BEFORE WE BEGIN

What is an AI chatbot, actually?

Three layers that shape every response you get

01

The Language Model (LLM)

A prediction engine trained on billions of words. It doesn't think or understand — it pattern-matches at massive scale.

Like: an autocomplete trained on all of human writing

02

Training & Alignment

Human reviewers teach the model values (RLHF). Anthropic also uses a "Constitution" — principles baked into training, not added as rules later.

Like: raising someone with values, not just a rulebook

03

The Consumer App Layer

Consumer apps layer safety guardrails, stress-testing, and a hidden system prompt on top of the model — shaping it into something user-friendly.

Like: seats, airbags, and crash tests turn an engine into a usable car

Today's demo lives inside all three of these layers — and shows what happens when each one changes.

VOCABULARY

Key terms for today's demo

THE MODEL

LLM (Large Language Model)
Trained on massive amounts of text from the internet, books, and code to predict language — not to understand meaning. ChatGPT, Claude, and Gemini are all LLMs at their core.

BAKING IN VALUES

RLHF (Reinforcement Learning from Human Feedback)
Human reviewers score thousands of responses. The model learns to prefer helpful, less harmful answers.

Constitution (Constitutional AI)
Anthropic's approach: a written set of principles the model is trained to follow — like a moral framework baked in during training, not added later.

BEFORE IT SHIPS → POST-TRAINING SAFETY

Guardrails
Hard limits on what the model will say or do — refusing harmful content, flagging crisis language. The model's "no-go zones."

Red-Teaming
Adversarial testers deliberately try to break the model before release. Weaknesses found get fixed.

WHEN YOU USE IT

System Prompt ⚠ not controlled by you
Hidden instructions loaded automatically by the app before you type anything. Written by the company. You can't see, change, or disable it.

User Prompt
What you actually type. The model sees BOTH the system prompt AND your message — but you only control one of them.

The demo will make each of these concrete — watch for them as we walk through each tier.