The concept of large language models (LLMs), explaining their importance and functionality in the realm of AI. LLMs, such as ChatGPT, Gemini, and Llama, utilize a deep learning structure known as the transformer to understand and generate text that resembles human language.

Key points covered in this include:

1. **Architecture**: The transformer architecture allows LLMs to effectively manage long texts and comprehend word relationships.
2. **Popular Models**: Notable models discussed include OpenAI’s GPT series (like GPT-3, GPT-4, and GPT-5), Google’s Gemini, Meta’s Llama, and the Chinese model Deep Seek.
3. **Applications**: LLMs are employed in various domains such as chatbots, document question answering (leveraging retrieval augmented generation), and coding assistance tools like GitHub Copilot.

The term "large" in LLMs signifies the vast number of parameters they contain, which enhances their language processing abilities. For instance, GPT-3 has 1.75 billion parameters. However, larger models necessitate more resources, leading to higher operational costs.

The working process of an LLM involves tokenizing input text into smaller segments, which the transformer model then uses to predict and generate output tokens, facilitating contextually relevant responses.

Furthermore, LLMs can be divided into proprietary and open-source models. Proprietary models (like the GPT series) restrict access to their underlying mechanisms, while open-source models offer more flexibility for modification and control, making them suitable for applications with sensitive data.

In essence, LLMs are robust AI tools capable of producing human-like text by harnessing extensive parameters and advanced architecture. Their selection depends on specific project requirements, infrastructure, and budget constraints. The next lecture will further explore the transformer architecture, which is fundamental to LLMs.