Yogesh Haribhau Kulkarni • You

AI Advisor (Helping people/organizations in their AI j...

1m •

Mixtral by Mistral AI, unlike traditional Large Language Models like GPT from OpenAI and Gemini from Google, takes a different path, leveraging an ensemble of specialized neural networks called "experts.", ie Mixture of Experts (MoE) approach.

At inference time, Mixtral dynamically selects and combines the outputs of just a couple of experts (out of 8, as shown in the pic below 😂), delivering the performance of large models with an inference cost comparable to smaller models. This radical advantage is just the tip of the iceberg! 🧊

Here is my blog post that explores the core concepts behind Mixtral, including sparse MoE layers, gate networks for token routing, and the potential for hierarchical MoE architectures.

https://lnkd.in/gY5QGc5r

It also dives into implementation details, model selection, function calling for external tool integration, and building a Retrieval-Augmented Generation (RAG) system using Mixtral's embedding model.

The blog is based on the recently launched course "Getting Started with Mistral" by DeepLearning.AI, created by Sophia Yang, Ph.D. and Andrew Ng.

https://lnkd.in/gtBBpdTZ

This fantastic resource covers:
✅ Prompting Mistral models via API calls
✅ Native function calling
✅ Building a basic RAG system
✅ Creating a chat interface on your documents

Highly recommended! 🚀

Let me know if you have used Mistral models and your thoughts on the same, in the comments below.

Mixtral by Mistral

medium.com • 3 min read