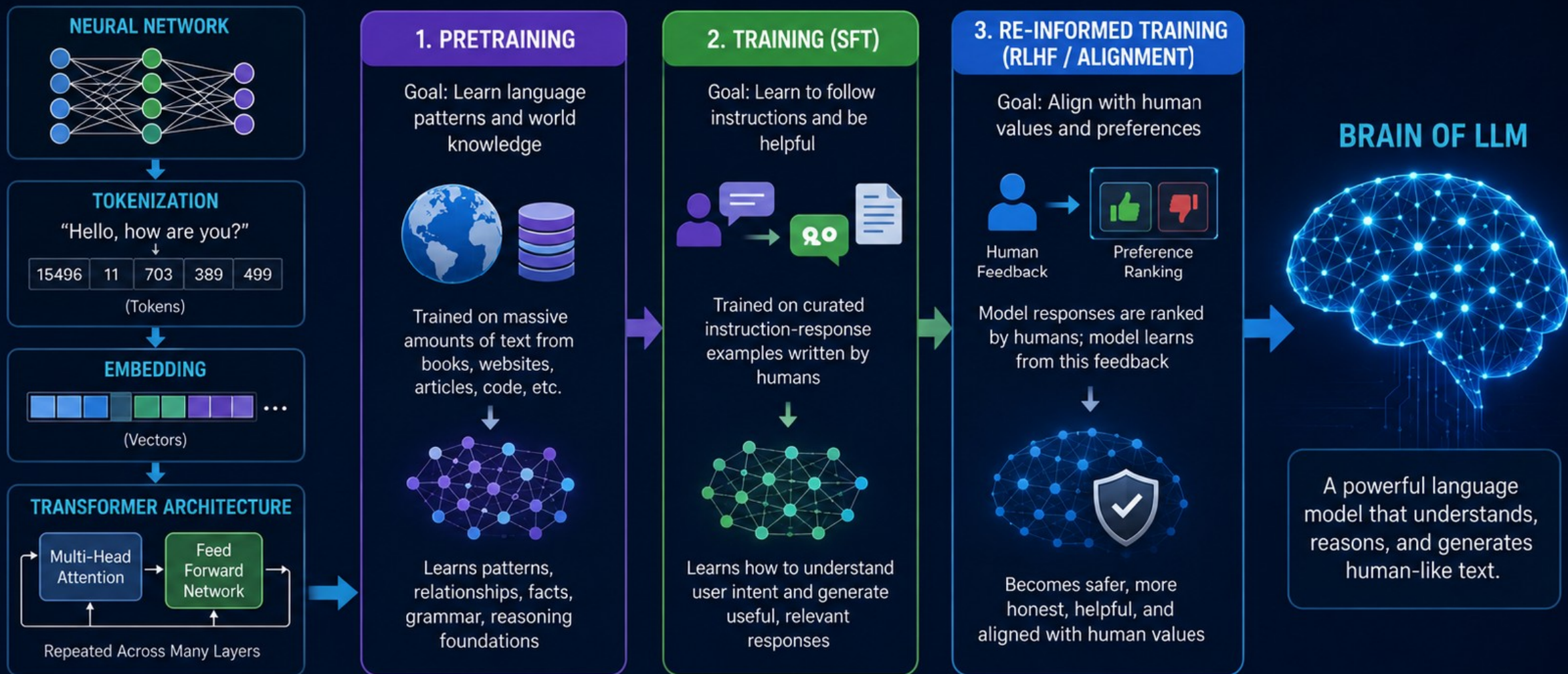
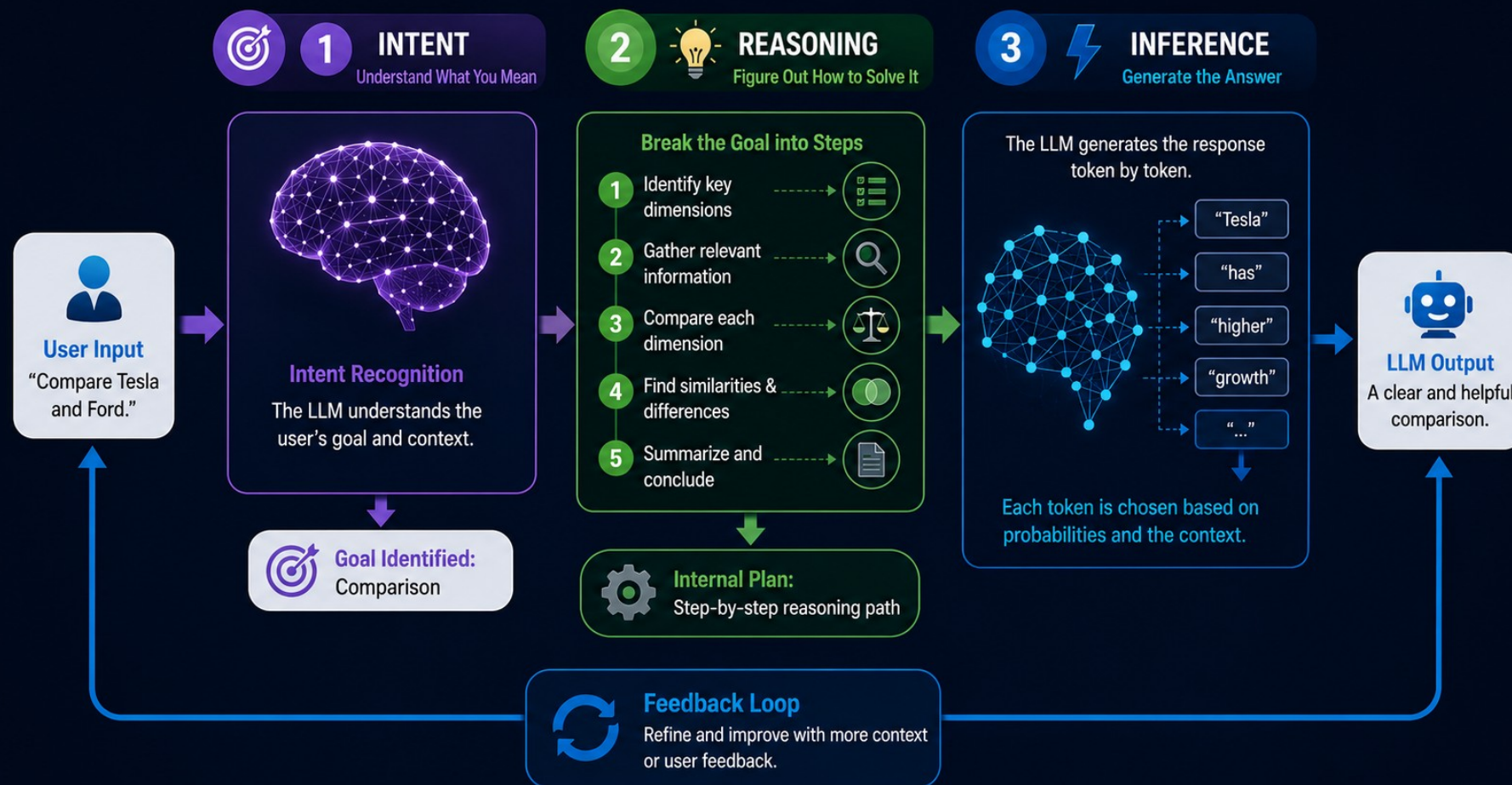


