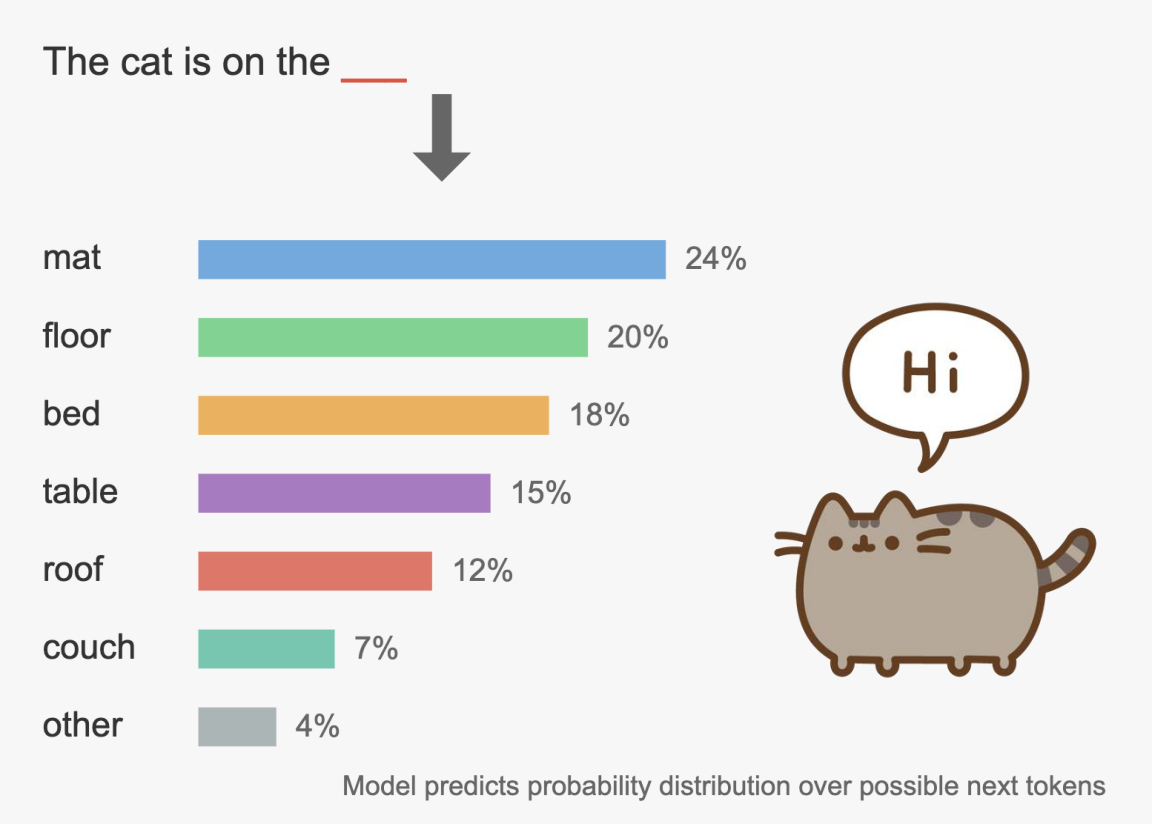
1. LLM basics

VIB AI Summer School 2025

Francesco Carli

EMBL-EBI

Next token predicton



Generated with Claude Opus 4

EMBL-EBI

Pre-training: a simple loss function

[START]	Target: "The"	loss_token_1
The	Target: "cat"	loss_token_2
The cat	Target: "is"	loss_token_3
The cat is	Target: "on"	loss_token_4
The cat is on	Target: "the"	loss_token_5
The cat is on the	Target: "mat"	loss_token_6

Total Loss

Σ = loss_token_1 + ... + loss_token_6

Average = Σ ÷ 6

Generated with Claude Opus 4

EMBL-EBI

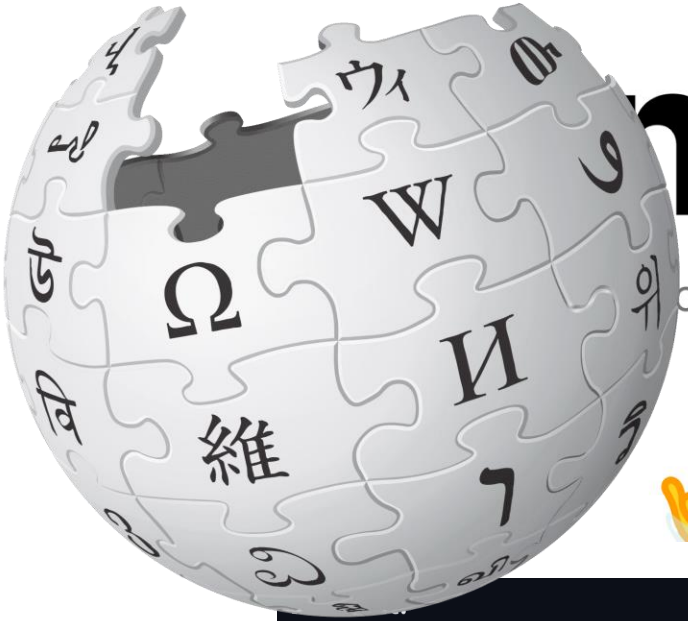
Scrape the whole internet and train on it

