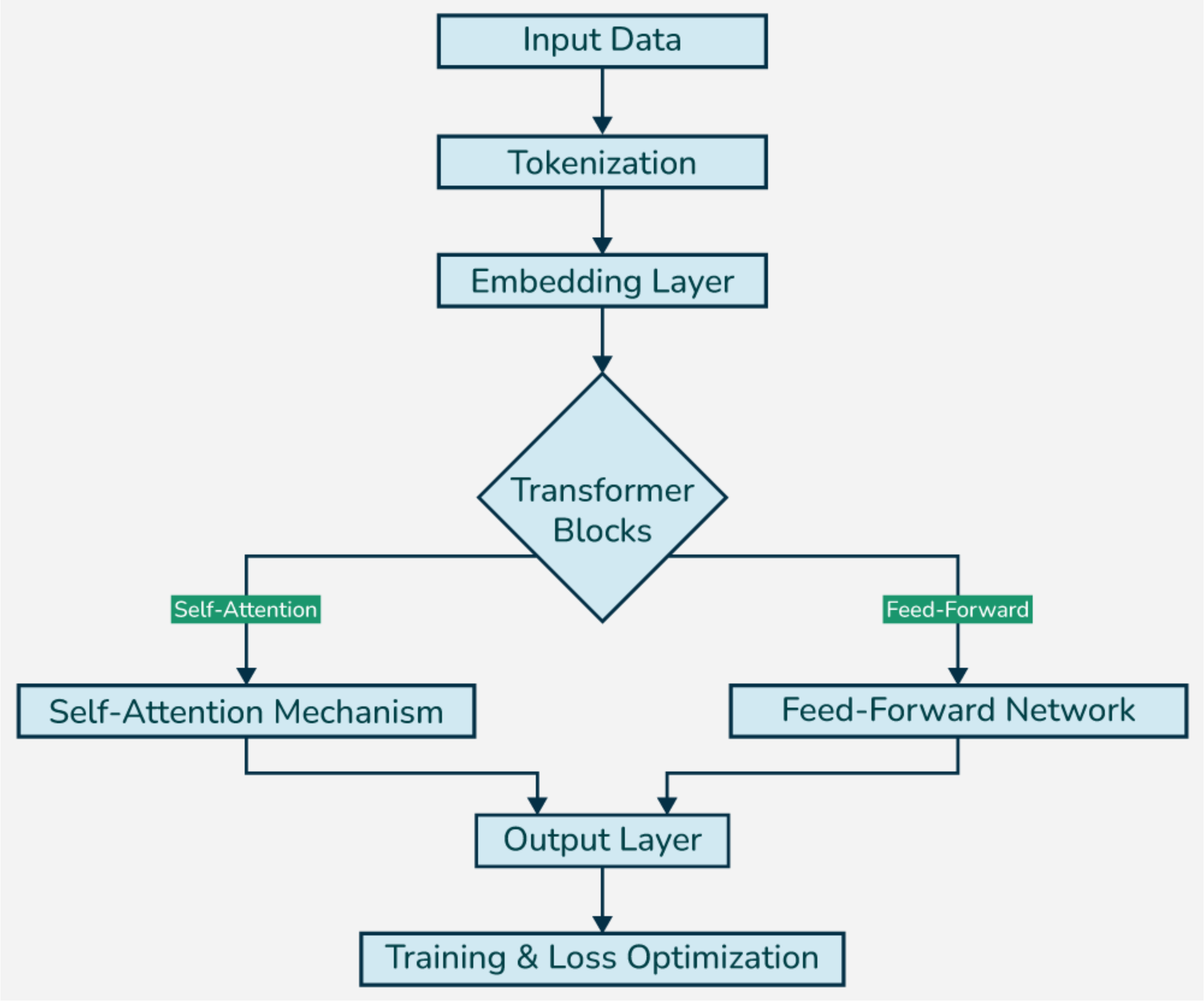

Here's a **simple explanation** of how a Large Language Model (LLM) like GPT works:

🔄 1. Training on Text Data

- It learns from **lots of text** (books, websites, etc.).
- It sees billions of sentences and learns **patterns in language**.

🔢 2. Tokenization

- Text is broken into **tiny pieces** called **tokens** (like words or parts of words).
- Example: "Hello!" → ["Hello", "!"]

🧠 3. Transformer Architecture

- GPT uses a special neural network called a **Transformer**.
- It uses **self-attention** to understand **which words are important** in a sentence.

🔮 4. Predict Next Token

- The model learns to **predict the next token** in a sentence.
- Example: "I like to eat" → predicts "pizza" (if that's common in training data).

💬 5. Generate Text

- It keeps predicting the next word/token until it finishes the sentence or paragraph.

🧪 6. Fine-Tuning (Optional)

- Sometimes, it's trained again on **specific tasks or data** (like legal, medical, or chatbot data).

🏁 Result:

You type a prompt like:
"Write a story about a robot."
And it continues with a relevant, fluent story.
