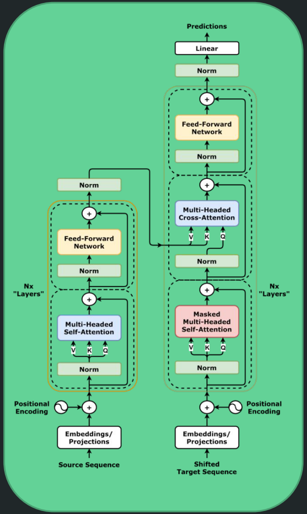
2.2. Large Language Model (LLM)

Introduction to LLMs

A Large Language Model (LLM) is a type of AI model trained to understand and generate human-like text.



Transformer Architecture

- GPT-5
- Gemini 2.5
- Llama 4
- DeepSeek

Popular LLMs

- Chatbots
- Doc QA
- Coding
- Agents

Applications

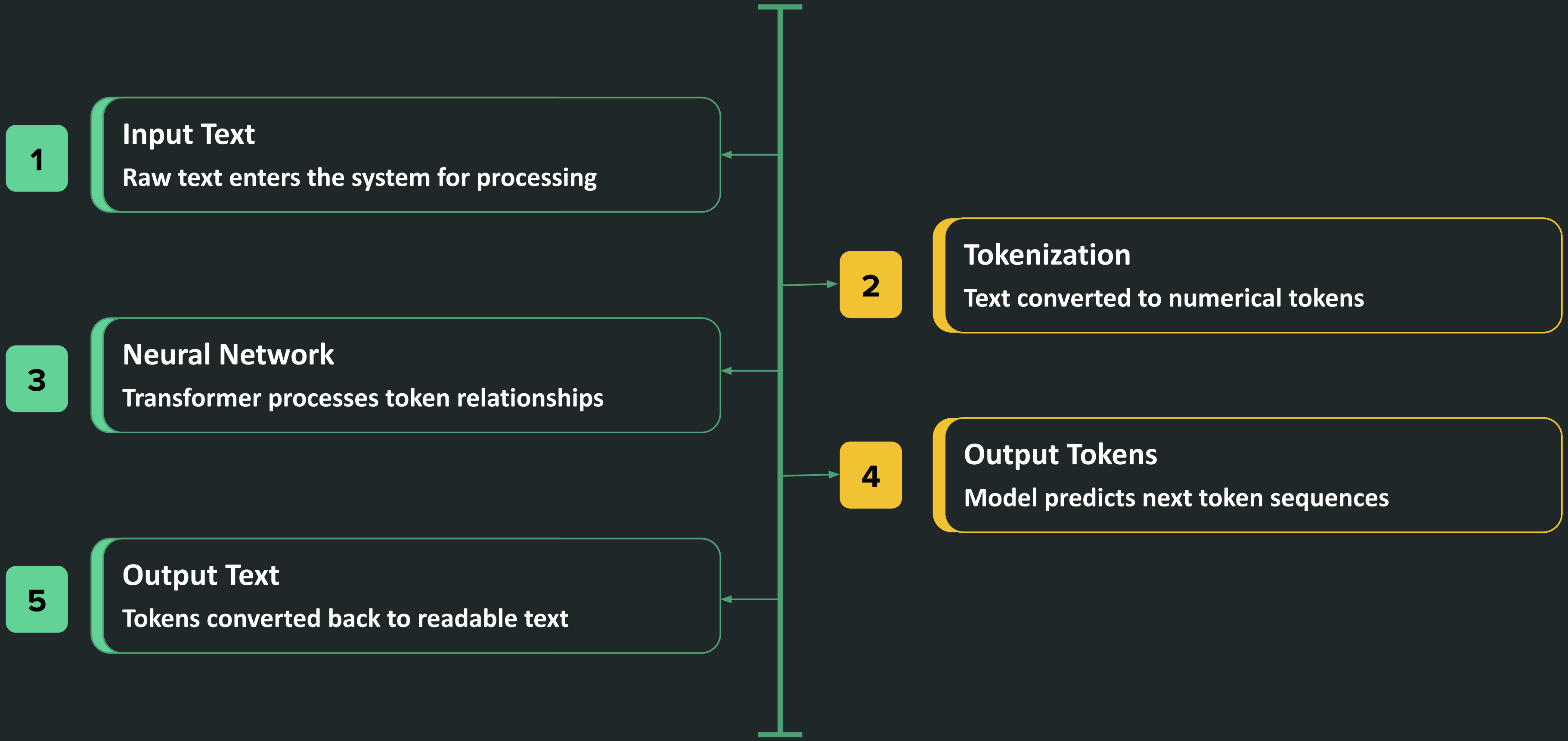
What makes it "Large"?

175B Parameters

Scale & Complexity

GPT-3 has 175 billion parameters, while newer models like GPT-4 & GPT-5 likely exceed this. More parameters enable better understanding but require significantly more computational resources.

The LLM Workflow



1 **Input Text**

Raw text enters the system for processing

2 **Tokenization**

Text converted to numerical tokens

3 **Neural Network**

Transformer processes token relationships

4 **Output Tokens**

Model predicts next token sequences

5 **Output Text**

Tokens converted back to readable text

Tokenization

Tokenization Process

Tokenization breaks text into smaller units like words, subwords, or characters. This allows LLMs to process language efficiently by converting these units into numeric IDs, enabling better generalization and reduced vocabulary size.

- Splits text into smaller semantic chunks
- Converts tokens into numerical IDs
- Enables processing of unseen words

Example

The sentence “ChatGPT is amazing!” becomes
["Chat", "G", "PT", " is", " amazing", "!"]
via subword tokenization, showing how even complex or unfamiliar words are broken into meaningful parts.

- "Chat" stays as "Chat"
- "GPT" splits into "G" and "PT"
- Semantic meaning is preserved through subword units

Note: Actual token splits may vary depending on the model and tokenizer used (e.g., GPT-2, GPT-4, LLaMA, etc.)

Proprietary vs Open-Source LLMs

Proprietary LLMs

- OpenAI - GPT-3, GPT-4, GPT-5, GPT-o4
- Google - Gemini-2.5
- Anthropic - Claude Opus 4.1, Claude Sonnet 4

Open-Source LLMs

- Meta - Llama-2, Llama-3
- Google - Gemma-2, Gemma-3
- OpenAI - GPT-oss
- DeepSeek