The 12 questions we ask before any engagement.

Most teams shipping "AI features" today are running a demo with extra steps. The 12 questions below are what we ask before we agree to help move a system from *"it works in the demo"* to *"it works in production."* If you can answer each in a sentence, you do not need us. If three or more come back fuzzy, you have a project — not a feature.

PRODUCT SURFACE

QUESTION 01

What user job does the AI do?

Name the single job-to-be-done in one sentence. If you need three, the surface is not ready.

QUESTION 02

What is the failure mode the user sees?

When the model is wrong, slow, refuses, or unavailable — what does the screen look like and what does the user do next?

QUESTION 03

What is the ground truth?

How do you know the answer was right? Who labels, who arbitrates, how fast does the loop close?

SYSTEM ARCHITECTURE

QUESTION 04

Where does the model live, and who pays?

API, fine-tune, self-hosted, on-device. Per-call cost, per-user cost, and who absorbs a 3× usage spike.

QUESTION 05

What is the prompt boundary?

Where do user inputs end and your instructions begin? Have you red-teamed injection from the inputs you actually accept?

QUESTION 06

What is in the context window, and who put it there?

RAG sources, retrieval policy, freshness, and what happens when retrieval returns nothing.

OPERATIONS

QUESTION 07

What is the p95 latency budget?

Not the average. The 95th percentile a user will feel. Have you measured it under load with a real provider, not a local mock?

QUESTION 08

What is the rate-limit and quota story?

Provider limits, account limits, and what happens when a single tenant burns the quota for the rest.

QUESTION 09

How do you roll back a prompt?

Prompts are code. What is the revert path, the audit trail, the time-to-recover when a new prompt regresses?

TRUST, SAFETY, AND THE BORING STUFF

QUESTION 10

What data leaves your perimeter, and where does it land?

Provider retention, sub-processors, training opt-outs, and the legal review you actually completed — not the one in the slide deck.

QUESTION 11

Who is accountable for a bad answer?

A named human, a Slack channel, an on-call rotation, a real SLA. *The model* is not an accountable party.

QUESTION 12

What does success look like in 90 days?

A metric, a baseline, a target, a review cadence, and a kill criterion if it does not work.

Production AI, not demoware.