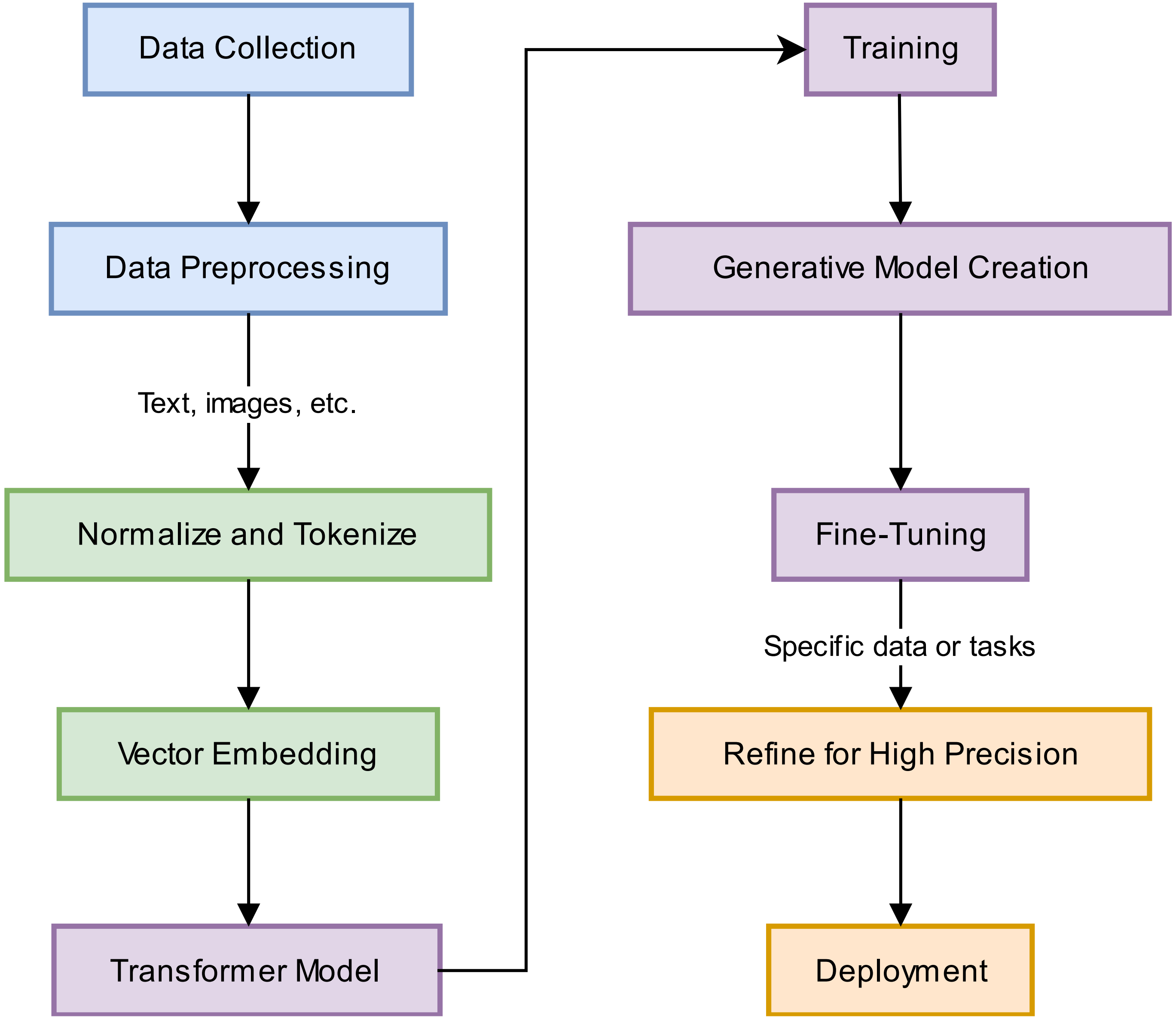
Introduction to LLMs

Workshop 27/05/2024, lesson 1

Artificial intelligence, ML and genAI

- **Artificial Intelligence (AI):** technology to mimic human capabilities such as reasoning, learning, problem-solving, and perception.
  - **Machine Learning (ML):** algorithms and statistical models to perform specific tasks without explicit instructions (pattern recognition and inference)
  - **Statistics:** backbone of AI that provides methodologies for collecting, analyzing, interpreting, and presenting empirical data
  - **Deep Learning (DL):** machine learning based on layered neural networks. DL excels in tasks such as speech recognition, image recognition, and natural language processing
  - **Generative AI:** creating new content: images, sounds, text and code, based on specific training data. Input type does not need to match output type



Data Collection

Data Preprocessing

Text, images, etc.

