Why 90 Percent of AI Agent Startups Will Fail, and the Architecture That Survives

What the last year of agent demos has taught us about durable products, and the layer underneath the ones that hold up.

Rohit Purohit, Founder and CEO, ViitorCloud Technologies

Enterprise practice of ViitorCloud Technologies

ISO 27001 certified since 2009

enterprise.viitorcloud.com

For the past year, everyone seems to be building AI agents.

Repositories, frameworks, and demos are everywhere. The morning feed is full of multi-agent systems, autonomous copilots, prompt libraries, and AI teams. The excitement is justified. The category is real.

After working closely with these systems through several client programs, I keep coming back to a simple conclusion.

Most AI agent products today are not products. They are prompt orchestration layers sitting on top of large models.

The uncomfortable part is that prompts, personas, and many of the agent frameworks themselves will commoditize quickly as the underlying models improve. We have already seen this pattern play out.

- Early prompt marketplaces appeared in 2023 and disappeared inside a year.

- Many generic AI copilots launched in 2024 struggled the moment the foundation model layer caught up with them.

- Sophisticated demos keep failing when they meet real operational workflows. The reason is simple. AI alone is not the product. Execution infrastructure is.

The Market Says the Same Thing

The 2026 numbers underline this.

The agentic AI market grew from roughly $5.25 billion in 2024 to $7.84 billion in 2025, with credible projections crossing $52 billion by 2030. Capital is flowing in fast.

At the same time, Gartner has now published the widely-circulated estimate that more than 40 percent of agent projects will be cancelled or fail to ship by 2027. Vertical AI startups raised over $15 billion in 2025 alone, and the verticals are already separating. Compliance, support, deal-making, incident response, and developer tooling each have a recognizable top three by mid-2026.

The signal is consistent. Capital is concentrated in the agent layer. Survival is concentrated in the layer underneath.

The Architecture That Survives

The systems that will outlast the next wave will likely share the same five-layer shape.

1. Context layer. A real understanding of the environment the agent operates in. The repositories, the documents, the data, the historical decisions. Most failures live here.
2. Tool execution layer. Agents have to actually interact with the systems where work happens. GitHub. CRM. Analytics. Cloud infrastructure. Internal platforms. Without this layer, the agent has nowhere to act.
3. Workflow orchestration. AI embedded inside execution pipelines. Idea, specification, architecture, implementation, testing, deployment. The agent is a step in the workflow, not the destination.
4. Memory and learning. Systems that improve over time by learning from organizational patterns and past outcomes.
5. Governance and human oversight. AI proposes actions. Humans approve execution where it matters. Auditable, reversible, and explainable. Notice what is not on this list. The model itself. The orchestration framework. The prompt library. Those layers are real and important, but they are not the moat. This is why the most durable AI platforms in production today are not chat assistants. They are embedded inside workflows. - GitHub Copilot succeeds because it lives inside the developer workflow. - Notion AI works because it sits inside the knowledge workspace. - Figma AI is

powerful because it operates within design systems and files. In each case, the moat is the vertical environment where the work actually happens. The model can be swapped. The environment cannot.

Execution Operating Systems for Verticals

The next generation of durable AI companies will build what is starting to be called the execution operating system for a specific vertical.

A few examples we are watching and shipping inside our own practice:

- AI Legal OS for law firms. SimplerToday.ai is one example we have been close to. Strong results inside the workflow, not at the surface.

- AI delivery OS for software teams. ViitorCloud has been building this for years inside our own delivery operations.

- AI marketing execution platforms. Already shipping with several of our clients.

- AI financial operations infrastructure. Quiet but real, especially for mid-market enterprises. The pattern is the same in each case. Agents are not the product. They are workers inside a larger execution engine.

A Simple Test A test I have started running on agent products, including our own.

If removing the underlying LLM breaks your product entirely, you probably do not have a durable product yet. The model should be a replaceable engine, not the foundation of the business.

This test sounds harsh. It is meant to be. A product that depends entirely on the current capabilities of a single model is not a product. It is a feature, and the feature lives at the discretion of whoever ships the next model.

Where This Goes Next

We are still very early in the agent infrastructure cycle. The next 24 months will likely shift the conversation from "AI assistants" to "AI execution systems," and from horizontal agent platforms to vertical operating systems with deep integrations.

The companies that understand this shift early will build the platforms others depend on. The ones that do not will become case studies for the next vintage of investors.

The interesting question, for any team building in this space right now, is the one underneath the obvious one. Not "what can our agent do?" The question is "what is the system our agent lives inside, and would that system still hold value if the agent did not exist?"

Curious to hear how others are thinking about this.

ViitorCloud Enterprise

Deep expertise for enterprise programs that have to hold up in production.

enterprise.viitorcloud.com | enterprise@viitorcloud.com | (+91) 84889 64723