

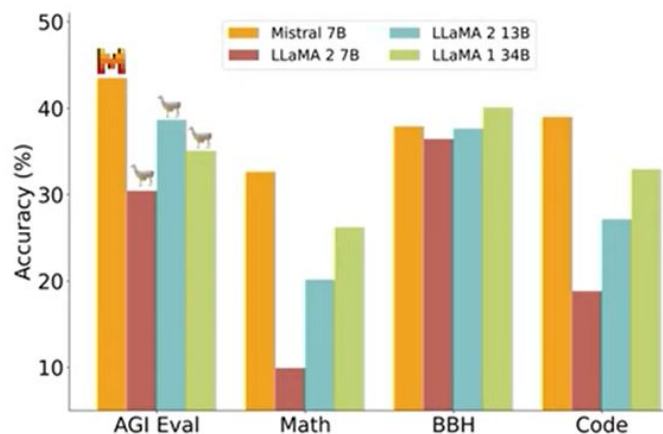
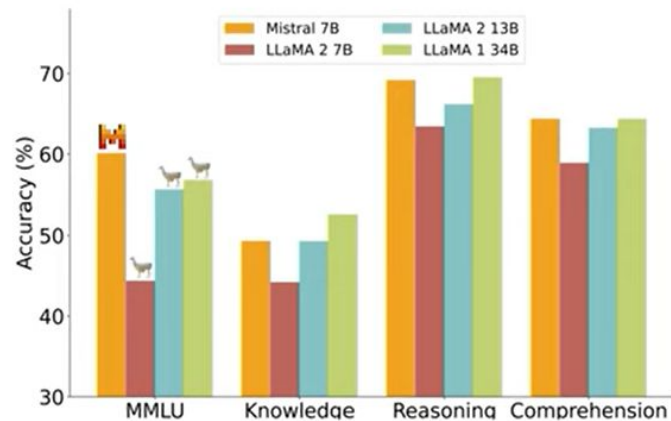
Mistral 7B

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford,
Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel,
Guillaume Lample, Lucile Saulnier, L lio Renard Lavaud, Marie-Anne Lachaux,
Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth e Lacroix,
William El Sayed

How does Mistral-7B perform?

- Mistral 7B
 - > best open 13B model (Llama 2)
 - > best released 34B model (Llama 1) in reasoning, mathematics, and code generation.
 - \equiv coding perf. of Code-Llama 7B
- Mistral 7B – Instruct
 - > Llama 2–chat 13B model

Model	Chatbot Arena ELO Rating	MT Bench
WizardLM 13B v1.2	1047	7.2
Mistral 7B Instruct	1031	6.84 +/- 0.07
Llama 2 13B Chat	1012	6.65
Vicuna 13B	1041	6.57
Llama 2 7B Chat	985	6.27
Vicuna 7B	997	6.17
Alpaca 13B	914	4.53



How does Mistral-7B perform?



Llama 2 13B Chat

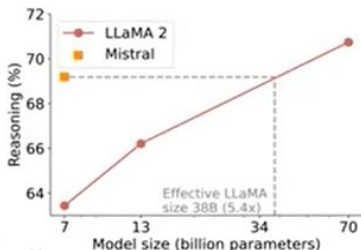
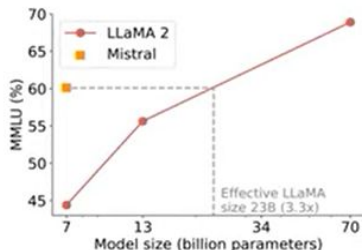
15571 🏆



Mistral 7B Instruct

18561 🏆

- <https://lmboxing.com/leaderboard>
- System prompt to enforce guardrails
 - Always assist with care, respect, and truth. Respond with utmost utility yet securely. Avoid harmful, unethical, prejudiced, or negative content. Ensure replies promote fairness and positivity.



Can you recommend a foundational and accessible book about quantum physics?