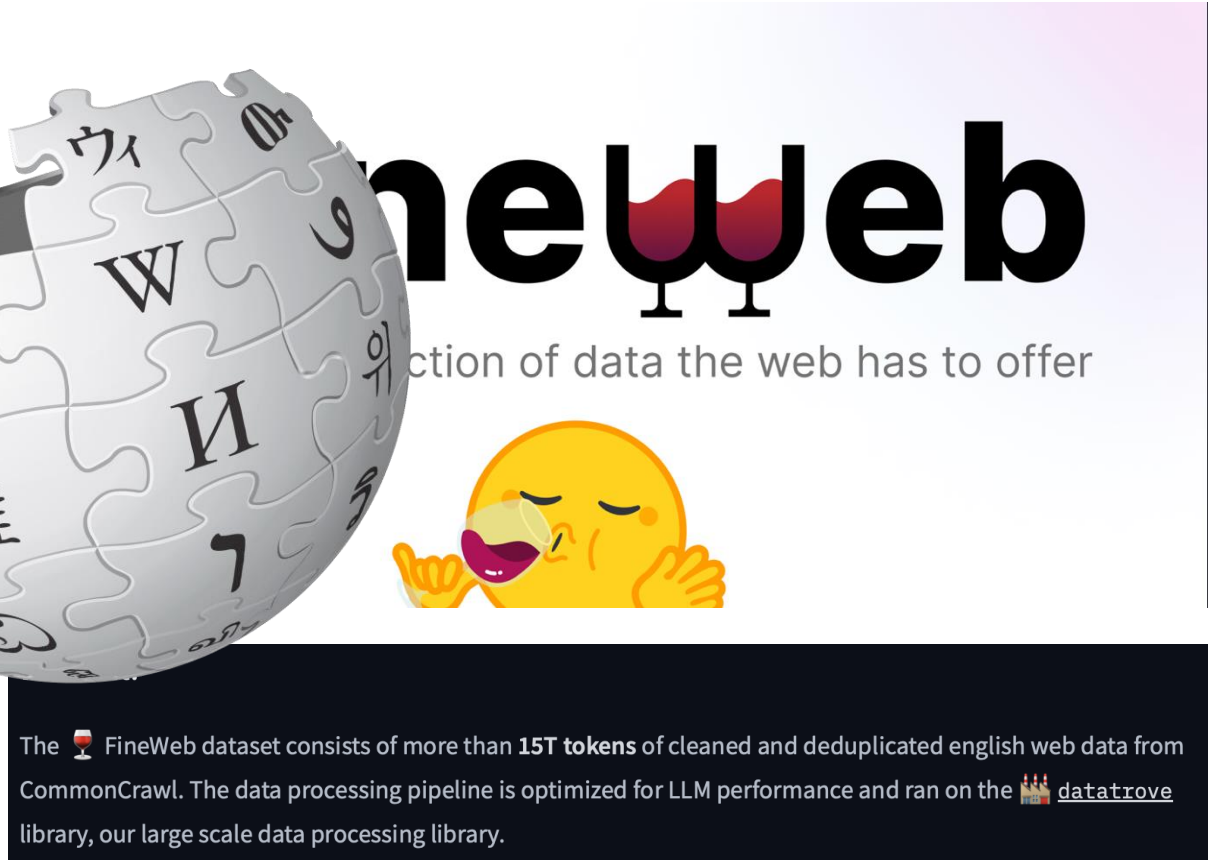
What is it?

The 🍷 FineWeb dataset consists of more than **15T tokens** of cleaned and deduplicated english web data from CommonCrawl. The data processing pipeline is optimized for LLM performance and ran on the 🏭 `datatrove` library, our large scale data processing library.

EMBL-EBI

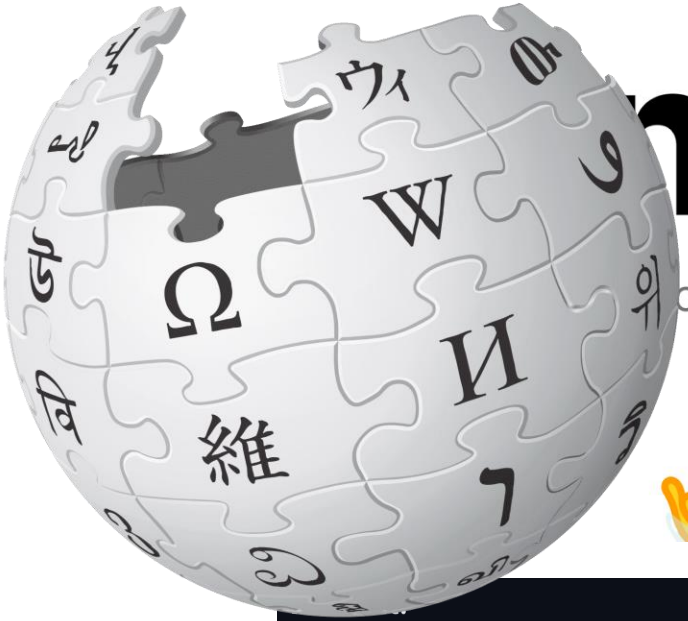
Scrape the whole internet and train on it

