AI ANALYSIS

Getting Retrieval Right: The Hidden Work Behind AI That Works

THE AI SUMMIT LONDON

MEIBEL

WHERE COMMERCIAL AI COMES TO LIFE

VOXO

SUMMARY

Spencer Torene's presentation delved into the complexities of ingesting, processing, and retrieving data in AI pipelines, particularly for generative AI systems. He began by explaining the three main components of augmented generation systems: ingest, storage, and retrieval. Torene highlighted the importance of matching the data type to the question type, differentiating between structured, semi-structured, and unstructured data. He utilised examples like legal documents, BBC News transcripts, and product specification sheets to illustrate the need for careful parsing and embedding creation. He emphasised that embedding creation should consider semantic differences within texts to avoid inaccurate representations.

Torene elaborated on the challenges of parsing semi-structured data, such as code and web pages, and the necessity of extracting relevant information while discarding irrelevant tags and scripts. He introduced the concept of structural data, which involves deriving structured data from unstructured sources, using examples like fiction novels and legal codes. By creating graphs of character relationships or citation networks, AI systems can improve retrieval accuracy and fill in gaps missed by traditional embedding matching. Torene stressed the importance of choosing the right database schema based on the type of question being asked, whether qualitative or quantitative, and the specific needs of the AI pipeline.

The session also covered techniques for verifying the accuracy of AI-generated responses. Torene discussed token grounding, OCR model confidence scores, and comparing multiple LLM generations to ensure consistency and reduce hallucinations. He acknowledged the limitations of LLM judgments, noting their potential biases, but pointed out their usefulness alongside other verification methods. The Q&A portion of the session brought up challenges in building ontologies and balancing generalized schemas with specific use cases. Torene shared insights into industry specialisations, noting that their focus is on integrating generative AI into established pipelines rather than creating new business cases. The session concluded with a discussion on the importance of domain expertise and collaboration with engineering teams to implement effective AI solutions.

TAKEAWAYS

Matching data types to question types

Torene emphasised the necessity of aligning the type of data to the nature of the question being asked. Structured, semi-structured, and unstructured data require different approaches to ensure accurate retrieval and processing. Understanding the distinction allows for more effective AI pipelines.

Structural data extraction

Extracting structure from unstructured sources can significantly improve retrieval accuracy in AI systems. Examples like creating character relationships graphs from fiction novels or citation networks from legal texts demonstrate the value of this approach. This method helps fill gaps missed by traditional embedding matching.

Verification techniques for AI responses

Ensuring the accuracy and consistency of AI-generated responses is crucial. Techniques such as token grounding, OCR model confidence scores, and comparing multiple LLM generations help mitigate hallucinations and biases. These methods provide a more reliable measure of AI output quality.

PARTICIPANTS

Spencer Torene
PhD Principal Research Scientist, Meibel

Powered by

voxoevent.ai