No	Unknown	128k
GPT-4	***	**	Limited access	Unknown	8k / 32k
GPT-4o	*****	****	No	Unknown	128k
Claude-3 (Opus)	****	***	No	Unknown	200k
Claude-3 (Sonnet)	***	****	No	Unknown	200k
Gemini 1.5 Pro	****	***	No	Unknown	1 mil
Gemini-1 Flash	***	***	No	Unknown	1 mil
Mixtral-8x22b	***	**	?	39B (141B)	2k
Llama-3	****	**	Yes	8B	8k
Phi-3-mini	*****	****	?	7B	8k

Speed

Model	Quality	Speed	Can you fine-tune?	Number of parameters	Context window
GPT-4 Turbo	****	***	No	Unknown	128k
GPT-4	***	**	Limited access	Unknown	8k / 32k
GPT-4o	*****	****	No	Unknown	128k
Claude-3 (Opus)	****	***	No	Unknown	200k
Claude-3 (Sonnet)	***	****	No	Unknown	200k
Gemini 1.5 Pro	****	***	No	Unknown	1 mil
Gemini-1 Flash	***	***	No	Unknown	1 mil
Mixtral-8x22b	***	**	?	39B (141B)	2k
Llama-3	****	**	Yes	8B	8k
Phi-3-mini	****	****	?	7B	8k

Anything but OpenAI

Model	Quality	Speed	Can you fine-tune?	Number of parameters	Context window
GPT-4 Turbo	****	***	No	Unknown	128k
GPT-4	***	**	Limited access	Unknown	8k / 32k
GPT-4o	*****	****	No	Unknown	128k
Claude-3 (Opus)	****	***	No	Unknown	200k
Claude-3 (Sonnet)	***	****	No	Unknown	200k
Gemini 1.5 Pro	****	***	No	Unknown	1 mil
Gemini-1 Flash	***	***	No	Unknown	1 mil
Mixtral-8x22b	***	**	?	39B (141B)	2k
Llama-3	****	**	Yes	8B	8k
Phi-3-mini	****	****	?	7B	8k

Open models

GPT-4 Turbo	****	***	No	Unknown	128k
GPT-4	***	**	Limited access	Unknown	8k / 32k
GPT-4o	*****	****	No	Unknown	128k
Claude-3 (Opus)	****	***	No	Unknown	200k
Claude-3 (Sonnet)	***	****	No	Unknown	200k
Gemini 1.5 Pro	***	***	No	Unknown	1 mil
Gemini-1 Flash	***	***	No	Unknown	1 mil
Mixtral-8x22b	***	**	?	39B (141B)	2k
Llama-3	****	**	Yes	8B	8k
Phi-3-mini	****	****	?	7B	8k



Open-source or proprietary?

What are the considerations?



OSS licensing

Non-commercial license

- Llama models (Meta)
- Creative Commons CC BY-NC-SA 4.0

Restricted license

- CC BY-SA- 3.0 restrictions on commercial use

Permissive license

- Apache 2.0



Open models

T	A	Model	Average	A	ARC	A	HellaSG	A	MMLU	A	TruthfulQA	A	WinoGrande	A	GSM8K	A	Hub License
◆	81.22	decidit/llm20B/Hfme-72b-v0.5	79.70	91.15	77.95	74.5	87.85	76.12	apache-2.0								
🕒	81	HTS432/multiVerse_7B	78.67	89.77	78.22	79.18	87.53	76.05	other								
🕒	88.98	HTS432/multiVerse_7B	78.58	89.74	78.27	79.89	87.37	76.8	other								
◆	88.48	abecassis/7mug-72B-v0.1	76.82	89.27	77.15	76.67	85.98	78.7	other								
◆	79.3	1bit1bit/alpaca-dragon-72B-v3	73.89	88.16	77.4	72.69	86.83	77.63	other								
🕒	79.19	microsoft/MiGral-8x22B-Instruct-v0.1	72.7	89.08	77.77	66.14	85.16	82.93	apache-2.0								
🕒	78.96	MaizuyaPanda/Llama-3-7B8-Instruction-PP0-v0.2	72.53	86.22	80.41	63.57	82.79	88.25	llama3								
🕒	78.89	MaizuyaPanda/Llama-3-7B8-Instruction-PP0-v0.8	72.46	86.03	80.5	63.26	83.58	87.34	llama3								
🕒	78.74	MaizuyaPanda/Llama-3-7B8-Instruction-PP0-v0.3	72.36	86	80.47	63.45	82.95	87.19	llama3								
🕒	78.6	menpo/Llama-3-29B-Japanese-Surume-Vector-v0.2	72.35	85.81	80.28	63.93	82.79	87.41	llama3								
🕒	78.55	nozeh/PM06-72B-Lora-1.6-7-PP0	70.82	85.96	77.13	74.71	84.96	78.62	mit								
...

