Comparing LLMs – The Basics (1)

Feature	Description / Notes
Open-source or Closed	Whether the model's weights, architecture, and training code are publicly available. Examples: GPT-4 (closed), LLaMA 2 (open).
Release Date & Knowledge Cut-off	The date the model was released and the latest point in time the training data covers. Important for up-to-date responses.
Parameters	Number of trainable weights in the model (e.g., billions of parameters). Determines model capacity.
Training Tokens	Amount of text data used during training, usually in billions of tokens. More tokens generally improve performance.
Context Length	Maximum input length (in tokens) the model can handle in a single prompt. Longer context allows better understanding of large inputs.

Comparing LLMs – The Basics (2)

Feature	Description / Notes
Inference Cost	Computational resources required to generate output (GPU, CPU usage).
API Charge / Subscription / Runtime Compute	How much it costs to access the model through a cloud service.
Training Cost	Cost to pretrain the model, including compute, electricity, and infrastructure.
Build Cost	Cost to fine-tune, deploy, or customize the model.
Time to Market	How quickly you can deploy and use the model in production.
Rate Limits	Restrictions on API usage, e.g., calls per minute or per day.
Speed & Latency	How fast the model responds, depends on model size, hardware, and context length.
License	Terms for use, redistribution, and commercial deployment. Some open-source licenses restrict commercial use.

Chinchilla Scaling Law

Key Principle: Number of model parameters should scale roughly proportional to the number of training tokens.
Implications:

- Increasing model size without enough data leads to plateaued or degraded performance.
- Having lots of training data with a small model underperforms; parameters must scale up.

Rule of Thumb: Let N = model parameters, D = training tokens. For optimal training:

$$N \propto D$$

Example:

- Doubling model parameters $\rightarrow$ need roughly double the training tokens.
- Insufficient tokens $\rightarrow$ diminishing returns.

Additional Notes:

- Smaller, well-trained models can outperform larger, undertrained ones.
- Helps determine compute-efficient configurations for new LLMs.

Common LLM Benchmarks

Benchmark	**Focus / Task**	**Description**
MMLU	Knowledge across multiple subjects	Assess general knowledge and reasoning; used to compare models like GPT-3, Chinchilla, and Gopher.
BIG-bench	Broad suite of diverse reasoning tasks	Tests reasoning, factual knowledge, ethics, math, code; hundreds of tasks evaluating beyond narrow QA.
HellaSwag	Commonsense reasoning	Multiple-choice questions for everyday situations; measures ability to predict plausible outcomes.
TruthfulQA	Factual accuracy / truthfulness	QA tasks designed to detect hallucinations; evaluates honesty of LLM answers.
WinoGrande / Winograd Schema Challenge	Pronoun resolution / coreference	Tests commonsense reasoning and context understanding; resolves ambiguous references.
ARC	Science and reasoning	Multiple-choice science questions; evaluates problem-solving and reasoning in STEM.
HumanEval	Coding and code generation	Tests Python programming ability; measures functional correctness of generated code.

Specific Benchmarks

Benchmark	What's Being Evaluated	Description
ELO	Model ranking / performance consistency	Evaluates models via pairwise comparisons; creates a relative ranking of LLMs across multiple tasks.
HumanEval	Code generation / functional correctness	Tests an LLM's ability to write Python functions that pass unit tests; measures coding logic and correctness.
Multipl-E	Multimodal reasoning	Evaluates LLMs on tasks combining text and images (or multiple modalities); measures reasoning and comprehension across modalities.

Limitations of Benchmarks

Limitation	Explanation
Narrow focus	Many benchmarks test only specific skills (e.g., coding, factual QA, commonsense), not overall intelligence or adaptability.
Static datasets	Benchmarks are fixed in time, so models trained after the cut-off may have an unfair advantage or miss newer knowledge.
Lack of real-world context	Benchmarks often use idealized tasks, not messy, ambiguous, or multi-step real-world scenarios.
Gaming / overfitting	Models can be fine-tuned or prompted to specifically excel on benchmark tasks without improving general capabilities.
Limited multimodality	Most benchmarks focus on text-only tasks; few measure image, audio, or multimodal reasoning.
Subjectivity	Some benchmarks (e.g., ethics, creativity, hallucination detection) are hard to score objectively.
Compute bias	Larger models may perform better mainly due to size, not reasoning ability, skewing benchmark results.
Hard to measure nuanced reasoning	Benchmarks mostly measure surface correctness; they often cannot capture multi-step reasoning, context-dependent judgment, creativity, or reasoning accuracy vs. fluency.

Advanced Benchmarks for Large Language Models

Benchmark	Focus / Task	Description / Meaning
GPQA	Graduate-level question answering	Evaluates performance on graduate-level tests with 448 expert questions. Non-PhD humans score only 34% even with web access. Measures LLM ability to handle highly specialized knowledge.
BBHard	Future capabilities	Includes 204 tasks previously thought beyond LLM capabilities. Designed to test reasoning, logic, and generalization at a next-level difficulty.
Math Lv 5	High-school math competition problems	Measures model's ability to solve advanced math problems requiring multi-step reasoning and problem-solving skills. Useful for chain-of-thought evaluation.
IFEval	Instruction following	Tests the model's ability to follow complex instructions, e.g., "write more than 400 words" and "mention AI at least 3 times." Evaluates comprehension and compliance with nuanced prompts.
MuSR	Multistep soft reasoning	Assesses logical deduction and multi-step reasoning. Example: analyzing a 1,000-word story and identifying "who has means, motive, and opportunity." Tests reasoning beyond surface facts.
MMLU-PRO	Harder MMLU version	Advanced, cleaned-up version of MMLU with questions having 10 possible answers instead of 4. Evaluates deeper knowledge, multi-choice reasoning, and generalization.

