

We built it on ourselves first.

The Outcome

Presidio deployed an agentic AI operating layer across its own pre-sales and delivery workflows — proving cloud launch, scaled innovation, on-prem migration, and the hybrid model that reduces token spend. The result is a reference architecture Presidio now brings to every client engagement.

How We Got There

We started in the cloud, proved the model at scale, then migrated to on-premise infrastructure — reducing token spend by an order of magnitude. 700+ users. 15B+ tokens processed. Every pattern tested on live workflows before it was recommended to a client.

Why It Worked

You can't recommend an AI operating model you haven't lived. Every client engagement Presidio leads in AI consulting, implementation, infrastructure, and managed services is grounded in what we built — and stress-tested — on ourselves first.

The Situation

Like most large professional services organizations, Presidio's institutional knowledge was distributed across systems that didn't connect: CRM records, SharePoint, Teams threads, email archives, delivery standards, and years of project-accumulated spreadsheets. Each held value. None of it was available at the moment decisions needed to be made - in the middle of a sales pursuit, during a delivery handoff, or when a compliance question surfaced three days before signature.

The Approach

We started in the cloud - the right environment to move fast, iterate, and prove the model without upfront infrastructure investment. Once the operating model was validated at scale across 700+ users and 15B+ tokens, we migrated the workload to on-premise infrastructure.

The Result

The result was a significant reduction in token spend and a hybrid architecture that gave the organization both the flexibility of cloud and the economics of on-prem at scale. Workflow agents handle opportunity qualification, statement-of-work drafting, and discovery preparation. Delivery agents surface standards and compliance requirements at authoring time. Shared-context infrastructure turns meeting transcripts into structured inputs for the next five downstream decisions.

You can't build conviction in an operating model you haven't lived. Presidio is its own Customer Zero — the proof that the end-to-end AI journey, from cloud inception to on-prem scale, works in production.

700+

Employees enabled on the internal AI operating platform

15B+

Tokens processed — pre-sales through delivery workflows

10x

Order-of-magnitude productivity gain across pre-sales workflows