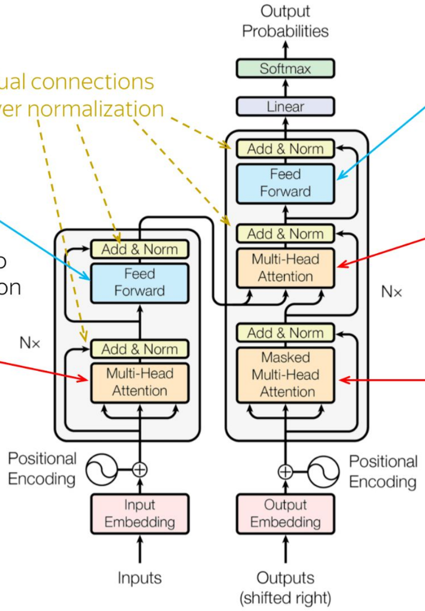
The Complete Prompt Engineering for AI Bootcamp (2023)

What is Mistral?

Discover the open-source LLM by Mistral, for a cheaper alternative to OpenAI.

Transformer Models



Residual connections and layer normalization

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

↑

Encoder self-attention: tokens look at each other

queries, keys, values are computed from encoder states

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

↑

Decoder-encoder attention: target token looks at the source

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

↑

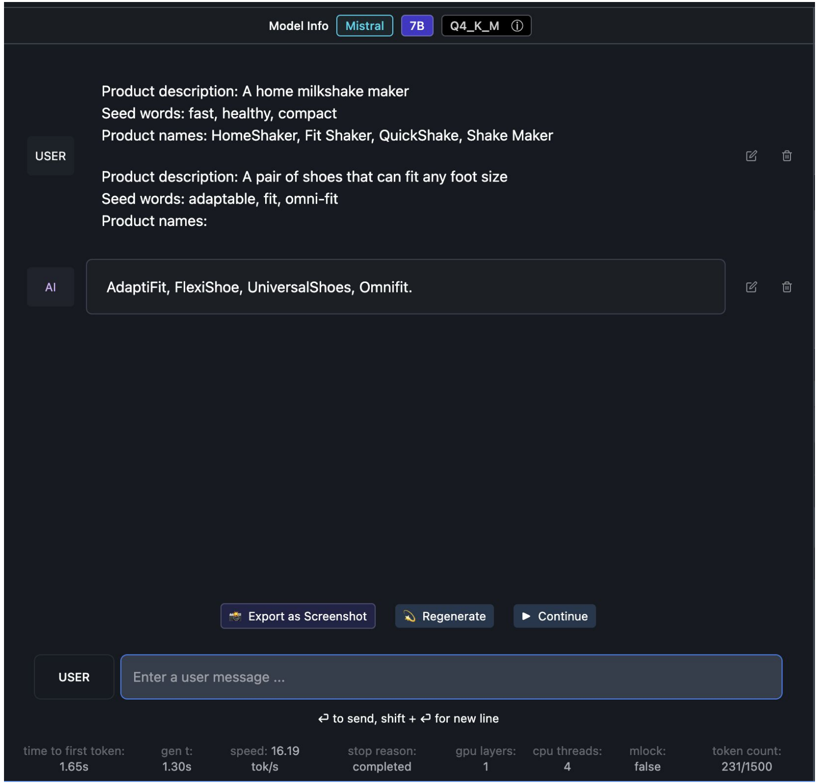
Decoder self-attention (masked): tokens look at the previous tokens

queries, keys, values are computed from decoder states

Source

Mistral 7B

- Created by french startup Mistral and open-sourced under the Apache license
- Access via hugging face or run locally
- Also offers a fine-tuned chat version



