Navigating Challenges and Technical Debt in LLMs Deployment

Ahmed Menshawy, VP of AI Engineering, Mastercard
Zeeshan Nawaz, Principal AI Engineer
Mahmoud Fahmy, Senior AI Engineer
Pasquale Minervini, Researcher at the University of Edinburgh, School of Informatics
Essentials for building a generative AI application

- Access to a variety of foundation models
- Environment to Customize Contextual LLMS
- Easy-to-use tools to build and deploy applications
- Scalable ML infrastructure
Essentials for building a generative AI application

- Access to a variety of foundation models
- Environment to Customize Contextual LLMS
- Easy-to-use tools to build and deploy applications
- Scalable ML infrastructure

A lot of challenges and technical debt
Not so much
Vast and complex surrounding challenges and technical debt around ML deployment*

Challenges with Closed-book (Non-parametric) Approach

- Problems:
  - Hallucination
  - Attribution
  - Staleness
  - Revisions
  - Customization

- Solution:
  - Couple to external memory
Contextualization: Open-Book (Non-parametric approach)

- Grounding: improve factual recall*
- Swap in and out vector indices and make corrections as necessary

Contextualization

Credit: Douwe Kiela - Retrieval Augmented Language Models
Vast and complex Surrounding challenges and technical debt around LLM deployment*

“After reading, I began to wonder if LLMs are the correct way forward to achieve the tasks we are currently trying to solve using them. It seems we have to tackle large swaths of problems before we can maximize LLMs' efficiency.”

Anonymous Reviewer