Copy Link to Question

Next Round (Enter) -->

Round Winner: mistral-7b-instruct-v0.1



mistral-7b-instruct-v0.1 🏆 Round Winner 🏆

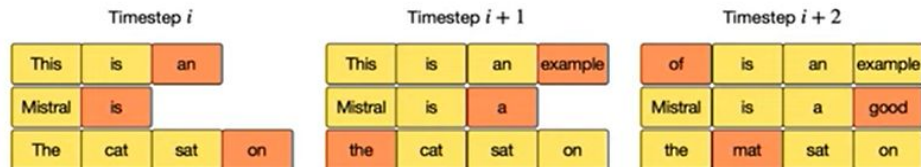
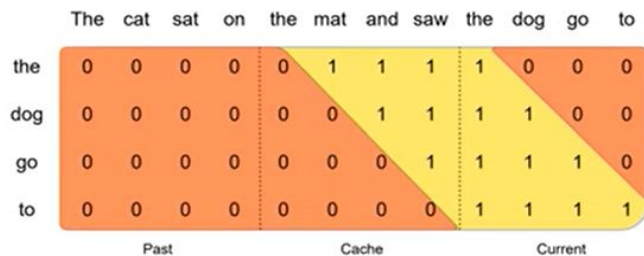
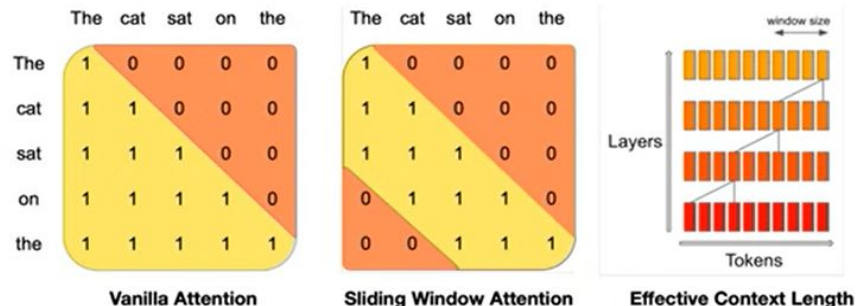
A great book for a foundational understanding of quantum physics is "The Quantum Universe: Everything that Can Happen Does Happen" by Brian Cox and Jeff Forshaw. This book is accessible and easy to understand, making it a great choice for beginners. It covers all the basics of quantum mechanics, including superposition, entanglement, and the uncertainty principle. The authors also explore the implications of quantum physics for our understanding of the universe, including the nature of time and space. Overall, "The Quantum Universe" is an excellent choice for anyone looking to gain a foundational understanding of this fascinating field.

llama-2-13b-chat

Certainly! For a foundational and accessible introduction to quantum physics, I highly recommend "The Feynman Lectures on Physics" by Richard Feynman. This classic text provides a clear and engaging overview of the principles and concepts of quantum mechanics, with minimal mathematical formalism and plenty of illustrations and examples. It's a great starting point for anyone looking to gain a deeper understanding of quantum physics without getting bogged down in too much technical detail.

What is the architecture for Mistral-7B?

- Transformer decoder: 32L, $|V|=32K$, dim=4096, #heads=32, context len=8192.
- Grouped-query attention (GQA): high inference speed, low RAM during decoding, higher batch sizes.
- Sliding Window Attention.
 - Each token can attend to at most W tokens from the previous layer.
 - Tokens outside the sliding window still influence next word prediction. At each attention layer, information can move forward by W tokens.
- Pre-fill and Chunking.
 - Pre-fill the (k, v) cache with the prompt.
 - Chunk long prompts into smaller pieces, and pre-fill the cache with each chunk.
 - With chunk size=window size, for each chunk, compute attention over the cache and over the chunk.
- Rolling Buffer Cache.
 - Cache has a fixed size of W
 - Keys and values for the timestep i are stored in position $i \bmod W$ of the cache.



Summary

- Mistral-7B
 - Grouped-query attention
 - Sliding Window Attention
 - Pre-fill and Chunking
 - Rolling Buffer Cache
- Perf
 - > Llama 2-13B
 - > Llama 1-34B.
 - \equiv coding perf. of Code-Llama 7B
 - Mistral 7B – Instruct > Llama 2–chat 13B model

Thank you