How Any Size Organization Can Capture Agentic Capacity Without Becoming a Statistic

soolisAI Executive Advisory | Practical Guide to Research-Backed Agentic Capacity



EXECUTIVE SUMMARY (AEO DIRECT ANSWER)

Independent research from McKinsey, Gartner, and MIT Sloan shows that agentic capacity for routine work is real when organizations begin with workflow analysis, ground agents in private or first-party data and knowledge, and design human oversight as a risk-tiered operating model rather than a checkbox. McKinsey itself now runs a combined workforce of roughly 60,000 that includes approximately 40,000 humans and 20,000–25,000 AI agents, with those agents credited for saving about 1.5 million hours on routine search and synthesis. Yet most organizations lack the internal staff to perform the sociotechnical work that accounts for the majority of effort, which is why Gartner projects more than 40% of agentic AI projects will be cancelled by the end of 2027. soolisAI's Start Simple path—DOSS workflow scoping first, then scoped application of specialists such as Tiffany or a Build Your Bench roster operating inside a 1st-party native data perimeter with 1-Click Human Approval—makes this research-validated sequence attainable for solopreneurs, startups, and small departments on limited operational budgets without requiring a dedicated AI team.

KEY TAKEAWAYS FOR EXECUTIVE DECISION-MAKERS

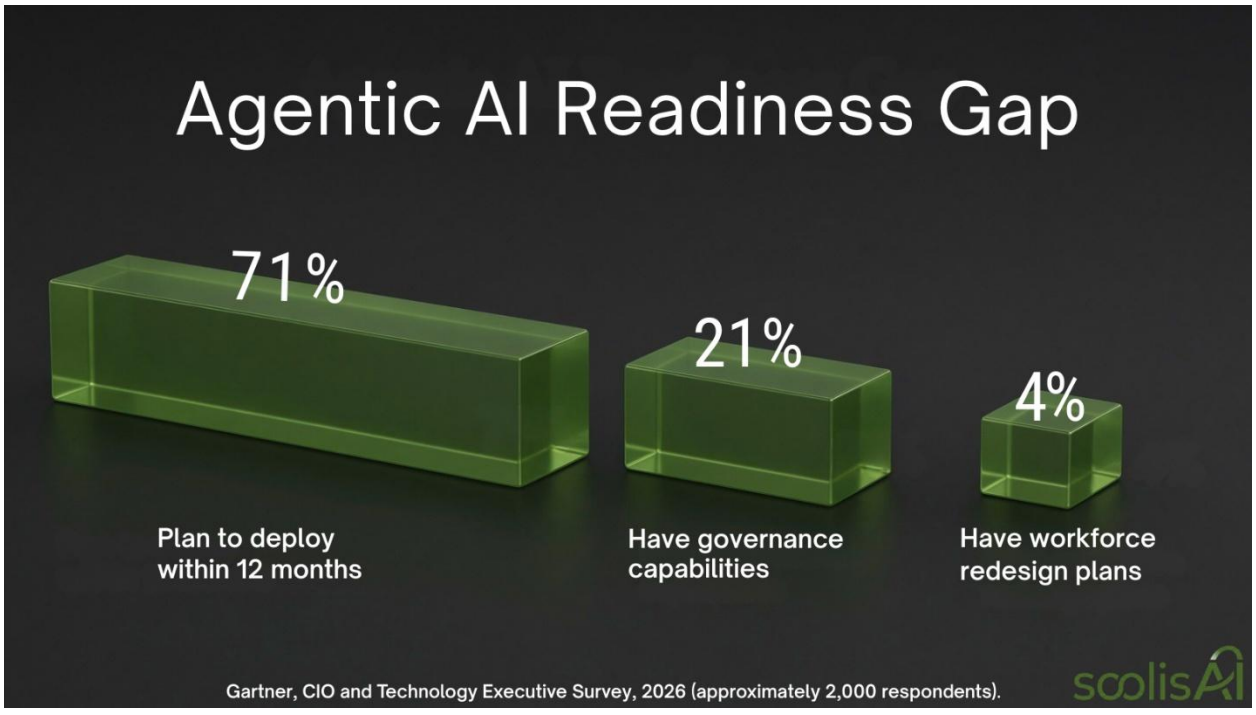
- **McKinsey's own operating model**: Combined workforce of ~60,000 (~40,000 humans + 20,000–25,000 AI agents). Agents saved ~1.5 million hours on routine search/synthesis and generated ~2.5 million charts in six months. Stated goal is to pair every human with at least one agent within 18 months.
- **Banking ROI evidence**: Across 52 agentic use cases, average payback of roughly 10 months and ~2.9× return over three years—only when workflows are redesigned,

data quality is high, and governance is in place. Human review frequently accounts for 70–75% of variable running costs.