Notes:

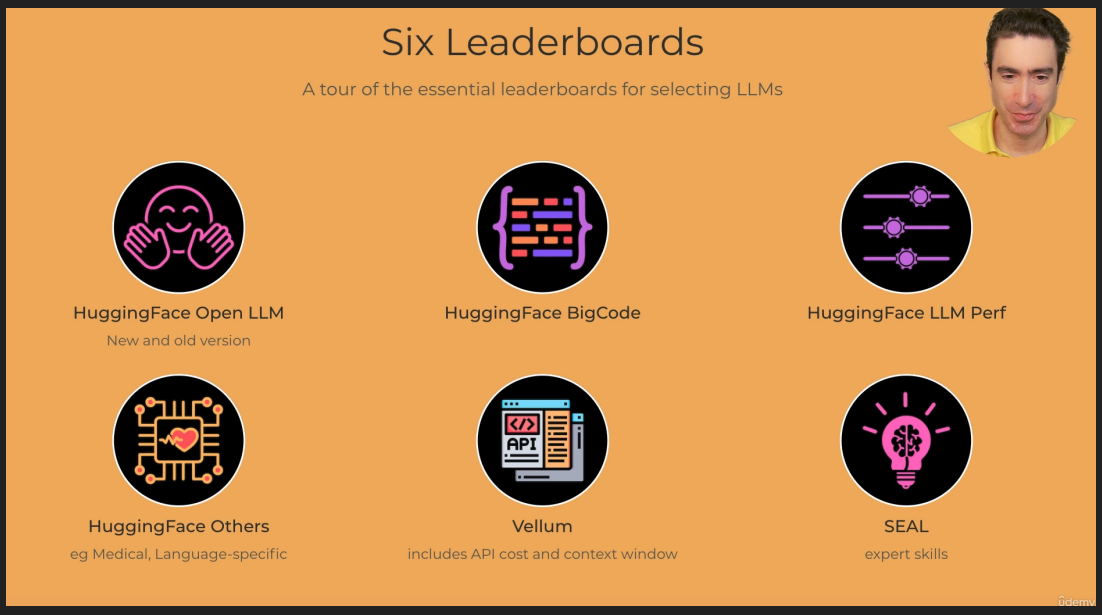
- These benchmarks are considered "next-level" because they test advanced reasoning, multi-step logic, and high-level knowledge.
- Useful for evaluating models that go beyond standard QA, commonsense reasoning, or basic coding tasks.
- Many of these benchmarks also measure compliance with complex instructions and reasoning under uncertainty.

Benchmark	Purpose / Examples
GPQA	Specialized knowledge, expert-level Q&A, academic reasoning, physics, chemistry, biology expertise. **Examples:** Explain the biochemical steps of glycolysis; Describe Newton's laws with practical applications; Explain photosynthesis in detail.
BBHard	Advanced reasoning, logic, generalization in challenging scenarios. **Examples:** Predict how a hypothetical AI agent could optimize a supply chain in a novel scenario; Design a strategy to minimize energy consumption in an unfamiliar industrial process; Solve a logic puzzle with multiple constraints.
Math Lv 5	Complex problem solving, chain-of-thought evaluation, mathematical reasoning. **Examples:** Solve: If $x^2 - 5x + 6 = 0$, find all integer solutions; Evaluate combinatorial problems like "How many ways can 5 books be arranged on a shelf?"; Solve calculus problems requiring multiple steps.
IFEval	Precise instruction compliance, comprehension, content generation with constraints. **Examples:** Write a 450-word essay on renewable energy mentioning AI at least 3 times; Summarize a research paper in 300 words including key terms; Rewrite a paragraph to follow a formal tone while keeping original meaning.
MuSR	Multi-step reasoning, deduction, understanding narratives beyond surface facts, solving prime puzzles and reasoning tasks. **Examples:** Analyze a 1,000-word mystery story to determine "who has means, motive, and opportunity"; Identify the next prime number in a complex sequence; Solve multi-step logic puzzles.
MMLU-PRO	Deep knowledge, advanced multi-choice understanding, broader knowledge evaluation, general language understanding. **Examples:** Choose the best explanation for why the sky is blue from 10 options; Identify the correct historical fact among 10 alternatives; Evaluate grammar and style in multiple-choice questions.

Tips to distinguish benchmarks:

- **GPQA** – focus on subject expertise.
- **BBHard** – focus on novel reasoning and logic.
- **IFEval** – focus on following instructions accurately.
- **MuSR** – focus on multi-step deduction and problem solving (e.g., prime puzzles, mysteries).
- **MMLU-PRO** – focus on broad knowledge and multiple-choice reasoning.

Leaderboard Image



Leaderboard	**Purpose / Use**
HuggingFace Open LLM	Tracks performance of open-source LLMs (new and old versions). Useful for selecting models for research or deployment with full weight access.
HuggingFace BigCode	Focused on code generation LLMs. Helps compare models for coding tasks, programming language coverage, and code quality.
HuggingFace LLM Perf	Benchmarking general LLM performance across various NLP tasks. Useful for measuring speed, accuracy, RAM memory usage and general reasoning ability.
HuggingFace Others	Specialized leaderboards, e.g., medical, domain-specific, or language-specific models. Guides selection for niche applications.
Vellum	Includes context window, API cost, and usage constraints. Useful for developers to choose models based on practical deployment factors.
SEAL	Focused on expert skills and high-level reasoning. Useful for selecting LLMs that excel in professional or complex knowledge domains.