Bridging the Gap: RAG, Local Models, and the Future of Enterprise AI

Bard

September 2, 2024

Introduction

- LLMs are transforming how we interact with information and automate tasks.
- This presentation explores three key aspects of this transformative landscape:
  - Retrieval Augmented Generation (RAG)
  - The rise of local models
  - GenAI at scale in a financial institution (Citibank)

Retrieval Augmented Generation (RAG)

- RAG overcomes limitations of LLMs by integrating external knowledge bases.
- Four key stages:
  - Indexing: Transforming unstructured data into a searchable representation (embeddings, vector databases).
  - Storage: Efficiently storing data for rapid retrieval.
  - Querying: LLM translates user queries into vector representations to search for relevant information.
  - Evaluation: Assessing performance using tailored metrics (factuality, coherence, synthesis).

The Rise of Local Models

- Local models offer advantages:
  - Data privacy: Keeping sensitive data within the organization's infrastructure.
  - Flexibility and control: Tailoring computing resources for specific needs.
  - Cost savings and pricing stability: Reducing reliance on external providers.
- Challenges:
  - Computational requirements: Needing significant hardware resources.
  - Maintenance and updates: Demanding specialized expertise.

GenAI at Scale: A Bank's Perspective (Citibank)

- Focus on robust security, compliance, and governance.
- Staggered approach:
  - Fostering organizational awareness.
  - Prioritizing thematic use cases.
  - Embracing a multi-model strategy.
- Emphasis on human-in-the-loop processes for deterministic validation.
- Hybrid approach: Building core components in-house while incorporating vendor products.

The Road Ahead

- Continuous innovation, knowledge sharing, and adaptation are crucial.
- Organizations need to balance local and cloud-based deployments.
- Collaboration between AI engineers, domain experts, and business leaders is key.
- Prioritize a human-centric perspective to amplify human ingenuity with AI.

Conclusion

- RAG, local models, and GenAI at scale represent interconnected trends shaping the future of enterprise AI.
- By embracing a balanced approach, organizations can harness the power of LLMs to transform their operations and drive innovation.