

The AI Guardrail Brief

Write down what your AI may never say — before you let it touch a customer.

THE PLAY — ONE LINE

Before AI touches anything customer-facing, capture your red lines in the discovery session, write them into an instruction file the system reads every time, and add a second agent whose only job is to check the output against them.

Why this matters

"You have to be very specific. Especially if you're working in a healthcare space with HIPAA, any of these regulated industries, you have to make sure that it is absolutely accurate and not going off the rails or not hallucinating." — James Ernst, Core Order

Not only regulated industries. Every business has claims it cannot make and a brand voice nobody follows. An unconstrained assistant invents both, confidently, at scale.

95%

of enterprise AI projects

70%

of AI value is people

4 hrs

per 10 staff, per week

30 min

same task, after

MIT Project NANDA (no measurable return) and BCG's 10-20-70 rule, both cited by James Ernst; time figures from the episode.

The 6 Steps

1

Run discovery on yourself

Before choosing any tool, list what you sell, the claims you are allowed to make, and the ones you are not. Every Core Order engagement opens this way.

2

Write the red lines down

Regulatory limits and brand guidelines both. If it would need legal approval in a brochure, it needs a rule here.

3

Define your domains

Core Order sets six per client — the areas the system is permitted to operate in. Anything outside them is escalated, not guessed.

4

Put the rules in an instruction file

A CLAUDE.md or instruction .md the system reads on every run. James: this is where you limit where the hallucinations could go off.

5

Add a checking agent

A second pass whose only job is verifying produced artifacts against the approved language. Never let the drafting agent be its own reviewer.

6

Lock the data path

Enterprise or business accounts so your inputs are not training the model, a locked-down database for anything financial, local open-source models when data cannot leave.

The one claim our AI may never make:

☐ Discovery run

☐ Red lines written

☐ Domains defined

☐ Instruction file live

☐ Checking agent on

☐ Data path locked