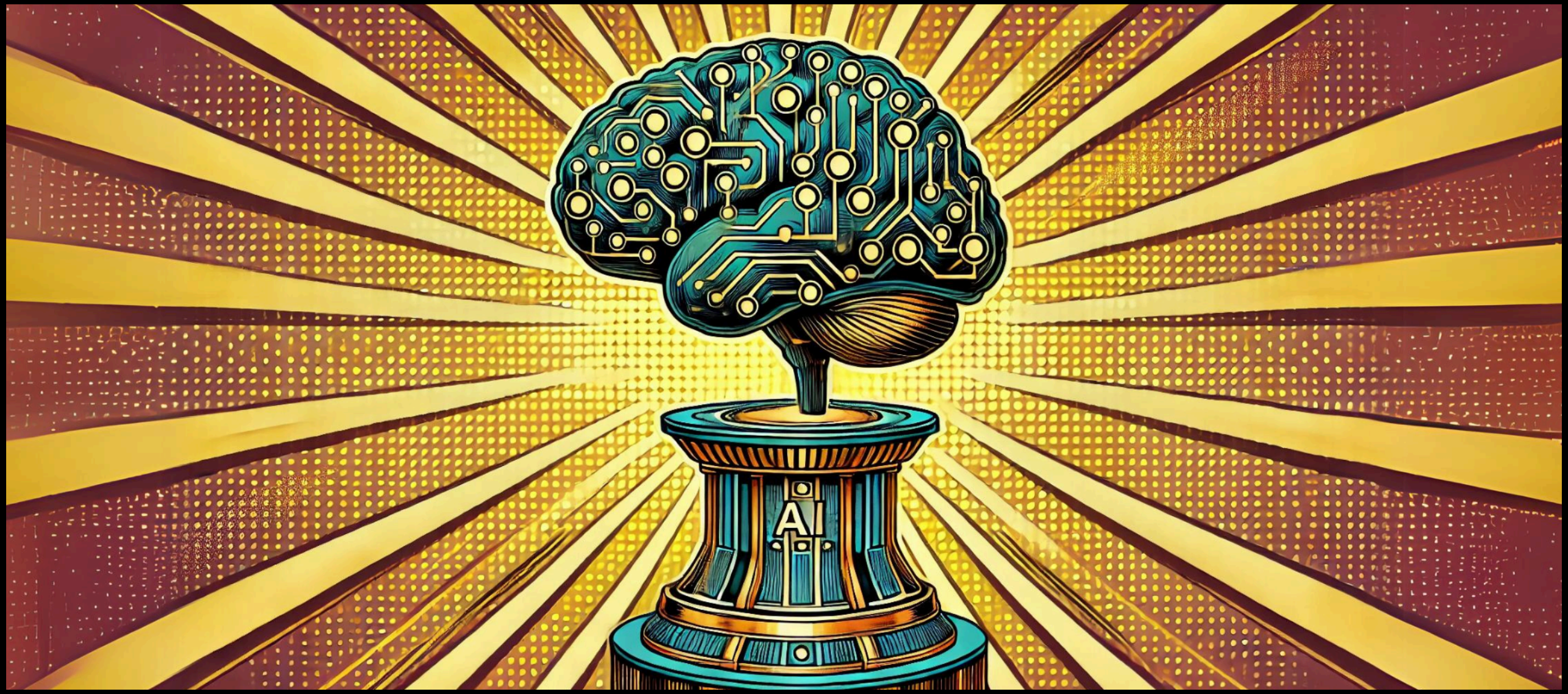
LLM Engineering

MASTER AI & LARGE LANGUAGE MODELS



WEEK 4 DAY 1

An essential week of building expertise

What you can now do

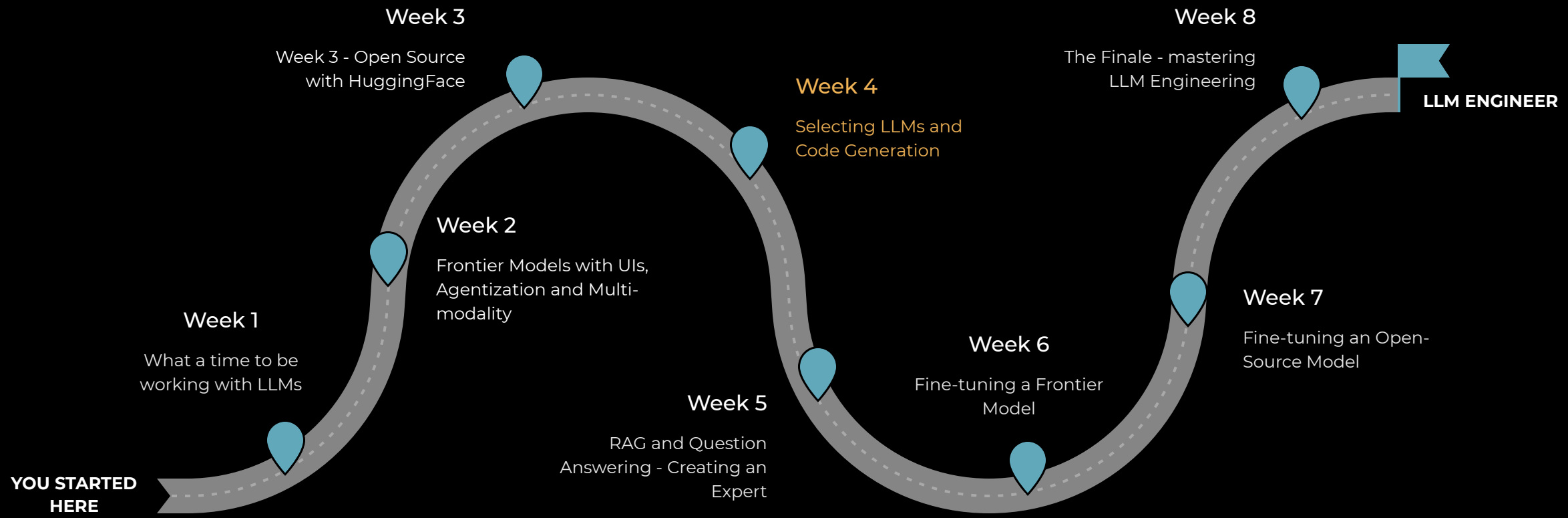
- Confidently code with Frontier Models
- Build a multi-modal AI Assistant with Tools
- Build solutions with open-source LLMs with HuggingFace transformers

---

After today you will be able to

- Discuss how to select the right LLM for the task
- Compare LLMs based on their basic attributes and benchmarks
