Generative AI

To Train, Fine-Tune, Prompt Engineer or Not is the Question!

This session discusses the three main approaches to using large language models (LLMs): training, fine-tuning, and prompt engineering. Training is the most time-consuming and expensive approach, but it can lead to the best performance. Fine-tuning is less time-consuming and expensive, but it can still lead to significant improvements in performance. Prompt engineering is the least time-consuming and expensive approach, but it can be less reliable.

Key Takeaways:

- Training:
  - Training LLMs requires a large dataset of text and code.
  - Training can take weeks or months, depending on the size of the dataset and the model architecture.
  - Training can be expensive, depending on the hardware used.
  - Training can lead to the best performance, but it is not always necessary.
- Fine-tuning:
  - Fine-tuning LLMs requires a small dataset of labeled data.
  - Fine-tuning can take hours or days, depending on the size of the dataset and the model architecture.
  - Fine-tuning is less expensive than training, but it can still be costly.
  - Fine-tuning can lead to significant improvements in performance, but it is not always necessary.
- Prompt Engineering:
  - Prompt engineering does not require any training data.
  - Prompt engineering can be done quickly and easily.
  - Prompt engineering is the least expensive approach to using LLMs.
  - Prompt engineering can be less reliable than training or fine-tuning.

This session concludes by discussing the trade-offs between the three approaches to using LLMs. The best approach depends on the specific task at hand and the resources available.

Sandeep Singh

Head of Applied AI/Computer Vision

BEANS .AI