7 Questions to Ask Before Hiring an AI Agency

The insider checklist most AI vendors hope you never read.

Before You Start

The AI agency market has a dirty secret: most of the companies pitching you have never shipped a production AI system. They've watched YouTube tutorials, built a few ChatGPT wrappers, and hired a designer to make it look legitimate.

By 2025, "AI agency" had become the most over-crowded, under-qualified category in tech services. The following seven questions were designed by engineers who've spent years rescuing projects that looked great in demos and failed catastrophically in the real world.

Ask every single one of them. The answers will tell you everything.

Question 1

"Can you show me something you built that is still running in production — not a demo?"

Why this question matters more than anything else on your checklist

Anyone can build a demo. A demo is a controlled environment with hand-picked inputs, no concurrent users, no edge cases, and no one's business on the line if it breaks. Production is a completely different discipline.

Production means the system handles 500 users simultaneously without degrading. It means the AI agent fails gracefully when an API goes down at 2am instead of taking your entire

platform with it. It means someone is being paged when error rates spike — and that someone has a documented runbook for fixing it.

The gap between a working demo and a production-grade AI system is where most agencies disappear.

 Red flag answer: They show you a demo environment, a video walkthrough, or a "client prototype." They talk about what they're building rather than what they've shipped. They pivot to showing you their tech stack instead of live URLs.

✅Green flag answer: They give you a live URL or a client reference you can contact directly. They can tell you the uptime record, the incident history, and what broke during deployment and how they fixed it. They talk about what went *wrong* and what they learned — not just what went right.

> **The insider truth:** A surprising number of "AI agencies" have never maintained a system past launch day. They build it, hand it over, and wish you luck. Ask specifically: "What does your post-launch support look like, and how many of your clients from 12 months ago are still on retainer with you?"

Question 2

"What happens when your AI agent hallucinates — and how have you handled it in a live system?"

The question that separates engineers from prompt engineers

Hallucinations aren't a bug that gets patched in the next release. They're a fundamental property of large language models that every production AI system must be architected to handle, not hope away. Any agency that doesn't have a direct, detailed answer to this question has never shipped a real AI system.

The question isn't *whether* your AI will hallucinate. It will. The question is whether the agency has built validation layers, confidence thresholds, human-in-the-loop checkpoints, and output sanitisation into the architecture from the beginning — or whether they'll bolt something on after your users start complaining.

 Red flag answer: "Our model is very accurate" or "We use GPT-4 so it barely hallucinates." Both statements reveal a dangerous misunderstanding of how LLMs work. Accuracy is not a fixed property. It degrades with edge cases, adversarial inputs, prompt drift, and model updates.

✓Green flag answer: They describe specific mechanisms: output schema validation, fallback routing when confidence is below a threshold, human review queues for high-stakes decisions, and how their monitoring stack catches hallucination patterns before users do.

> **The insider truth:** The agencies who wave away hallucination concerns are the ones whose clients call Susea six months later with broken systems and furious customers. Ask them to describe a specific hallucination incident from a previous project and walk you through exactly what they did. No story = no production experience.

Question 3

"Who actually writes the code — and can I talk to them?"

The question the sales rep really doesn't want you to ask

The AI agency industry has a tiered outsourcing problem that almost no one talks about publicly. A polished US-based agency website, a well-spoken account manager, a sleek case study PDF — and then the work gets handed off to an overseas team of junior developers who've never met each other and are juggling 12 other projects simultaneously.

This isn't a geographic bias. It's a quality control, accountability, and communication chain problem. When your production system breaks at 11pm on a Friday, you need to reach someone who has context on every architectural decision made during the build. Outsourced relay chains collapse under that pressure.

 Red flag answer: Vague answers about their "global team" or "talent network." Reluctance to introduce you to the engineers before signing. A sales process where you only ever speak to non-technical people. Contracts that don't specify who owns post-launch accountability.

✓Green flag answer: The engineer you'll work with is in the room during the sales conversation or available for a pre-contract technical call. They can speak to specific architectural choices, have opinions about your stack, and push back on unrealistic timelines with specifics. You're hiring engineers, not a brand.

> **The insider truth:** Ask directly: "Will the engineers I meet before signing be the engineers who build and maintain my system?" Get it in the contract if the answer is yes. The handoff from "sales team" to "delivery team" is where most client-agency relationships begin to deteriorate.

Question 4

"What's your security review process for AI-generated code?"

The question that could save you from a breach you don't know is coming

AI-generated code — whether it came from Cursor, Copilot, GPT-4, or your agency's internal tooling — has a consistent and well-documented security profile: it produces functional code that is frequently insecure by default.

Studies of AI-generated codebases consistently find hardcoded credentials, missing input validation, CSRF vulnerabilities, SQL injection exposure, and authentication gaps. These aren't theoretical risks. They're in production systems right now, owned by companies who paid a premium price to an agency that never ran a security review.

An agency that uses AI tooling to build faster (which every good agency should) without a mandatory security audit layer is shipping you a ticking clock.

 Red flag answer: "We follow best practices" with no specifics. No mention of penetration testing, OWASP guidelines, or static analysis tooling. Treating security as a post-launch add-on or a separate service you can purchase later.

✅Green flag answer: They describe a specific security checklist applied to every codebase — authentication hardening, secrets management, input sanitisation, dependency vulnerability scanning. They can name the tools they use (Snyk, Semgrep, etc.) and have a documented process for handling discovered vulnerabilities. Bonus: they've done this before the contract is signed, not after.

> **The insider truth:** Ask specifically: "Has any system you've built ever been involved in a data breach or security incident?" A good agency will answer honestly and tell you what they changed. An agency that says "never" without any qualification either hasn't built enough to know, or isn't being truthful.

