[PPT-02] LARGE LANGUAGE MODELS

What is ChatGPT?

- It is an assistant or **AI chatbot** that has conversations (chats).
- It is not a **large language model** (LLM), but a user interface built around one.
- Chatbot arena: ChatGPT (OpenAI), Claude (Anthropic), Gemini (Google), DeepSeek (High-Flyer), Grok (X AI), Coral (Cohere), Le Chat (Mistral), Copilot (Microsoft).

Early ChatGPT critique

- Not factual.
- Limited to training data timeframe.
- Limited to public training data, no access to proprietary information.
- No reasoning.
- Not good at math.
- Hallucination.

State of the art

- Advanced usage:
  - Instruction following.
  - In-context learning.
  - Chain-of-Thought (CoT) reasoning.
- Augmentation:
  - Retrieval-augmented generation (RAG).
  - Tool usage: calculator, web search, Python interpreter.

Tokens

- The **tokens** are the “atoms” in which text is split by the LLM.
- They are typically words, subwords or punctuation. There are also tokens for the beginning and the end of a text.
- The model has a **vocabulary** of tokens. For every token, there is an **embedding vector**.

What do LLMs really do?

- They produce a reasonable continuation of any text, based on what people have written on billions of webpages (the training data).
- The model performs mathematical calculations with the input vectors, using a **neural network** architecture whose parameter values have been determined during the training of the model.
- The resulting vector is a set of **probabilities** for the output token, which is chosen according to these probabilities.

Continuation

- The output token is added to the input tokens. Then, a new output token is generated, and so on, until the end token is generated. The set of tokens generated is the “answer” of the model.
- **Context window**: the maximum number of tokens that the model can manage to respond a single prompt. For “reasoning” models, this is a relevant parameter.

How do we interact with a language model?

- Chat app: ChatGPT, Gemini, DeepSeek, Perplexity, LM Studio.
- Programmatic way:
  - Remote model (API): OpenAI, Gemini, DeepSeek, Groq.
  - Local model: Ollama, Hugging Face.