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Pricing

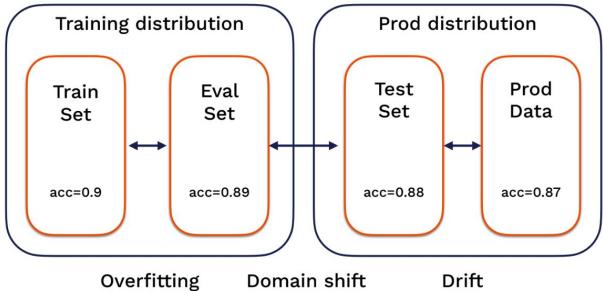
https://docs.google.com/spreadsheets/d/1NX8ZW9Jnfpv88PC2d6Bwla87JRiv3GTeqwXoB4mKU_s/edit#gid=0



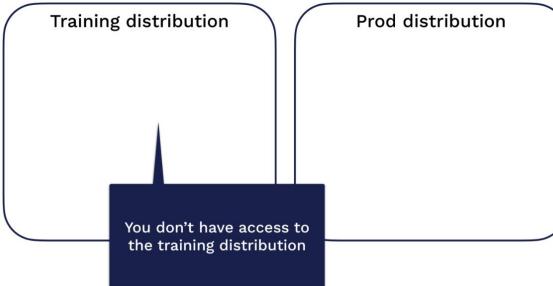
Why comparing LLMs is difficult



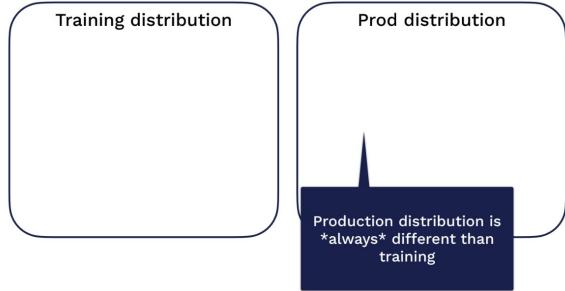
Using Traditional ML techniques



Why this won't work for LLMs



Why this won't work for LLMs



Why this won't work for LLMs

Traditional ML

```
pred=["cat", "dog", "dog", "cat", "dog", "cat", "dog", "cat", "dog"]
label=["dog", "dog", "dog", "cat", "dog", "cat", "dog", "dog", "cat", "dog"]
```

► acc=0.9

Generative

```
pred=["this is an image of a tabby cat"]
label="photograph of a cat"]
```

► What metric?

It's hard to define quantitative metrics

Comparing Large Language Models



Holistic Evaluation of
Language Models



HELM

HELM Paper - <https://arxiv.org/pdf/2211.09110.pdf>

HELM results - https://crfm.stanford.edu/helm/latest/?group=core_scenarios

Limitations of HELM

- Feature completeness and fine-tuning
- Price
- Latency
- Platform uptime



Jonathan A. Fernandes • You

1w • ④

This is one of the worst-kept secrets in the field of Generative AI and Large Language models.

Imagine preparing for a test and having the exact questions beforehand. That's essentially what happens in the AI world.

The models often get trained on materials that they'll be tested on in benchmarks like HELM. So, when they ace these tests, it's impressive but not entirely surprising.

That's why it's always important to not only consider different benchmarks as a guiding light but to have your own set of tests that makes sense for your business for you to determine if a Large Language Model performs well for your business.

OpenLLM Leaderboard

Model	Average	ARC	HellaSwag	MMLU	TruthfulQA
ValiantLabs/ShiningValiant	74.17	72.95	87.88	79.97	64.88
ICBU-NPU/FashionGPT-70B-v1.2	74.11	73.04	88.15	79.11	65.15
sequelbox/StellarEight	74.1	72.95	87.82	71.17	64.46
Riiid/sheep-duck-llama-2-70B-v1.1	74.07	73.04	87.81	78.84	64.58
AIDC-ai-business/Marcozoni-70B-v1	74.06	73.55	87.62	70.67	64.41
ICBU-NPU/FashionGPT-70B-v1.1	74.05	71.76	88.2	70.99	65.26
adonlee/LLaMA_2_70B_LoRA	73.9	72.7	87.55	70.84	64.52
uni-tianyan/Uni-TianYan	73.81	72.1	87.4	69.91	65.81
Riiid/sheep-duck-llama-2	73.69	72.35	87.78	70.82	63.8
Riiid/sheep-duck-llama-2	73.67	72.27	87.78	70.81	63.8
fangloveskari/RCRA_LLaMA_70B_QLoRA	73.4	72.27	87.74	70.23	63.37
ICBU-NPU/FashionGPT-70B-v1	73.26	71.08	87.32	70.7	63.92

https://huggingface.co/spaces/HuggingFaceH4/open_llm_leaderboard

Getting started for your application



A system for tracking prompts

- Naive - Prompt history in playground
- Storing them in a spreadsheet
- Track prompts in git
- Solution for non-technical stakeholders





How do you compare LLMs?