- Use the Open LLM Leaderboard to evaluate LLMs

Reminder of the 8 weeks to mastery



How to compare LLMs

Importantly, LLMs need to be evaluated for suitability **for a given task**

Start with the basics

Parameters

Context length

Pricing

Then look at the results

Benchmarks

Leaderboards

Arenas

COMPARING LLMs

The Basics (1)

Compare the following features of an LLM:

- Open-source or closed
- Release date and knowledge cut-off
- Parameters
- Training tokens
- Context length

COMPARING LLMs

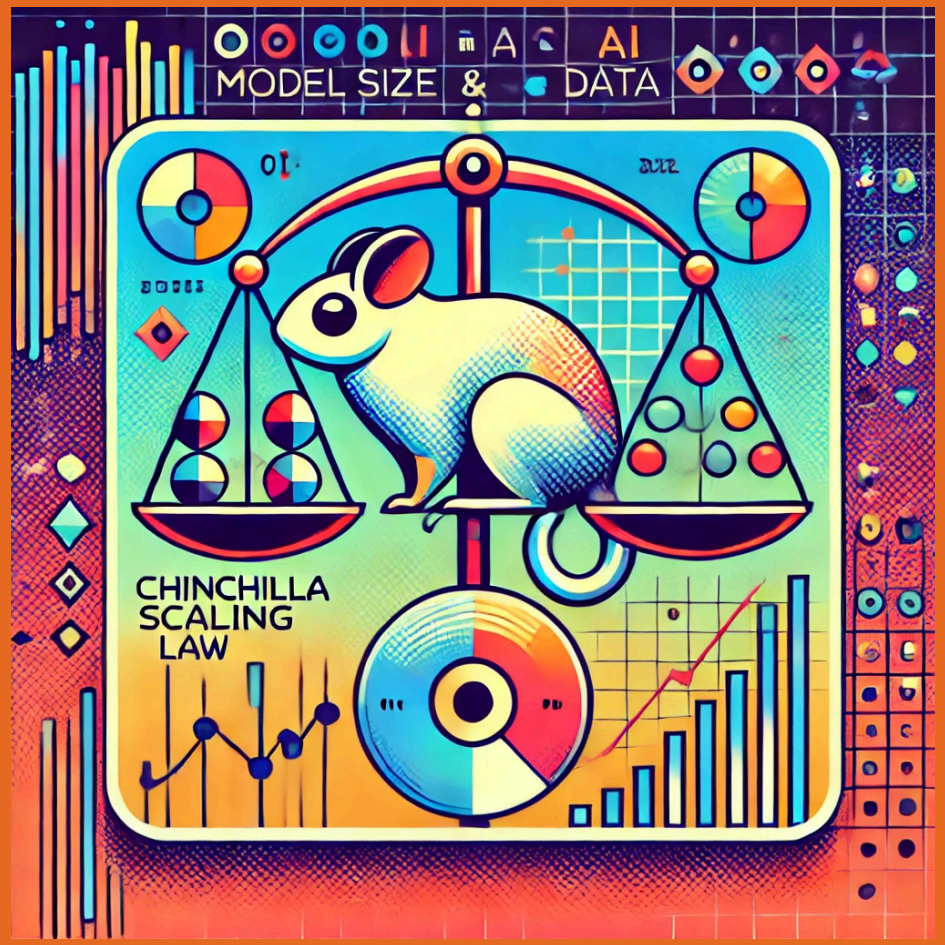
The Basics (2)