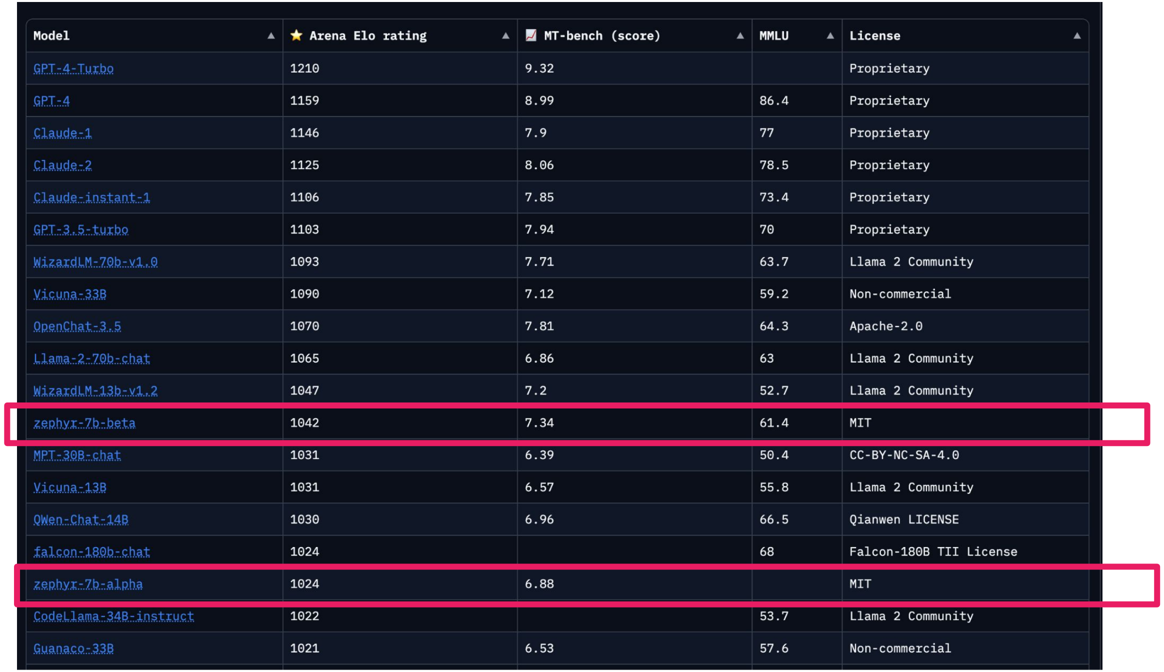
Product description: A home milkshake maker
Seed words: fast, healthy, compact
Product names: HomeShaker, Fit Shaker, QuickShake, Shake Maker

Product description: A pair of shoes that can fit any foot size
Seed words: adaptable, fit, omni-fit
Product names:

AdaptiFit, FlexiShoe, UniversalShoes, Omnifit.

```
time to first token:   gen t:   speed: 16.19   stop reason:   gpu layers:   cpu threads:   mlock:   token count:
1.65s                  1.30s    tok/s          completed      1             4              false    231/1500
```

GPT-3.5 Level Performance



Model	⭐ Arena Elo rating	📈 MT-bench (score)	MMLU	License
GPT-4-Turbo	1210	9.32		Proprietary
GPT-4	1159	8.99	86.4	Proprietary
Claude-1	1146	7.9	77	Proprietary
Claude-2	1125	8.06	78.5	Proprietary
Claude-instant-1	1106	7.85	73.4	Proprietary
GPT-3.5-turbo	1103	7.94	70	Proprietary
WizardLM-70b-v1.0	1093	7.71	63.7	Llama 2 Community
Vicuna-33B	1090	7.12	59.2	Non-commercial
OpenChat-3.5	1070	7.81	64.3	Apache-2.0
Llama-2-70b-chat	1065	6.86	63	Llama 2 Community
WizardLM-13b-v1.2	1047	7.2	52.7	Llama 2 Community
zephyr-7b-beta	1042	7.34	61.4	MIT
MPT-30B-chat	1031	6.39	50.4	CC-BY-NC-SA-4.0
Vicuna-13B	1031	6.57	55.8	Llama 2 Community
QWen-Chat-14B	1030	6.96	66.5	Qianwen LICENSE
falcon-180b-chat	1024		68	Falcon-180B TII License
zephyr-7b-alpha	1024	6.88		MIT
CodeLlama-34B-instruct	1022		53.7	Llama 2 Community
Guanaco-33B	1021	6.53	57.6	Non-commercial

Mistral 7B Features

Sliding Window Attention

Higher layers in the transformer stack can access information from further in the past than what is immediately visible in their attention window. As a result, the model can maintain a broader context over longer sequences, as well as offering linear compute cost which is a significant improvement over traditional attention mechanisms with quadratic compute complexity.

Mistral 7B Use Cases

Private Data

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

Efficiency

The sliding window mechanism increases speed and decreases cost, making it one of the cheaper models to run in production, particularly for larger tasks.