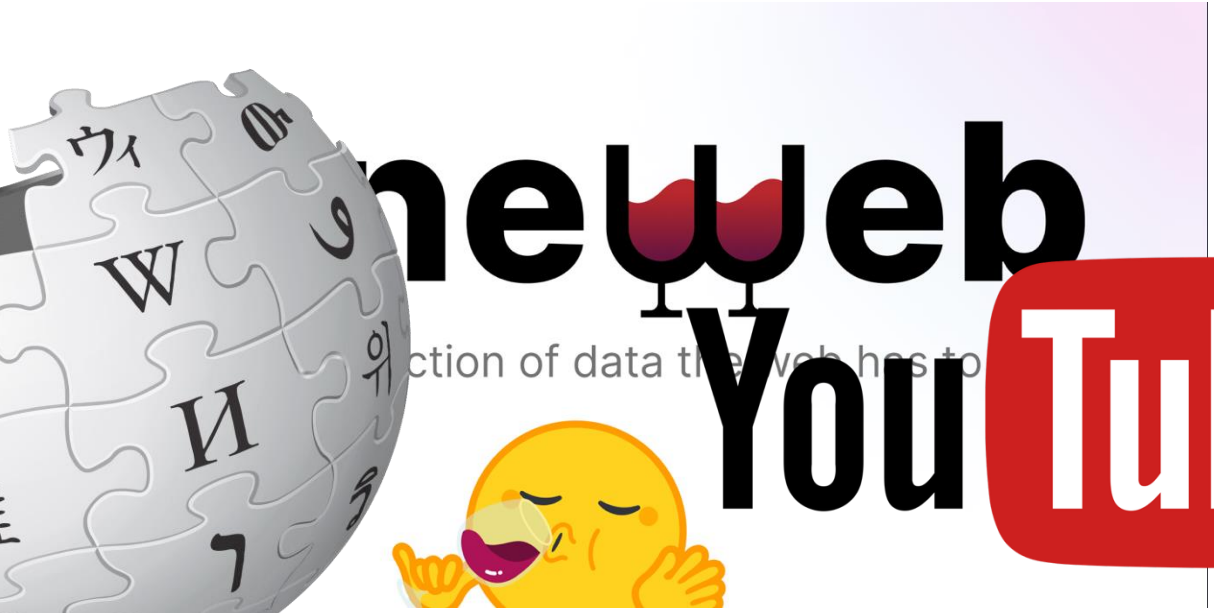




EMBL-EBI

Scrape the whole internet and train on it





The 🍷 FineWeb dataset consists of more than **15T tokens** of cleaned and deduplicated english web data from CommonCrawl. The data processing pipeline is optimized for LLM performance and ran on the 🏭 `datatrove` library, our large scale data processing library.

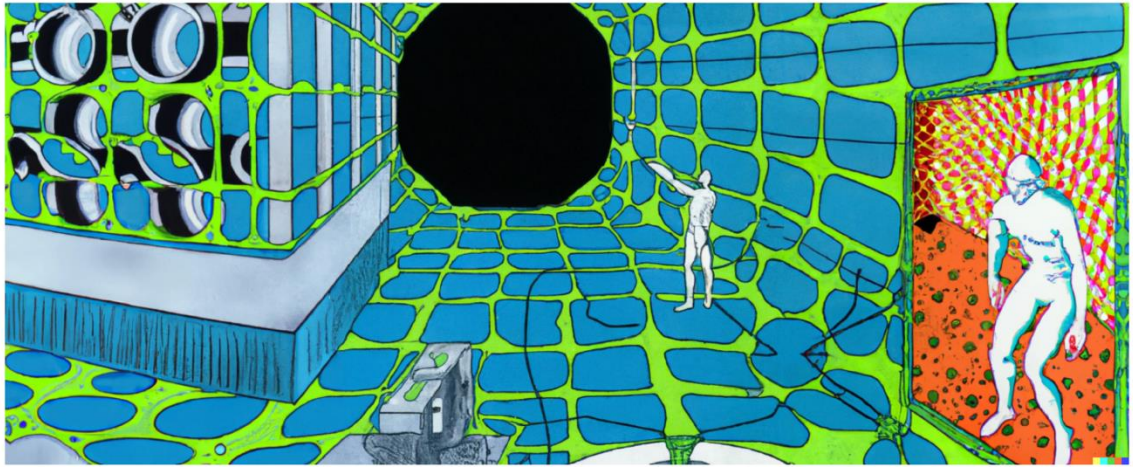
EMBL-EBI

Scrape the whole internet and train on it

The 🍷 FineWeb dataset consists of more than **15T tokens** of cleaned and deduplicated english web data from CommonCrawl. The data processing pipeline is optimized for LLM performance and ran on the 🏭 `datatrove` library, our large scale data processing library.

EMBL-EBI

Is “just a simulator”



"Moebius illustration of a simulacrum living in an AI-generated story discovering it is in a simulation" by DALL-E 2

