UNDERSTANDING LARGE LANGUAGE MODELS (LLMS)

🎯 What this lesson is about:

In this video, you learned what large language models (LLMs) are, how they work, how they're built, trained, improved, and why understanding them helps you create better automations and agents in n8n.

🔍 What is an LLM?

An LLM (Large Language Model) is a type of AI trained on huge amounts of text from the internet, books, and more. It can answer questions, write stories, summarize content, translate languages, generate code, and more. Tools like ChatGPT, Claude, Gemini, and DeepSeek are all LLMs.

LLMs don't think like humans – they predict the next word in a sentence based on everything they've seen during training. They use probabilities, not understanding.

🧩 How are LLMs built?

LLMs are made of two main files:

- **Parameters file** – The brain. It stores everything the model has learned (can be over 140GB).
- **Runner file** – The body. A small script that loads the parameters and runs the model so you can interact with it.

Think of it like this: parameters = memory, runner = action.

🤔 How do LLMs generate answers?

LLMs work by predicting the next word based on what came before. They don't understand meaning – they choose the most likely next word. You can adjust a setting called **temperature:**

- Low temperature = more predictable and safe
- High temperature = more creative and surprising

UNDERSTANDING LARGE LANGUAGE MODELS (LLMS)

🛠️ How are LLMs trained?

Training an LLM means compressing a huge part of the internet into a model using powerful computers (thousands of GPUs for days or weeks). This creates a massive parameters file full of patterns and knowledge – not an exact copy of the internet, but a compressed version.

Some models like DeepSeek were trained differently – not on internet data, but on other models, making training cheaper and faster.

🔓 Open vs Closed Models:

- **Open-source** models like LLaMA 2, Mistral, Zephyr – You can download and run them locally. They're free to use, and your data stays private.
- **Closed-source** models like ChatGPT, Claude, and Gemini – You access them through APIs or websites, and your data goes to companies' servers.

🔧 Fine-tuning: How LLMs get better

After training, LLMs can be improved with custom data (like Q&A pairs or documents). This makes them better at specific tasks – like customer support, coding, or generating content. Even user feedback helps them improve over time.

Fine-tuning makes a general AI assistant into a helpful expert.

🤖 LLMs as Real Agents:

LLMs are not just chatbots anymore. When combined with tools like Zapier or n8n, they become **AI agents** that can:

- Answer emails
- Summarize documents
- Control apps
- Automate workflows
- Retrieve real-time data using RAG (Retrieval-Augmented Generation)

They're used in business, content creation, development, and much more.