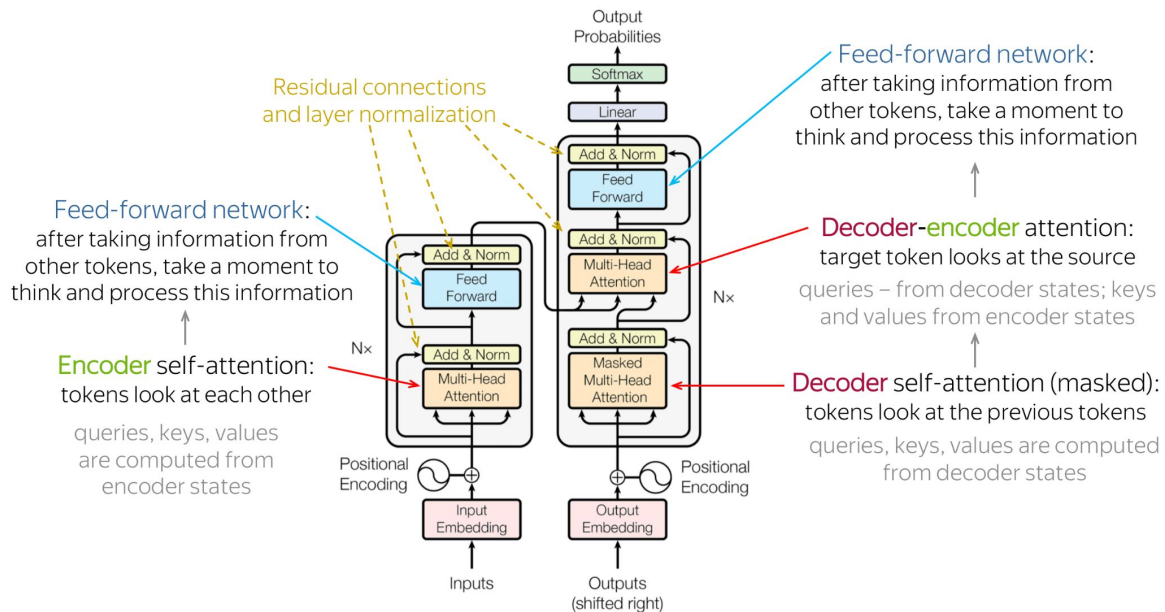# What is Meta LLaMA 2?

Discover the open-source LLM by Meta, for a viable alternative to OpenAI.