$$P(new\ token | \{old\ tokens\})$$

https://www.lesswrong.com/posts/vJFdjigzmcXMhNTsx/

EMBL-EBI

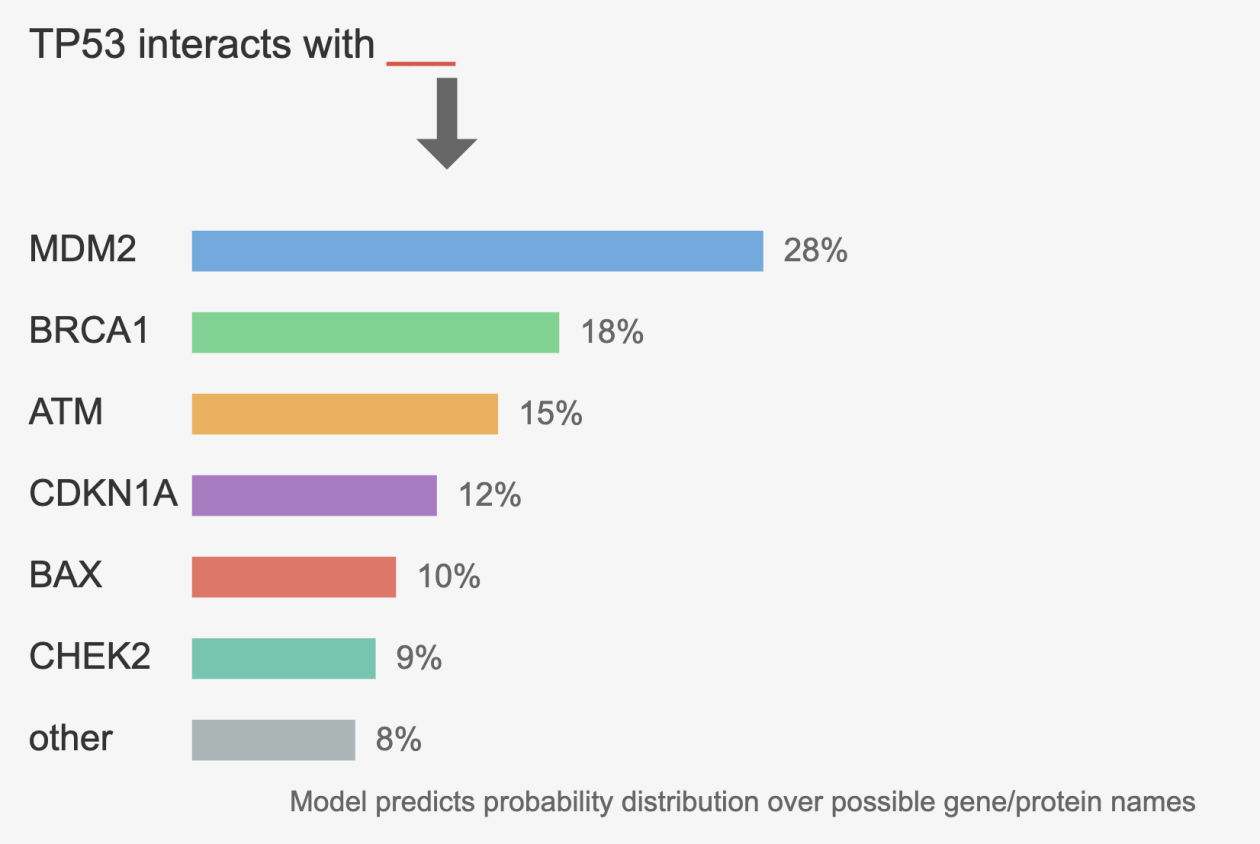
Understanding hallucinations

Understanding hallucinations

Sub-Type	User Input	Model Output	Explanation
Factual Inconsistency	Tell me about the first person to land on the Moon.	Yuri Gagarin was the first person to land on the Moon	The LLM's response is factual inconsistency as **Neil Armstrong was the first person to land on the Moon**, not Yuri Gagarin, who was the first human in space
Factual Fabrication	Tell me about the historical origins of unicorns	Unicorns were documented to have roamed the plains of Atlantis around 10,000 BC, where they were considered sacred creatures and were often associated with royalty.	The LLM's response is a fabricated claim since **there's no verified, real-world evidence supporting the existence of unicorns in Atlantis**, especially in connection with royalty

https://www.lakera.ai/blog/guide-to-hallucinations-in-large-language-models

Understanding hallucinations

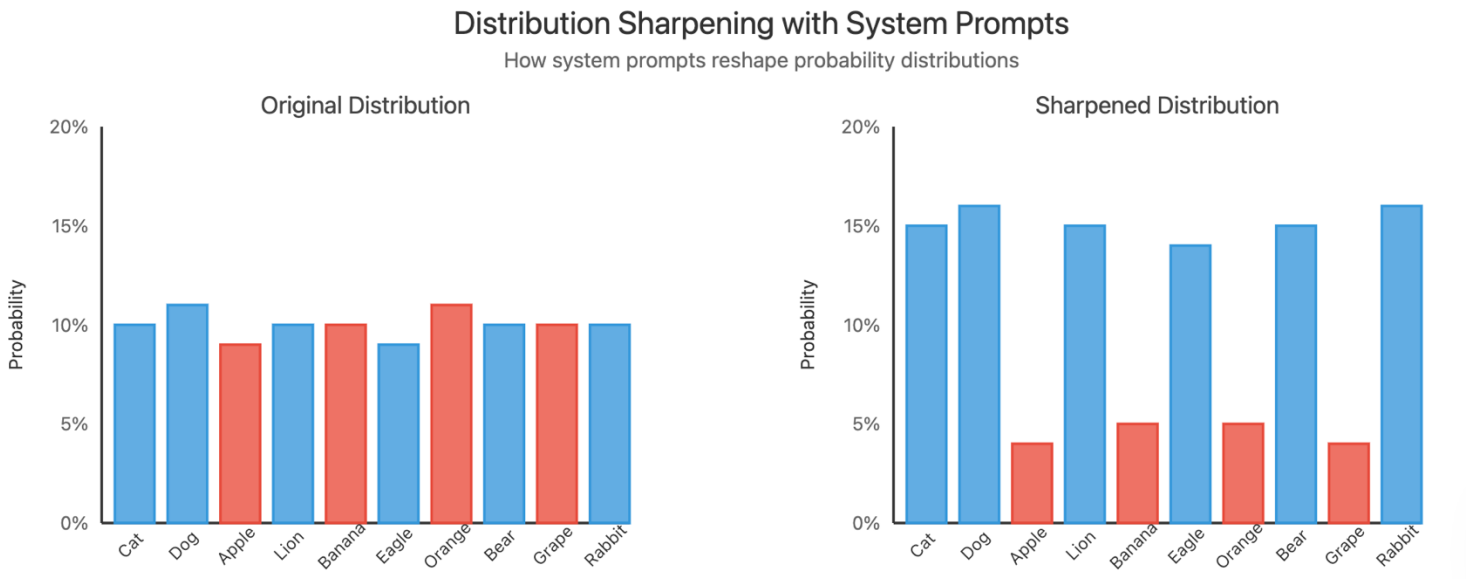


Generated with Claude Opus 4

EMBL-EBI

The role of the system prompt

'We are talking about animals"



