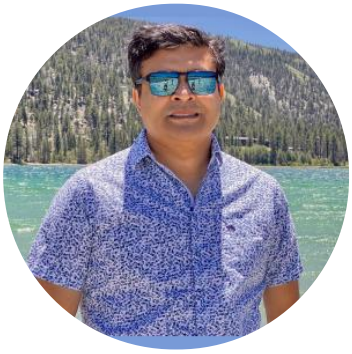
SHOW NOTES

@mjkabir Notes



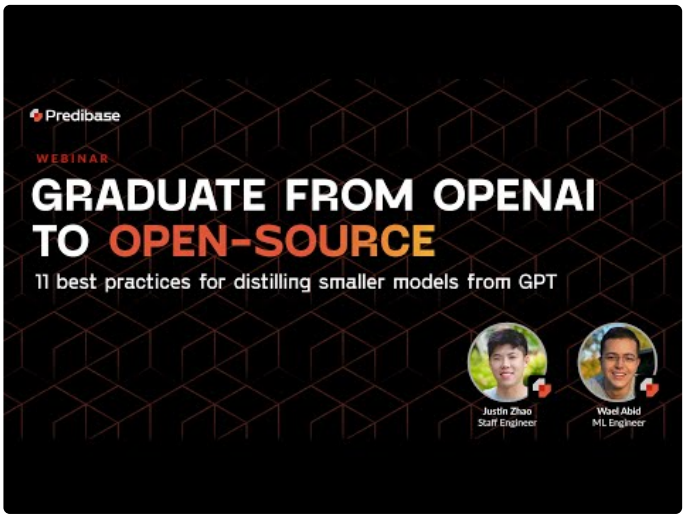
https://shownotes.app/show/mfnv3

Migrating From Closed Source to Open Source LLM

My notes on tips and tricks on switching to OSS LLM.

AI REVIEW PASSED.

Bengali Spanish Japanese Chinese



12 Best Practices for Distilling Smaller LLMs with GPT

As large language models (LLMs) become increasingly integral to business applications, the need for smaller, more efficient models has never been greater, according to the speaker. This shift is driven by the compelling performance of open-source LLMs, juxtaposed with the high costs, resource demands, and slower speeds of larger commercial models like GPT-4.

In this session, the speaker states that the audience will learn how to graduate from OpenAI to open-source with model distillation. Drawing on the speaker's experience distilling language models at Google and Predibase, combined with a review of recent LLM research, the speaker presents 12 best practices to help the audience get started distilling large models into smaller, more cost-effective LLMs.The speaker uses the Jigsaw comment toxicity dataset and provides code snippets while walking through each of the 12 distillation best practices:

- Understand the limitations of smaller models;
- Build good logging infrastructure;
- Define clear evaluation criteria;
- Maximize the quality of your teacher model;
- Use auxiliary techniques to maximize data quality offline;
- Account for data diversity, representation, and balance;
- Consider how you want to serve your fine-tuned models;
- Start small, even smaller;
- Assess the marginal utility of having more data;
- Actually look at the model's mistakes;
- Experiment graciously; and Deploy and monitor your models in production.

Whether the audience consists of LLM experts, AI enthusiasts or developers looking to learn more, the speaker states that these strategies are both practically viable and grounded in academic theory, to help get started with distillation.
228 days ago

Website:
https://www.youtube.com/watch?v=vELvSKZOREA