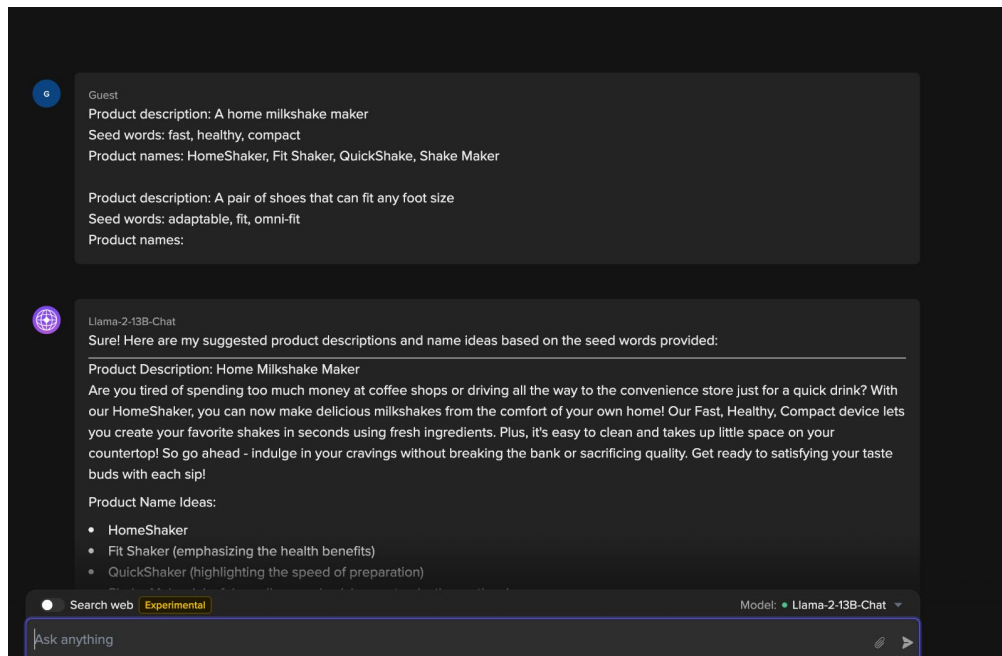# Transformer Models

— — —

Output
Probabilities

Softmax

Linear

Residual connections and layer normalization

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Add & Norm
Feed Forward

Feed-forward network: after taking information from other tokens, take a moment to think and process this information

Add & Norm
Feed Forward

Add & Norm
Multi-Head Attention

Decoder-encoder attention: target token looks at the source

queries – from decoder states; keys and values from encoder states

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states

N×

Add & Norm
Multi-Head Attention

N×

Add & Norm
Masked Multi-Head Attention

Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

Positional Encoding

Positional Encoding

Input Embedding

Output Embedding

Inputs

Outputs (shifted right)

Source

Meta LLaMA 2

- Created by Meta and open-sourced / leaked
- Access via hugging face or [lmsys](#)
- Use one of the fine-tuned models (i.e. Vicuna) for better results
- Try for free at https://chat.nbox.ai/

GPT-3.5 Level Performance

Leaderboard

| Blog | GitHub | Paper | Dataset | Twitter | Discord |

🏆 This leaderboard is based on the following three benchmarks.

○ Chatbot Arena - a crowdsourced, randomized battle platform. We use 50K+ user votes to compute Elo ratings.

○ MT-Bench - a set of challenging multi-turn questions. We use GPT-4 to grade the model responses.

○ MMLU (5-shot) - a test to measure a model's multitask accuracy on 57 tasks.

💻 Code: The Arena Elo ratings are computed by this notebook. The MT-bench scores (single-answer grading on a scale of 10) are computed by fastchat.llm_judge. The MMLU scores are computed by InstructEval and Chain-of-Thought Hub. Higher values are better for all benchmarks. Empty cells mean not available.

Model ▲	★ Arena Elo rating ▲	📝 MT-bench (score) ▲	MMLU ▲	License ▲
GPT-4	1206	8.99	86.4	Proprietary
Claude-1	1166	7.9	77	Proprietary
Claude-instant-1	1138	7.85	73.4	Proprietary
Claude-2	1135	8.06	78.5	Proprietary
GPT-3.5-turbo	1122	7.94	70	Proprietary
Vicuna-33B	1096	7.12	59.2	Non-commercial
Vicuna-13B	1051	6.57	55.8	Llama 2 Community
MPT-30B-chat	1046	6.39	50.4	CC-BY-NC-SA-4.0
WizardLM-13B-v1.1	1040	6.76	50	Non-commercial
Guanaco-33B	1038	6.53	57.6	Non-commercial
PaLM-Chat-Bison-001	1015	6.4		Proprietary
Vicuna-7B	1006	6.17	49.8	Llama 2 Community

Meta LLaMA 2 Features

Open-Source

As of yet the only real benefit is that this model is
open-source and therefore free (less the cost of compute).
You'd need to build a UX and any advanced functionality
around it, and may need to fine-tune it.

Meta LLaMA 2 Use Cases

———

Privacy / Data Protection

Many use-cases in the enterprise can't use OpenAI for fear of sensitive data leaking or being used to train the model (though OpenAI claims to keep API data private).

Fine-Tuning

If you have [200+ examples](#) fine-tuning beats prompt engineering for a specific defined task.