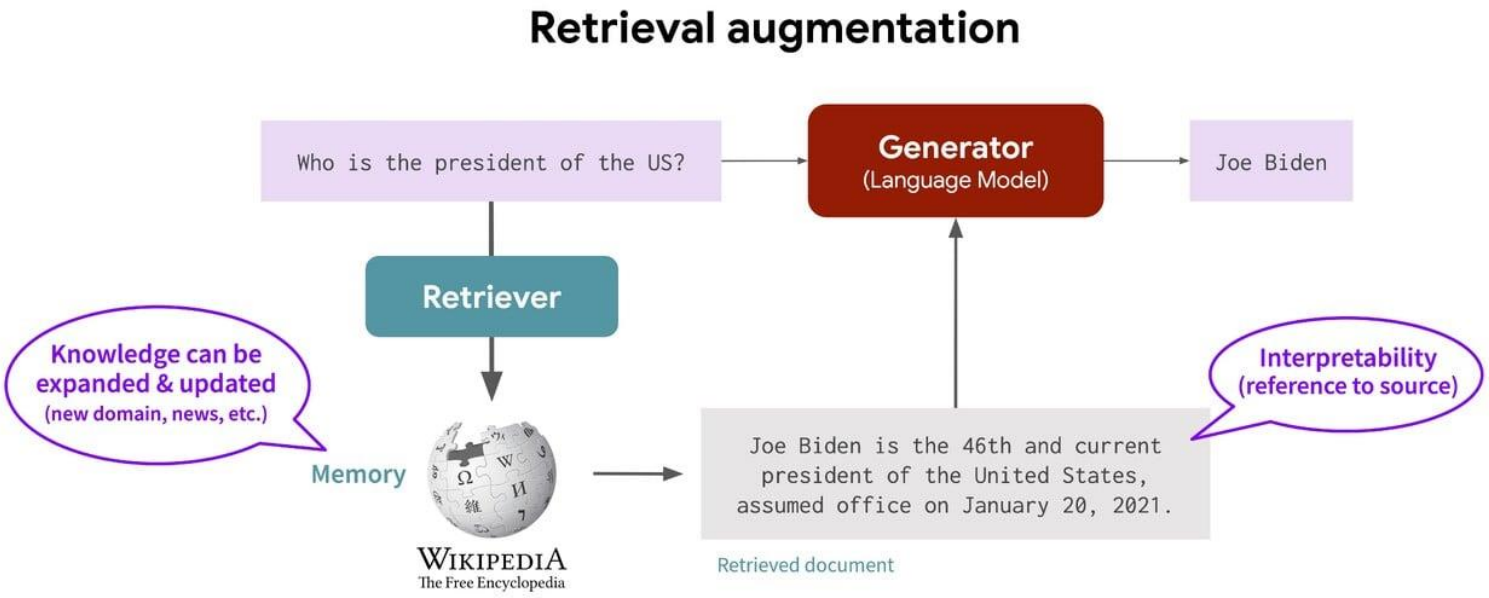
$$P(new\ token | \{old\ tokens\})$$

$$\{old\ tokens\} = \{sys\ \ prompt\} + \{already\ sampled\ tokens\}$$

Generated with Claude Opus 4

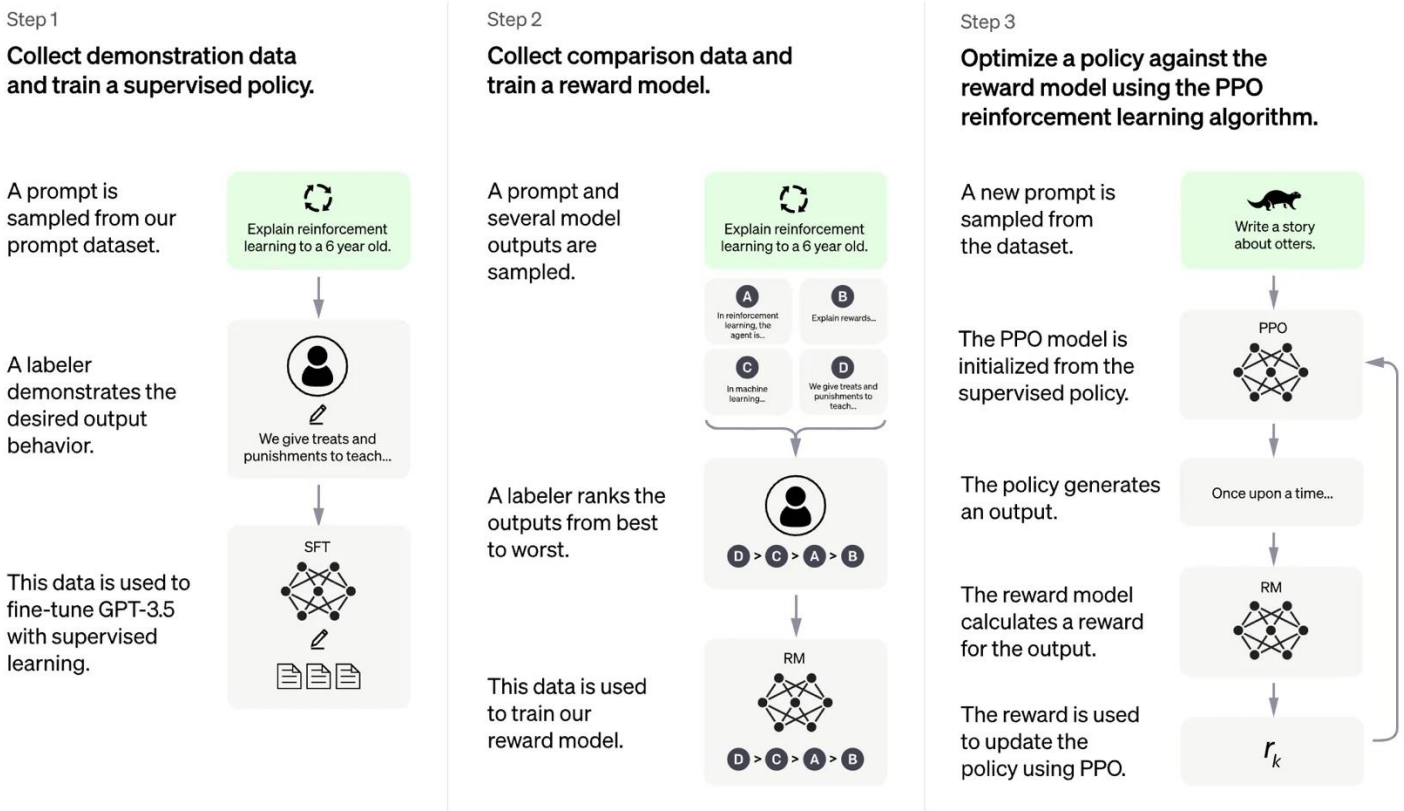
EMBL-EBI

Likewise for RAG



https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html#copying-literal-text