Normalize and Tokenize

Vector Embedding

Transformer Model

Training

Generative Model Creation

Fine-Tuning

Specific data or tasks

Refine for High Precision

Deployment

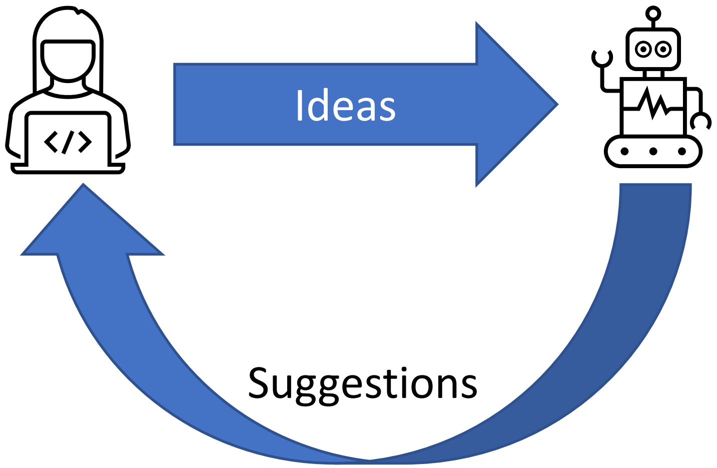
Tokenizing and embedding

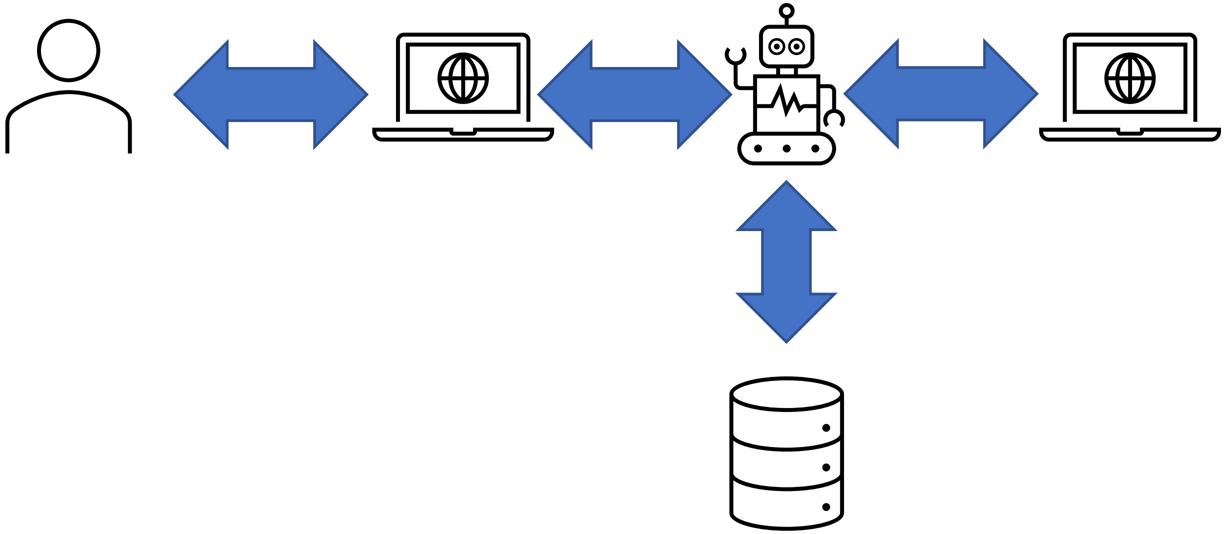
- Tokenization is the process of breaking down text into smaller pieces:
  - Sentence: "The green plant is growing in a beautiful blue pot"
  - Tokens: ["The", "green", "plant", "is", "growing", "in", "a", "beautiful", "blue", "pot"]
- After tokenization, each token is converted into a numerical form 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]

Generative AI

- Processing in the Neural Network:
  - These embeddings are then fed into a neural network, typically a transformer model
  - The transformer uses layers of attention mechanisms to weigh the importance of different tokens relative to each other
- Contextual Understanding:
  - Embeddings they are updated based on the surrounding context within the sequence
  - Calculate overall meaning and how each token relates to the others
- Token Prediction:
  - The final layer of the transformer model outputs a new vector for each input token
  - A softmax function is applied to convert it into a probability distribution over all possible next tokens
- Generating Output:
  - The token corresponding to the highest probability is typically chosen as the next token in the sequence
  - This process is repeated for generating each subsequent token, using the newly generated tokens as additional context

Generative AI developments





Conceptualize, elaborate and refine

Summarize, explain and communicate

Conceptualize, elaborate and refine

- A.k.a. prompt engineering or prompt hacking
- Chat interface with for example ChatGPT, Midjourney, Gemini, etc
- This is how most people know of and interact with generative AI
- Prompts:
    - Could you rewrite the following text to adjust for better reading and also in a less formal one of voice, but not to informal: \`text\`
    - The text will be placed on a blog post for a company. The text you provided is a little bit to informal for a company blog. Could you adjust it a little bit?

Conceptualize, elaborate and refine

- Example usage:
  - Grant proposals
  - Blog posts
  - Ideas for apps (Weather and activity app)
  - Programming, debugging, code generation, technology explanation, etc, etc
  - Logo creation
  - Image generation
  - Etc, etc, etc