Fundamentals

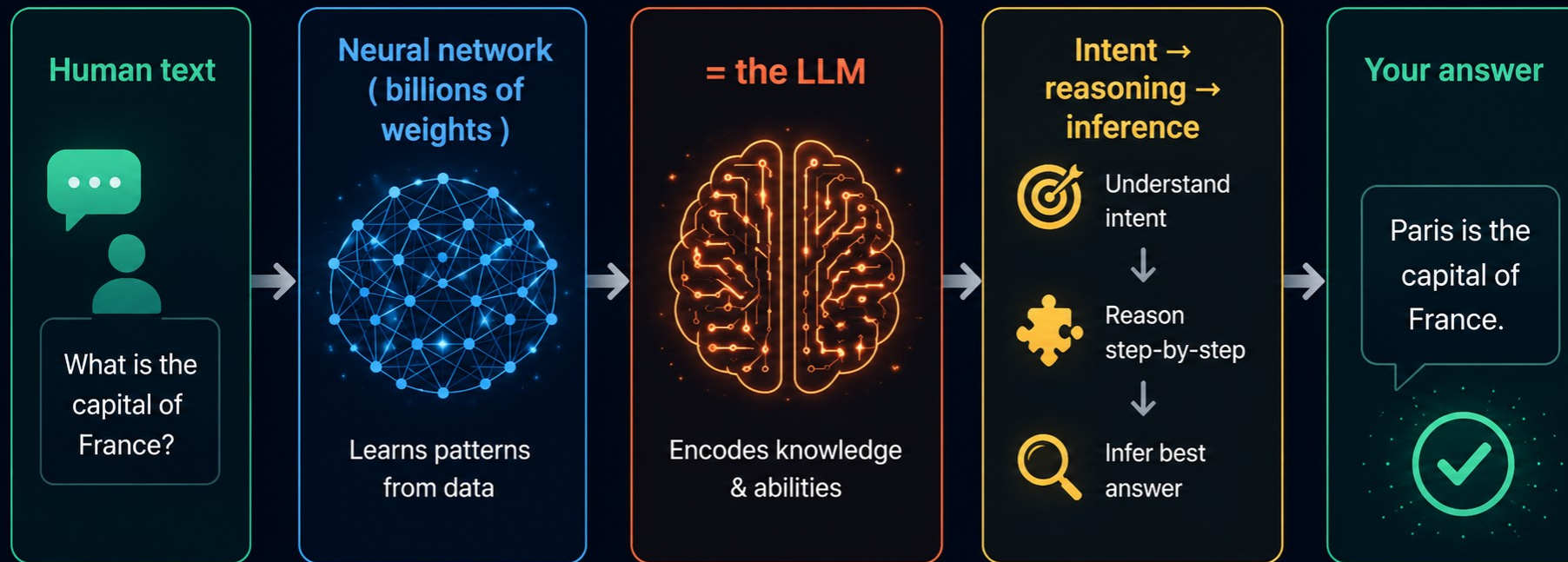
How an AI is actually built



From your question to its answer



So what is really happening?



AI doesn't **know** — it **predicts**.

The Model is the Engine. Apps / Assistants / Agents Sit on Top.

Training a constitution into Claude

THE IDEA • CONSTITUTIONAL AI

Claude is given a written set of values and learns to **check and fix its own answers** against them — the **why**, not just rules.



Amanda Askeff

Philosopher (NYU PhD) who leads Anthropic's work on Claude's character & fine-tuning; lead author of the "soul doc."

HOW IT GETS TRAINED IN • SELF-CRITIQUE + AI FEEDBACK (RLAIF)



THE DEBATE Who decides the values? One company writes the rules — and **grades itself**.