* Transformer model vs. MoE (Mixture-of-Experts Layer) model?
* GPT model behind? Is it Transformer? Predicting the next word related?
* Decoder mode vs. others in Transformer model?
* Decoder-only architecture like GPT models?
* Gemini uses Transformer & MoE model researches
* Multi-headed attention vs. self-attention in Transformer?
* Does GPT use RNN (recurrent neural network)?
* GPT: “we fine-tune the model’s behavior using reinforcement learning with human feedback ([RLHF⁠](https://openai.com/index/learning-from-human-preferences/)).”
* Does Gemini have one?
* Convolutional vs. recurrent neural networks
* If all AI are developed based on Transformer model, then how some is better than others?
* <https://aws.amazon.com/what-is/transformers-in-artificial-intelligence/>
* <https://openai.com/index/gpt-4-research/>
* [MoE Research](https://arxiv.org/abs/1701.06538)
* [Transformer Research](https://research.google/blog/transformer-a-novel-neural-network-architecture-for-language-understanding/)
* <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#architecture>
* <https://www.larksuite.com/en_us/topics/ai-glossary/problem-that-ai-is-trying-to-solve>
* [Llama 3 Herd of Models](https://www.youtube.com/watch?v=uXt6rYXnV8U)
* <https://www.cloudflare.com/learning/ai/what-is-large-language-model/>
* <https://aws.amazon.com/what-is/large-language-model/> “ The underlying transformer is a set of [neural networks](https://aws.amazon.com/what-is/neural-network/) that consist of **an encoder and a decoder** with self-attention capabilities.”
* <https://medium.com/data-science-at-microsoft/how-large-language-models-work-91c362f5b78f>