Retrieval-Augmented Generation

Workshop 27/05/2024, lesson 5

Retrieval-Augmented Generation (RAG)

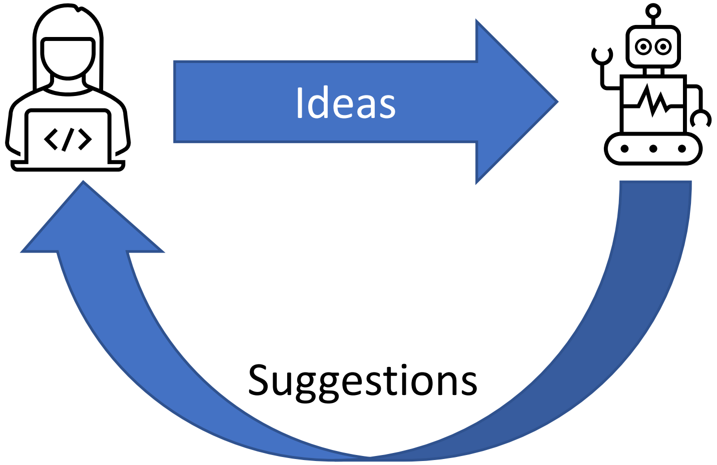
Generative models combined with retrieval mechanisms (a.k.a. database search):

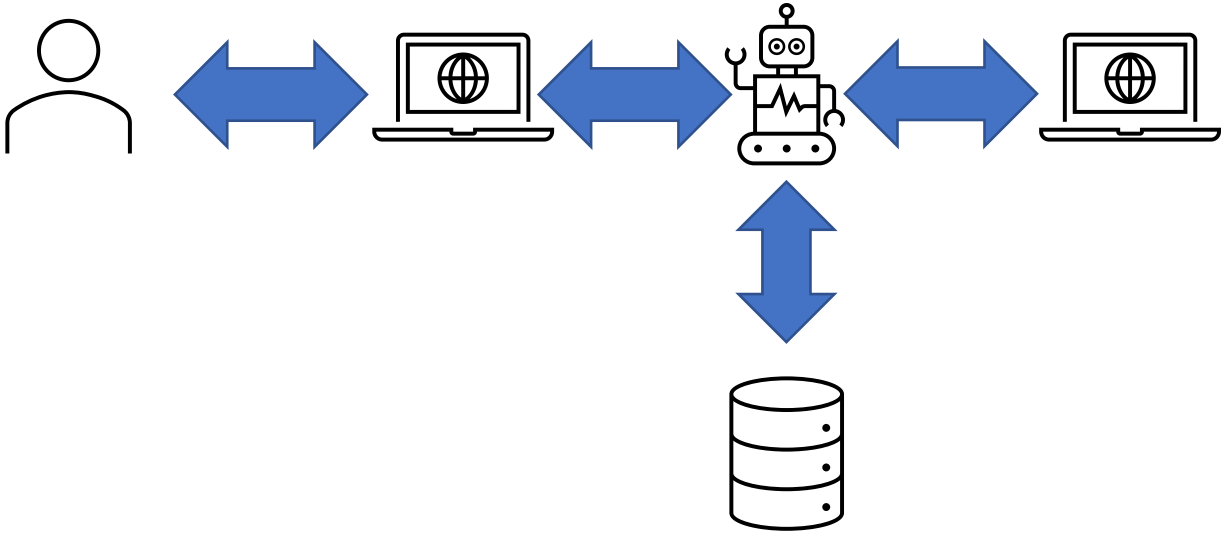
1. **Retrieval Phase**: When a query or prompt is received, the model first performs a search to retrieve relevant documents or data snippets from a structured database or a large corpus of text.
2. **Augmentation**: The retrieved content is then used to augment the input to the generative model.
    1. Additional context, facts, or examples that are not inherently known by the model
    2. Useful for generating accurate and contextually relevant responses.
3. **Generation Phase**: Original input and the retrieved information are combined and presented to the LLM.

The key advantages:

- Use vast amounts of information.
- When applications require factual correctness and depth, such as question answering, content creation, and summarization tasks.

Generative AI developments





Conceptualize, elaborate and refine

Summarize, explain and communicate

Tokenizing and embedding

- Tokenization is the process of breaking down text into smaller pieces:
  - Sentence: "The green plant is growing in a beautiful blue pot"
  - Tokens: ["The", "green", "plant", "is", "growing", "in", "a", "beautiful", "blue", "pot"]
- After tokenization, each token is converted into a numerical form known as an embedding
- These embeddings capture not just the raw word but also aspects of its meaning and its relationship to other words.
  - **The**: [0.1, -0.2, 0.3], **green**: [0.5, -0.4, 0.3] , **plant**: [0.6, 0.1, -0.3], **is:** [0.0, 0.0, 0.0]
  - **growing:** [0.4, 0.5, -0.6], **in** : [0.0, 0.0, 0.1], **a**: [0.0, 0.0, 0.0], **beautiful:** [0.6, 0.6, -0.2]
  - **blue:** [0.2, -0.3, 0.5], **pot:** [0.5, -0.2, 0.3]