Compare the following features of an LLM:

- Inference cost
  API charge, Subscription or Runtime compute
- Training cost
- Build cost
- Time to Market
- Rate limits
- Speed
- Latency
- License

COMPARING LLMs

The Chinchilla Scaling Law



Number of parameters ~ proportional to the number of training tokens

---

If you're getting diminishing returns from training with more training data, then this law gives you a rule of thumb for scaling your model

---

And vice versa: if you upgrade to a model with double the number of weights, this law indicates your training data requirement

7 common benchmarks that you will often encounter

Benchmark	What's being evaluated	Description
ARC	Reasoning	A benchmark for evaluating scientific reasoning; multiple-choice questions

7 common benchmarks that you will often encounter

Benchmark	What's being evaluated	Description
ARC	Reasoning	A benchmark for evaluating scientific reasoning; multiple-choice questions
DROP	Language Comp	Distill details from text then add, count or sort

7 common benchmarks that you will often encounter

Benchmark	What's being evaluated	Description
ARC	Reasoning	A benchmark for evaluating scientific reasoning; multiple-choice questions
DROP	Language Comp	Distill details from text then add, count or sort
HellaSwag	Common Sense	"Harder Endings, Long Contexts and Low Shot Activities"

7 common benchmarks that you will often encounter

Benchmark	What's being evaluated	Description
ARC	Reasoning	A benchmark for evaluating scientific reasoning; multiple-choice questions
DROP	Language Comp	Distill details from text then add, count or sort
HellaSwag	Common Sense	"Harder Endings, Long Contexts and Low Shot Activities"
MMLU	Understanding	Factual recall, reasoning and problem solving across 57 subjects

7 common benchmarks that you will often encounter

Benchmark	What's being evaluated	Description
ARC	Reasoning	A benchmark for evaluating scientific reasoning; multiple-choice questions
DROP	Language Comp	Distill details from text then add, count or sort
HellaSwag	Common Sense	"Harder Endings, Long Contexts and Low Shot Activities"
MMLU	Understanding	Factual recall, reasoning and problem solving across 57 subjects
TruthfulQA	Accuracy	Robustness in providing truthful replies in adversarial conditions

7 common benchmarks that you will often encounter

Benchmark	What's being evaluated	Description
ARC	Reasoning	A benchmark for evaluating scientific reasoning; multiple-choice questions
DROP	Language Comp	Distill details from text then add, count or sort
HellaSwag	Common Sense	"Harder Endings, Long Contexts and Low Shot Activities"
MMLU	Understanding	Factual recall, reasoning and problem solving across 57 subjects
TruthfulQA	Accuracy	Robustness in providing truthful replies in adversarial conditions
Winogrande	Context	Test the LLM understands context and resolves ambiguity

7 common benchmarks that you will often encounter

Benchmark	What's being evaluated	Description
ARC	Reasoning	A benchmark for evaluating scientific reasoning; multiple-choice questions
DROP	Language Comp	Distill details from text then add, count or sort
HellaSwag	Common Sense	"Harder Endings, Long Contexts and Low Shot Activities"
MMLU	Understanding	Factual recall, reasoning and problem solving across 57 subjects
TruthfulQA	Accuracy	Robustness in providing truthful replies in adversarial conditions
Winogrande	Context	Test the LLM understands context and resolves ambiguity
GSM8K	Math	Math and word problems taught in elementary and middle schools

3 specific benchmarks

Benchmark	What's being evaluated	Description
ELO	Chat	Results from head-to-head face-offs with other LLMs, as with ELO in Chess

3 specific benchmarks

Benchmark	What's being evaluated	Description
ELO	Chat	Results from head-to-head face-offs with other LLMs, as with ELO in Chess
HumanEval	Python Coding	164 problems writing code based on docstrings

3 specific benchmarks

Benchmark	What's being evaluated	Description
ELO	Chat	Results from head-to-head face-offs with other LLMs, as with ELO in Chess
HumanEval	Python Coding	164 problems writing code based on docstrings
MultiPL-E	Broader Coding	Translation of HumanEval to 18 programming languages