- Benchmarks
- Company objectives and requirements
 - Chatbot
 - Must reference company documentation
 - Max size of response so can be used/read by agent while on call



Build a dataset for your task

Chatbot arena

The screenshot shows a comparison between two models, Model A and Model B. Both models respond to the input "Correct grammar: I are happy." with "Correct grammar: I are happy." Model A's response is followed by a message asking for more sentences to correct. Model B's response is followed by a message providing the correct grammar ("The correct grammar would be 'I am happy.'"). Below the messages are four buttons: "A is better", "B is better", "Tie", and "Both are bad".

LMSYS Chatbot Arena Leaderboard

Rank# (UB)	Model	Arena Elo	95% CI	Votes	Organization	License	Knowledge Cutoff
1	GPT-4o-2024-05-13	1289	+6/-5	12808	OpenAI	Proprietary	2023/10
2	GPT-4-318K-preview	1250	+3/-3	81015	OpenAI	Proprietary	2023/4
2	GPT-4-Turbo-2024-04-09	1252	+4/-2	59114	OpenAI	Proprietary	2023/12
2	Gemini_1.5_Pro_API-840K-Preslim	1248	+3/-3	62929	Google	Proprietary	2023/11
3	Claude_3.0ous	1246	+2/-2	116342	Anthropic	Proprietary	2023/8
4	GPT-4-6128-preview	1244	+3/-2	75206	OpenAI	Proprietary	2023/12
7	Bard_Gemini_Pro	1208	+5/-5	12387	Google	Proprietary	Online
7	Llama-3-70b-Instruct	1203	+3/-2	118358	Meta	Llama 3 Community	2023/12
8	Claude_3_Sonnet	1199	+3/-3	94050	Anthropic	Proprietary	2023/8
9	Reka-Core-20240501	1195	+4/-3	33583	Reka AI	Proprietary	Unknown
10	Deno-Max-0428	1187	+5/-4	29794	Alibaba	Proprietary	Unknown
11	GPT-4-0314	1189	+3/-3	54137	OpenAI	Proprietary	2021/9
11	Command_R+	1189	+3/-3	60230	Cohere	CC-BY-NC-4.0	2024/3

<https://huggingface.co/spaces/lmsys/chatbot-area-leaderboard>

Compare models side by side

[Nat.dev](#)



Elo Ratings

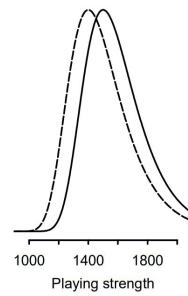


Elo rating

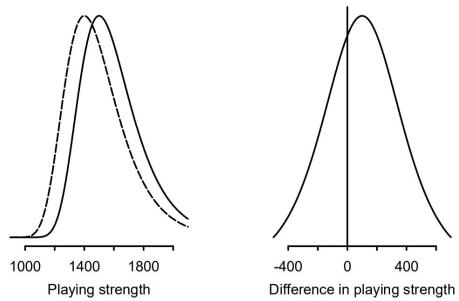
- Arpad Elo (Hungarian-American physics professor)
- Used in chess but also in other games (football (soccer), basketball, american football, baseball, board games)
- Not measured in absolute terms. Inferred from wins, losses, and draws against other players.



Elo rating



Elo rating



Expected scores given Elo rating

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}$$

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/400}}$$



Updating a player's Elo rating

$$R'_A = R_A + K(S_A - E_A)$$



Elo Rating Lab



Elo Rating Notebook

https://colab.research.google.com/drive/1rSGJq_kQNZ-tMafcZHE2CXESEZBPeJUE?usp=sharing



Wrapping up



 Live Course



Hands-on Retrieval Augmented Generation (RAG)

With [Jonathan Fernandes](#)

⌚ 3h 0m 🗓 June 27 • 5pm-8pm GMT+1

 Live Course



Hands-on GPT-4-Turbo

With [Jonathan Fernandes](#)

⌚ 3h 0m 🗓 June 20 • 5pm-8pm GMT+1



What we covered today

Fundamentals of LLMs

- What are LLMs?
- Parameters
- Tokens
- Context length
- LLM trade-offs
- Getting started for your application
- Why are comparing LLMs difficult?

Benchmarking LLMs

- HELM
- OpenLLM Leaderboard
- Using your own data
- Q&A
- Break

Elo Ratings

- Comparing LLMs
- Notebook
- Exercise
- Q&A



The logo consists of a solid black rectangular background containing a graphic element. This graphic element features three overlapping circles: a large orange circle at the bottom left, a smaller red circle above it to the right, and a medium-sized red circle above the orange one. To the right of this graphic, the word "O'REILLY" is written in a bold, white, sans-serif font. A registered trademark symbol (®) is positioned at the top right of the "Y".

O'REILLY®