Instruction tuning: reinforcement learning from human feedback (RLHF)



Post training: reinforcement learning with verifiable rewards (RLVR)

deepseek

DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models

Zhihong Shao[1,2*†], Peiyi Wang[1,3*†], Qihao Zhu[1,3*†], Runxin Xu[1], Junxiao Song[1]
Xiao Bi[1], Haowei Zhang[1], Mingchuan Zhang[1], Y.K. Li[1], Y. Wu[1], Daya Guo[1*]

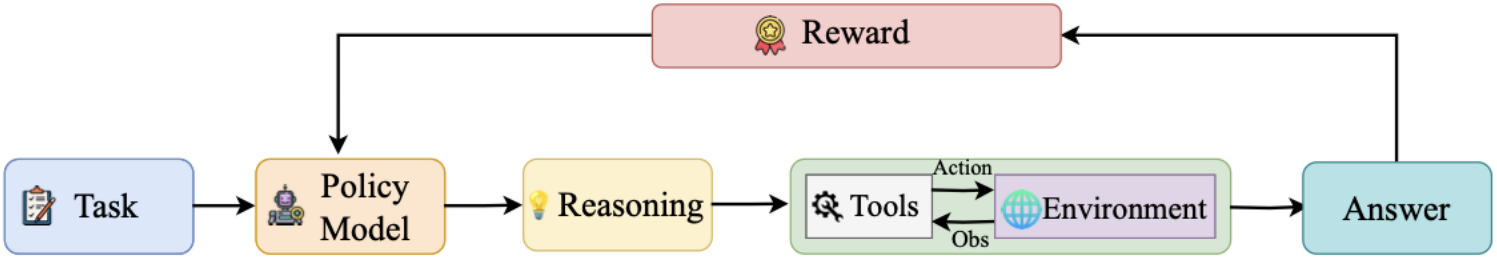
[1]DeepSeek-AI, [2]Tsinghua University, [3]Peking University

EMBL-EBI

Post training: reinforcement learning with verifiable rewards (RLVR)

Agentic Reasoning and Tool Integration for LLMs via Reinforcement Learning

Joykirat Singh, Raghav Magazine, Yash Pandya, Akshay Nambi
Microsoft Research
corresponding author: akshayn@microsoft.com

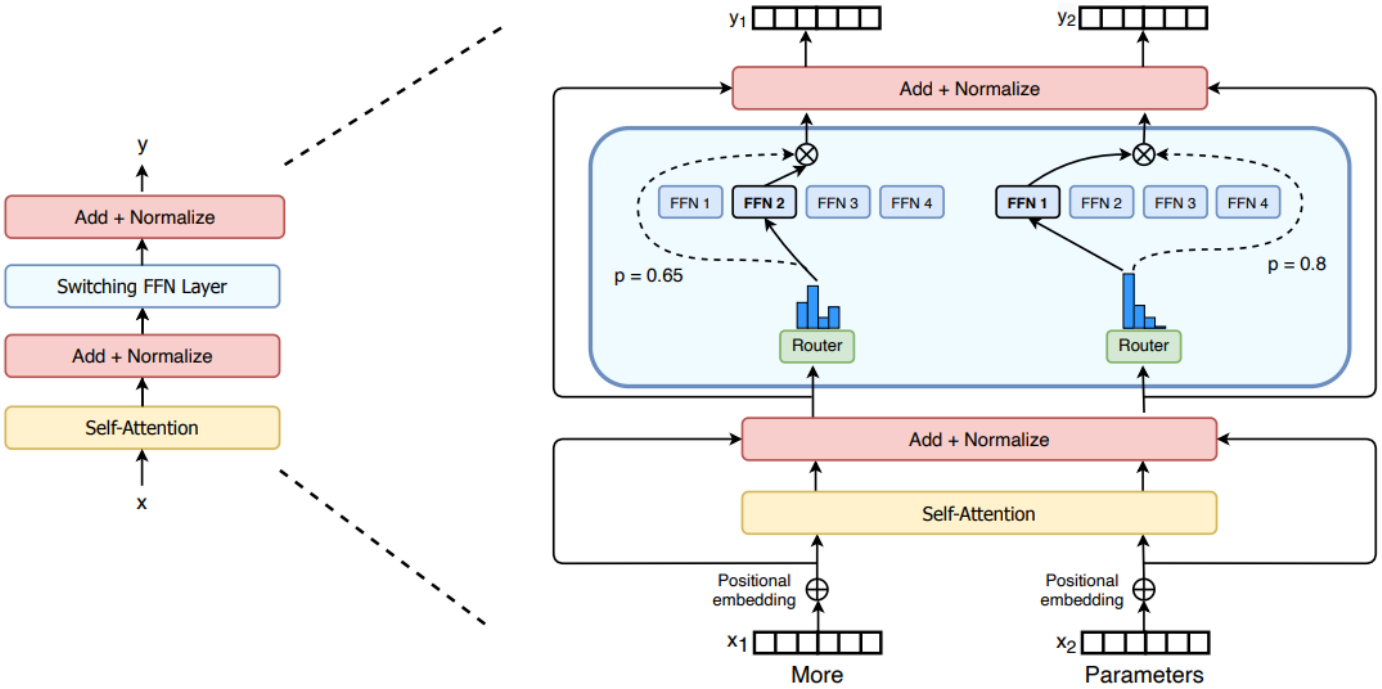


EMBL-EBI

Recap of the training steps

- **Pre-training:** raw next token prediction;
- **Instruction tuning (RLHF):** alignment with human preferences;
- **Post-training:** extension to specific thinking/agentic behaviors