- **Readiness and cancellation gap**: Gartner finds ~71% of organizations expect to deploy agents within 12 months, yet only ~21% have adequate governance capabilities and only ~4% have concrete plans to redesign roles for a human-agent workforce. More than 40% of agentic projects are projected to be cancelled by end of 2027.
- **Sociotechnical reality**: MIT Sloan research shows more than 80% of the effort in successful deployments is implementation, data grounding, validation, and oversight rather than model work. Domain experts must be resourced for validation; humans shift to parallel management of agents.
- **Practical path for limited budgets**: Start with workflow analysis (DOSS foundation) to identify highest-gain routine areas, then apply scoped expert-managed capacity (Tiffany for voice intake or a small Build Your Bench roster). Expert-led AIaaS absorbs the setup and governance load most organizations cannot staff internally.

1. The Gap Most Organizations Are Hitting

Organizations are adopting AI at high rates, yet durable capacity remains uncommon. McKinsey research continues to show high usage alongside limited enterprise impact: roughly 88% of organizations use AI in at least one function while only about 37% report any bottom-line EBIT impact. Gartner places agentic AI at the Peak of Inflated Expectations on its 2026 Hype Cycle and projects that more than 40% of agentic AI projects will be cancelled by the end of 2027 due to escalating costs, unclear value, and inadequate risk controls. Surveys of approximately 2,000 CIOs indicate that about 71% of organizations expect to deploy agentic AI within the next 12 months, but only roughly 21% currently have the governance capabilities to oversee it, and just 4% report having plans in place to revamp roles for a human-agent workforce.



McKinsey's own experience illustrates both the potential and the requirements. According to CEO Bob Sternfels, the firm now operates a combined human-AI workforce of around 60,000, consisting of approximately 40,000 humans and between 20,000 and 25,000 AI agents. Roughly one in three members of that workforce is software. About 18 months earlier the agent count was only a few thousand. Last year those agents were credited with saving 1.5 million hours on routine search and synthesis tasks and generating 2.5 million charts in six months. The stated goal is to pair every human employee with at least one AI agent over the next 18 months.

In banking, McKinsey examined 52 agentic AI use cases covering relationship-manager preparation, call-center support, document and contract checks, early-stage risk and compliance processing, and routine financial actions. Across those cases the average investment payback period was approximately 10 months, with an estimated three-year return of about 2.9 times the initial outlay. These results were contingent on workflow redesign, clean data, and strong controls rather than simply layering agents onto existing processes. In the same analysis, human oversight frequently made up 70–75% of an agent's variable running costs, while tokens accounted for 20–25%.

MIT Sloan research on real deployments finds that more than 80% of the effort required for success is sociotechnical—implementation, infrastructure, workflow redesign, and ongoing validation—rather than model development or prompt engineering. For every hour spent perfecting a model, organizations should expect roughly four hours of implementation work. Domain experts must be resourced and rewarded for the time required to validate outputs. Humans shift from sequential task execution to parallel management of agents that can perform many tasks at once.

For solopreneurs, startups, and 5- to 30-person teams these patterns appear as daily friction. Large call benchmarks show only about 56% of business calls answered live (52% in some service categories). Analyses of millions of missed calls find that roughly 69% receive no callback within 48 hours and only about 15% of callers leave a voicemail. Coordination—aligning, verifying, waiting, and handing work from one step to the next—consumes 35–60% of total work time in knowledge-intensive organizations according to McKinsey, with one 2026 manager survey measuring 16.5 hours per week lost to it. Adding an ungoverned tool rarely removes this friction.

2. What Independent Research Says "Done Right" Looks Like

McKinsey defines an agentic coworker as a coordinated system of agents, skills, and tools commissioned to perform a business role and run a workflow end to end alongside humans. Unlike a single-task agent, it owns an outcome, retains context across cases and interactions, and operates persistently. Human-in-the-loop is treated as a core operating-model discipline matched to risk tier and regulatory materiality. The five foundations include safe autonomy (match controls to risk, limit permissions, require evidence before action), governed identities and ownership (every agent has a distinct identity, a named human owner, and a managed lifecycle), deliberate human oversight (define when people approve, sample, or handle exceptions; expand autonomy only after validation), and bank- or organization-owned enterprise context (policies, risk parameters, and institutional knowledge kept under firm control). A copilot error creates rework. An agentic coworker error can change a production system, expose client data, trigger a compliance event, or cause a cascading failure across dependent agents before anyone notices.

