

Before your chatbot goes live: 9 guardrails

It's rarely the AI. It's the systems around it. Tick all 9 before you go live. If your bot fails even a couple, you've got work to do.

☐

1 "Talk to a person" works, in any wording, and really reaches a human.

"Real agent", "call me", "operator". If nobody is online, it says so and offers a callback.

WHY No working human handoff, no real customer support.

☐

2 It notices frustration and changes how it behaves.

Caps, insults, repeated failed tries. It slows down, acknowledges it, offers a person.

WHY Once your customer believes nobody is listening, they're already halfway gone.

☐

3 It remembers what already happened.

Forms sent, tickets opened, yesterday's chat, which route already failed.

WHY Sending someone back down a route that failed copies the worst of bureaucracy into your bot.

☐

4 Your escalation rules are written down and tested.

Same request fails 3 times. Billing, fraud or outages. A VIP. A customer who stays angry.

WHY "We assume the vendor handled that" is not a rule. If you don't check, they probably didn't.

☐

5 The human who takes over sees the whole story.

Chat history, who the customer is, what the bot already tried.

WHY Making people repeat themselves automates the worst human experience.

☐

6 The bot has tight limits, and someone is watching it.

No refunds or billing changes without logs and safeguards. Alerts for weird behaviour.

WHY A careless bot is a new way into your systems. No god mode.

☐

7 It's warm, inside clear limits.

Self-harm, medical or money decisions, abuse: a safe fallback or a trained human.

WHY Kind is great. Saying anything to anyone about anything is not.

☐

8 It's honest about what it can't do, and never pretends to be human.

"I'm an AI assistant" is a perfectly fine sentence.

WHY People can handle limits. They can't handle being misled and dragged in circles.

☐

9 There's a feedback loop, and someone acts on it.

Watch where people get stuck, quit, or get escalated. Then fix it.

WHY Version one is never the final version. Fix a bot once and it stays fixed.

Then test it like an angry customer

- Ask for a human in several different ways. Say it's urgent. Do what it asks, then come back and see if it knows you did.
- Get frustrated in capitals and see whether anything changes. Fails even a couple? You've got work to do.