Mixture of Experts

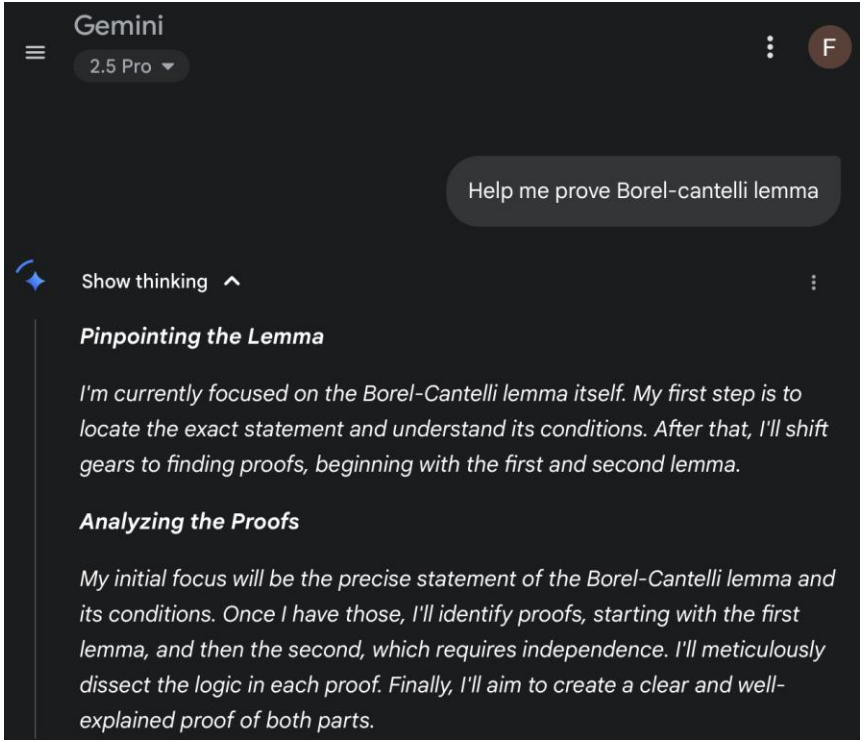


Key points:

- Have **faster inference** compared to a model with the same number of parameters;
- Require **high VRAM** as all experts are loaded in memory
- Face many **challenges in fine-tuning**

https://huggingface.co/blog/moe

Thinking models vs. standard models

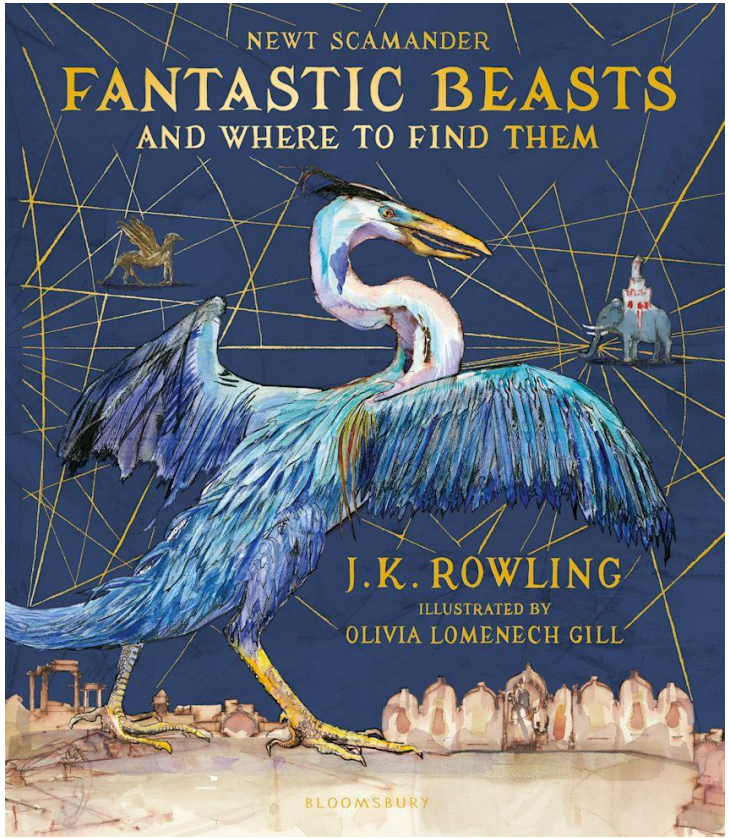


- Reasoning models produce **intermediate "chain-of-thought" traces**; regular models output only final answers.
- Thinking models are trained on or fine-tuned with step-by-step reasoning examples; regular models are trained on direct input–output pairs only
- Reasoning models perform better on **multi-step tasks , such as** logic, math, programming, and planning; regular models struggle or approximate.
- They often require **more compute and generate longer outputs**, due to reasoning steps; regular models are faster and leaner.