COMPARING LLMs

Limitations of Benchmarks

- Not consistently applied
- Too narrow in scope
- Hard to measure nuanced reasoning
- Training data leakage
- Overfitting

And a new concern, not yet proven

- Frontier LLMs may be aware that they are being evaluated

6 Hard, Next-Level Benchmarks

Benchmark	What's being evaluated	Description
GPQA	Graduate Tests	448 expert questions; non-PhD humans score 34% even with web access

6 Hard, Next-Level Benchmarks

Benchmark	What's being evaluated	Description
GPQA	Graduate Tests	448 expert questions; non-PhD humans score 34% even with web access
BBHard	Future Capabilities	204 tasks believed beyond capabilities of LLMs (no longer!)

6 Hard, Next-Level Benchmarks

Benchmark	What's being evaluated	Description
GPQA	Graduate Tests	448 expert questions; non-PhD humans score 34% even with web access
BBHard	Future Capabilities	204 tasks believed beyond capabilities of LLMs (no longer!)
Math Lv 5	Math	High-school level math competition problems

6 Hard, Next-Level Benchmarks

Benchmark	What's being evaluated	Description
GPQA	Graduate Tests	448 expert questions; non-PhD humans score 34% even with web access
BBHard	Future Capabilities	204 tasks believed beyond capabilities of LLMs (no longer!)
Math Lv 5	Math	High-school level math competition problems
IFEval	Difficult instructions	Like, "write more than 400 words" and "mention AI at least 3 times"

6 Hard, Next-Level Benchmarks

Benchmark	What's being evaluated	Description
GPQA	Graduate Tests	448 expert questions; non-PhD humans score 34% even with web access
BBHard	Future Capabilities	204 tasks believed beyond capabilities of LLMs (no longer!)
Math Lv 5	Math	High-school level math competition problems
IFEval	Difficult instructions	Like, "write more than 400 words" and "mention AI at least 3 times"
MuSR	Multistep Soft Reasoning	Logical deduction, such as analyzing 1,000 word murder mystery and answering: "Who has means, motive and opportunity?"

6 Hard, Next-Level Benchmarks

Benchmark	What's being evaluated	Description
GPQA	Graduate Tests	448 expert questions; non-PhD humans score 34% even with web access
BBHard	Future Capabilities	204 tasks believed beyond capabilities of LLMs (no longer!)
Math Lv 5	Math	High-school level math competition problems
IFEval	Difficult instructions	Like, "write more than 400 words" and "mention AI at least 3 times"
MuSR	Multistep Soft Reasoning	Logical deduction, such as analyzing 1,000 word murder mystery and answering: "Who has means, motive and opportunity?"
MMLU-PRO	Harder MMLU	A more advanced and cleaned up version of MMLU including choice of 10 answers instead of 4

The HuggingFace Open LLM Leaderboard

- Navigating model types and parameters
- Understanding the benchmarks, finding our models

T	Model	Average ⬆️	IFEval	BBH	MATH Lvl 5	GPQA	MUSR	MMLU-PRO
🟢	Qwen/Qwen2-72B	35.13	38.24	51.86	29.15	19.24	19.73	52.56
🟢	Qwen/Qwen1.5-110B	29.56	34.22	44.28	23.04	13.65	13.71	48.45
🟢	dnhkng/RYS-Phi-3-medium-4k-instruct	28.38	43.91	46.75	11.78	13.98	11.09	42.74
🟢	Qwen/Qwen1.5-32B	26.69	32.97	38.98	26.66	10.63	12.04	38.89
🟢	01-ai/Yi-1.5-34B-32K	26.4	31.19	43.38	13.44	15.1	14.08	41.21
🟢	meta-llama/Meta-Llama-3-70B	26.37	16.03	48.71	16.54	19.69	16.01	41.21
🟩	dnhkng/RYS-Medium	25.94	44.06	47.73	7.78	10.4	8.73	36.96
🟢	meta-llama/Meta-Llama-3.1-70B	25.91	16.84	46.4	16.69	18.34	16.58	40.6
🟢	mistral-community/mixtral-8x22B-v0.3	25.55	25.83	45.73	16.84	17	7.46	40.44

Note: only covers Open Source models - we will look at leaderboards combining open and closed source next time

WEEK 4 DAY 1

40% there

What you can now do

- Code with Frontier Models including AI Assistants with Tools
- Build solutions with open-source LLMs with HuggingFace transformers
- Compare LLMs to identify the right one for the task at hand

---

After next week you will be able to

- Navigate the most useful leaderboards to evaluate LLMs
- Give real-world use cases of LLMs solving commercial problems
- Confidently choose LLMs for your projects