Incorporating Advances in LLMs into the Machine Learning Lifecycle

Round table leader & Moderator: Navneet Rao, Senior Engineering Manager @ Thumbtack

Round Table Session Abstract

New advances in large language models (LLMs) have the potential to be one of the most disruptive technological advancements in recent history. The pace at which new innovations are being created & shared in the machine learning community far outstrips the pace of other innovative advancements in technology. There is an urgent need amongst leaders in the machine learning community to start building best practices around the creation & use of large language models as there is no precedent for the pace of innovation currently being witnessed.

Having spoken to ML leaders across organizations of various sizes, in this roundtable, we will discuss some of the pressing topics on the minds of machine learning leaders & practitioners who are trying to incorporate advances in LLMs into the machine learning lifecycle.

Goal: Gather insights and share learnings from the successes and failures of any recent experimentation / strategies.

Topics

1. How do you trade-off short term product commitments while trying to experiment with disruptive advancements in LLM technologies?
Discussion starting points:

a. How much effort are you dedicating to incorporating LLMs into product use cases?
b. Are you actively building new Proof of Concepts (PoCs)?
c. How are you managing expectations with your leadership?

2. How are you handling the hallucination problem?
Discussion starting points:

a. Using LLMs for internal-only use cases
b. Incorporating LLMs into product use cases

3. How are you internally sharing best practices to avoid duplication of efforts?
Discussion starting points:

a. For prompt creation
b. For fine-tuning model development

4. How are you dealing with licensing & legal concerns over using LLMs trained by various organizations?

Scenarios:

a. When they have commercially usable licenses (Apache V2 / MIT)
b. When they have commercially usable licenses with additional clauses around responsible AI (https://bigscience.huggingface.co/blog/the-bigscience-rail-license)
c. When they have commercially usable licenses with additional clauses around not using their materials or output or results of materials to improve any other large language model (https://ai.meta.com/resources/models-and-libraries/llama-downloads/)
d. When they have commercially usable licenses but you're unsure of the data lineage, do you train your own models

5. How are you thinking about the trade-off between fine tuning open source LLMs internally versus using APIs from companies like OpenAI?

Discussion starting points:

a. Have you successfully fine tuned open source LLMs for product / research?
b. Have you successfully used APIs from companies like OpenAI for product / research?
c. What challenges did you face when trying to do one vs the other?

Presenter Bio

Navneet Rao is a senior EM at Thumbtack, leading cross functional teams of software engineers (including ML eng) focusing on search experience (semantic search, search relevance), search ranking (personalization using Deep Cross Networks) & monetization (auction dynamics, supply efficiency). Recently, he's also been leading efforts around the use of generative AI at Thumbtack. Before that, he drove ML experimentation & productization on the algorithms team, for IBM's flagship, conversational AI product: Watson Assistant. He also holds several patents in this space. He is an alumni of the Language Technologies Institute (LTI), at the School of Computer Science at Carnegie Mellon University.

Navneet Rao
Senior Engineering Manager at Thumbtack

About Thumbtack

Thumbtack (www.thumbtack.com) is a local services marketplace where customers find and hire skilled professionals. Our app intelligently matches customers to electricians, landscapers, photographers and more with the right expertise, availability, and pricing. Headquartered in San Francisco, Thumbtack has raised more than 400 million USD from Baillie Gifford, Capital G, Javelin Venture Partners, Sequoia Capital, and Tiger Global Management among others. Available in all 3,143 U.S. counties, more than 4M customers have used Thumbtack in the last 12 months to find and hire professionals and local businesses.

Round Table Summary

Participants

Engineers, research scientists, PMs, EMs, data scientists from tech companies like Amazon, Google, Etsy, Workday, Thumbtack among others.

Discussion Summaries

1. How do you trade-off short term product commitments while trying to experiment with disruptive advancements in LLM technologies?

Discussion starting points

1. How much effort are you dedicating to incorporating LLMs into product use cases?
2. Are you actively building new Proof of Concepts (PoCs)?
3. How are you managing expectations with your leadership?

Discussion Insights

- Creating a XFN task force that spans across applied science, engineering, leadership, security, legal & eng can be helpful in working through challenges around prototyping, licensing & productization of LLMs
- When carving out a portion of your team's time to experiment with LLMs align efforts around the larger product imperatives
- Have applied science work alongside engineering so as to more quickly move forward on LLM experimentation
- If there is risk to your company's business model, start thinking through that and then consider introducing product differentiation or improvements around productivity

2. How are you handling the hallucination problem?

Discussion starting points

1. Using LLMs for internal-only use cases
2. Incorporating LLMs into product use cases

Discussion Insights

- Clarify what hallucination is, it's generally when LLMs make statements that sound plausible but are not. They are statements that do not have a basis in the training data that it's learned from.
- Consider constraining the space of output possibilities, so that this is less of a problem e.g. Ask the LLM to choose from option A, B or C

- Hallucination can be very domain specific, it's not always a problem since you might want more out of the box thinking e.g. applications in medicine
- Explicitly ground knowledge of the output space by using Retrieval Augmented Generation (RAG) so as to reduce hallucination
- Consider exploring chain of thought reasoning so you're making the LLM explain itself, so you can detect sources of bias

3. How are you internally sharing best practices to avoid duplication of efforts?

Discussion starting points

1. For prompt creation
2. For fine-tuning model development

Discussion Insights

- At this time point, given that we are still in the early days of experimenting with LLMS, maybe duplication of efforts might be the very thing needed to drive innovation
- Since open source research is rapidly changing the space & new innovations are coming in at quite a pace, do not over optimize your work in any one direction, since newer generation models might address some of the discrepancies you are currently noticing
- Introduce general software engineering best practices as early as possible, like creating libraries & creating a more plug & play framework of iteration using LLMs

Overall Summary

We are still in the early days around LLM adoption, but it's clear that the disruptions in technology introduced via new advances in generative AI are percolating across many areas of technology.