Question 5

"What's your honest timeline to production — and what causes the most slippage?"

The question that reveals whether they respect your intelligence

Six weeks. Every AI agency says six weeks. Then it becomes ten. Then fourteen. Then you're explaining to your board why the AI initiative that was supposed to launch in Q1 is now a Q4 maybe.

Timeline inflation isn't laziness — it's an incentive problem. Agencies win deals with optimistic timelines and absorb the friction of extended projects at your expense. The honest answer to "how long will this take?" always involves a list of dependencies, risks, and caveats. An agency that gives you a clean number without qualifications is telling you what you want to hear.

 Red flag answer: A specific timeline delivered with confidence and no caveats. No discussion of what could extend the timeline. No questions about your data readiness, existing systems, or internal resource availability before quoting a duration.

✅Green flag answer: They give you a timeline with explicit dependencies attached: "Four weeks, assuming your API credentials are ready on day one, your data is consolidated, and you have an internal stakeholder available for two hours per week for review. The most common causes of slippage are X, Y, and Z — here's how we mitigate each." That answer costs them deals. It should earn your trust.

> **The insider truth:** The fastest path to a destroyed timeline is an agency that under-scopes discovery. Ask: "What does your discovery and scoping process look like, and how long does it take before you write a single line of code?" Agencies that skip straight to building are the ones who rebuild the same feature three times.

Question 6

"Do you fix AI systems you didn't build — and what does that process look like?"

The question that reveals how deep their expertise actually goes

There's a particular kind of technical confidence that only comes from diagnosing someone else's mess. Any engineer can build a system to their own specifications. It takes a genuinely senior practitioner to walk into an unfamiliar codebase, map the architecture in their head, identify where the original decisions went wrong, and fix it without introducing three new problems.

Agencies that only build from scratch have comfortable, controlled experience. Agencies that regularly rescue broken systems have seen every failure mode up close. The latter is who you want building yours.

 Red flag answer: "We only work on greenfield projects" or visible discomfort with the question. Inability to describe a structured approach to auditing an unfamiliar system. Immediately pivoting to a rebuild proposal without asking to see the existing codebase first.

✅Green flag answer: They have a documented diagnostic process: codebase audit, architecture mapping, failure mode identification, prioritised remediation plan. They've done it before and can give you a reference from a client whose broken system they rescued. They ask smart questions about your existing setup before quoting anything.

> **The insider truth:** This question works as a reverse proxy for experience. Building is easy. Rescuing is hard. The agencies with genuine production depth fix things. The ones with polished decks and optimistic timelines build demos.

Question 7

"What does the relationship look like 12 months after launch — specifically?"

The question that separates partners from project vendors

The most important moment in your AI agency relationship isn't the launch. It's month four, when the model starts behaving unexpectedly on edge cases you didn't anticipate. It's month

seven, when your user volume triples and the system starts degrading under load. It's month eleven, when OpenAI updates their API and breaks three of your integrations simultaneously.

Every AI system in production requires ongoing care: model performance monitoring, prompt drift correction, dependency updates, retraining triggers, and architecture evolution as your use cases expand. An agency that doesn't talk about this before you sign either doesn't plan to be around for it or doesn't understand it's necessary.

 Red flag answer: A maintenance retainer offered as a checkbox on the contract with no specifics. Vague promises about "ongoing support." No proactive monitoring plan. An agency that frames launch as the finish line rather than the starting line.

✅Green flag answer: They can describe, specifically, what they monitor post-launch: error rates, latency percentiles, token costs, model output quality metrics, and user satisfaction signals. They have a documented incident response process with SLAs. They proactively schedule quarterly architecture reviews. They treat your success after launch as a portfolio asset, not a contractual obligation they tolerate.

> **The insider truth:** The AI systems that deliver real ROI are the ones that get progressively smarter and more efficient over time — because someone is watching the data, identifying opportunities, and iterating. Ask: "Can you show me what the monitoring dashboard looks like for a current client, and how often do you make changes post-launch?" The answer tells you everything about whether they see this as a partnership or a transaction.

The Scoring Guide

After each conversation, rate the agency 1–3 on each question:

	Question	❌ Avoid	⚠️ Proceed with caution	✅ Strong signal
1	Live production proof	Demo only	Some references	Live URL + reference
2	Hallucination handling	Dismisses it	Generic answer	Specific mechanisms

3	Who writes the code	Vague / offshore relay	Some transparency	Meet the engineers
4	Security process	"Best practices"	Partial process	Named tools + audit
5	Honest timelines	No caveats	Some dependencies listed	Risk-first scoping
6	Can they fix existing systems	No	Reluctant	Documented process
7	Post-launch specifics	Vague retainer	Some monitoring	Full documented plan

Score 18–21: Strong candidate. Move to technical reference checks. **Score 12–17:** Proceed carefully. Clarify every yellow flag in writing before signing. **Score below 12:** Walk away. The risk of a failed deployment significantly exceeds the cost of finding a better partner.

One Final Question They Won't Expect

Before you end the conversation, ask this:

"What's the most expensive mistake you've ever made on a client project — and what did you change because of it?"

The agency that answers this question honestly, with specifics and without defensiveness, is the agency that has earned the right to make mistakes on your behalf and fix them before they become your problem.

The agency that can't answer it has either never made a mistake — which means they haven't built enough — or they don't respect you enough to be honest about it.

Both are disqualifying.

This checklist was developed by the engineering team at Susea.ai based on 200+ AI system audits, 40+ production rescues, and conversations with founders who learned these lessons the expensive way.

If you'd like a second opinion on an agency you're evaluating — or a free 15-minute technical review of a proposal you've received — our engineers are available at susea.ai/audit

---

Susea.ai | Fix · Build · Deliver *AI agents that work in production, not just in demos.*