Nice resource: https://magazine.sebastianraschka.com/p/understanding-reasoning-llms

EMBL-EBI

Where to find models



EMBL-EBI

Two big families

“Open” models

- Llama3 (meta.ai)
- Qwen3 (alibaba)
- GLM 4.5 (z.ai)
- DeepSeek 3.1 (DeepSeek)
- Kimi-K2 (moonshot.ai)
- Mistral Small 3.1 (Mistral.AI)

Commercial models

- GPT5 (OpenAI)
- Gemini (Google DeepMind)
- Claude 4 (anthropicAI)
- Grok 4 (xAI)

Where to find open models

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Docs | Pricing

Qwen/Qwen3-235B-A22B-Thinking-2507 | like 326 | Follow Qwen 45.5k

Text Generation | Transformers | Safetensors | qwen3_moe | conversational | arxiv:5 papers | License: apache-2.0

Model card | Files xet | Community 7 | Train | Deploy | Use this model

Edit model card

Qwen3-235B-A22B-Thinking-2507

Qwen Chat

Highlights

Over the past three months, we have continued to scale the **thinking capability** of Qwen3-235B-A22B, improving both the **quality and depth** of reasoning. We are pleased to introduce **Qwen3-235B-A22B-Thinking-2507**, featuring the following key enhancements:

Downloads last month
37,840

Safetensors
Model size 235B params | Tensor type BF16
Chat template | Files info

Inference Providers NEW
Fireworks
Text Generation | Examples

https://huggingface.co

EMBL-EBI

Where to find open models

Models Turbo Search models Sign in Download

Embedding Vision Tools Thinking Popular

gpt-oss
OpenAI's open-weight models designed for powerful reasoning, agentic tasks, and versatile developer use cases.
tools thinking 20b 120b
1.3M Pulls 3 Tags Updated 1 week ago

deepseek-r1
DeepSeek-R1 is a family of open reasoning models with performance approaching that of leading models, such as O3 and Gemini 2.5 Pro.
tools thinking 1.5b 7b 8b 14b 32b 70b 671b
58.4M Pulls 35 Tags Updated 1 month ago

gemma3
The current, most capable model that runs on a single GPU.
vision 270m 1b 4b 12b 27b
12.8M Pulls 26 Tags Updated 1 week ago

Ollama

https://ollama.com

EMBL-EBI

Size calculator

Parameter Count

Enter the parameter count of the model in billions (1B to 1.8T)

e.g., 3, 7, 70, 1800

Vendor

Select your hardware manufacturer

NVIDIA

Quantization Method

Available quantization methods for your vendor

No Quantization (FP32)

Bit Precision

Available precision options

32-bit (FP32)

Hardware

Select GPU memory size

GPU Memory: 32 GB

8 GB 512 GB

Number of GPUs

Multiple GPUs can run larger models

1 GPU

RAM Usage

15%

COMPATIBLE

~4.80 GB

of 32.00 GB
(1 × 32 GB)

Memory Overhead 20%

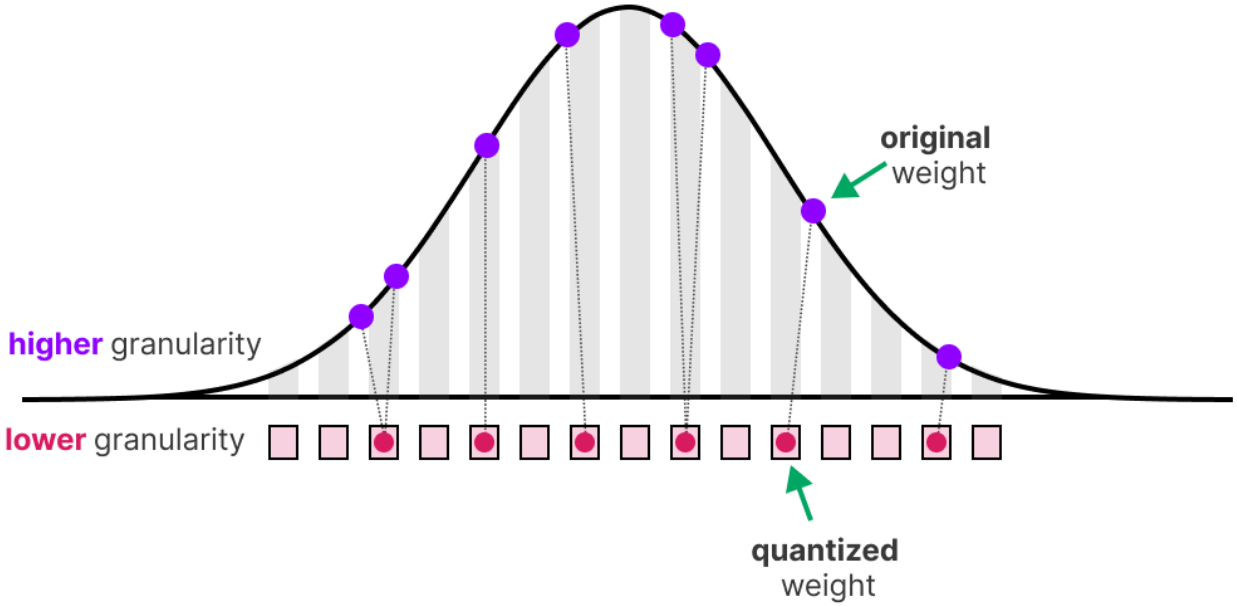
Additional memory required beyond model weights

Model Size	1B parameters
Quantization	32-bit (No Quantization)
Weights Memory	4.00 GB
Overhead	0.80 GB (20%)
Memory Per GPU	4.80 GB

https://research.aimultiple.com/self-hosted-llm/

EMBL-EBI

Quantization



https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-quantization

EMBL-EBI

Other libraries for open models

https://github.com/vllm-project/vllm

https://github.com/ggml-org/llama.cpp

https://github.com/sgl-project/sglang

Where to find closed models

OpenRouter

The Unified Interface For LLMs

Better prices, better uptime, no subscription.

Start a message...

Featured Models

View Trending

Model	Tokens/wk	Latency	Weekly growth
Gemini 2.5 Pro by google	136.1B	2.6s	-18.54%
GPT-5 Chat by openai	16.8B	761ms	-27.9%
Claude Sonnet 4 by anthropic	520.9B	1.8s	-10.96%

https://openrouter.ai

EMBL-EBI