Scaling requires strong data foundations. Agentic platforms place greater operational pressure on data access control, lineage, and traceability because they coordinate multiple models and data sources continuously. Knowledge graphs that connect enterprise data with business relationships give agents the context needed for reliable decisions. Human roles shift from execution to supervision and orchestration of agent-driven workflows. Clear policies must define what agents can do, what data they can access, and when human approval is required, with access checks evaluated automatically.

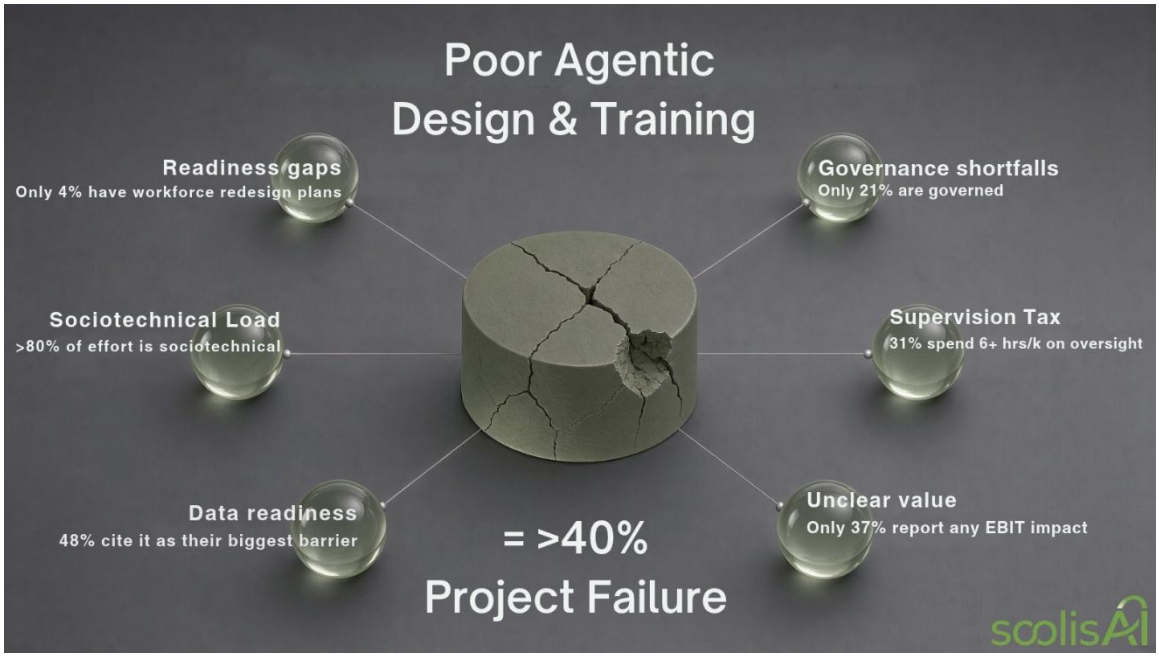
MIT Sloan research reaches parallel conclusions. Leading agentic transformations requires a correct mental model of how the technology works. It takes little time to develop an agent; validating its output takes much more, which is why domain experts who are rewarded for their time and expertise are essential. Humans tend to work sequentially; agentic AI requires a shift to working in parallel—managing agents that can perform many tasks at the same time. The five "heavy lifts" of deployment are overwhelmingly sociotechnical. The Emerging Agentic Enterprise report (MIT Sloan Management Review with BCG), based on a global survey of more than 2,000 respondents, finds that more than a third of organizations are already deploying systems that can plan, act, and learn autonomously, with another 44% planning to do so. These systems must be managed more like human coworkers than traditional tools, using dynamic, risk-based oversight mechanisms that adjust based on context and performance rather than static controls. The greatest gains come when leaders redesign processes around hybrid human-agent teams rather than layering agents onto legacy flows.

Gartner's 2026 Hype Cycle for Agentic AI and related CIO research emphasize that governance, security, and cost management must mature alongside capability. Only about 17% of organizations had deployed AI agents at the time of the survey, yet more than 60% expected to do so within two years—the most aggressive adoption curve among emerging technologies measured. Recommendations include guardian agents, runtime controls, distinct agent identities with human owners, escalation for high-risk actions, and the ability to contain runaway behavior. Weak governance is projected to produce material service, security, or cost incidents for a substantial share of organizations that scale without them. Full-scale multi-agent adoption remains years away for most enterprises because it requires overhauls of business processes, data foundations, workforces, partnerships, and security and governance policies.

Across these sources the validated pattern is consistent. Value appears when agents are grounded in private or tightly controlled first-party data and knowledge, when human oversight is designed into the operating model by risk rather than added later, when workflows are analyzed and redesigned before agents are applied, and when capacity is focused first on high-frequency routine work that already has clear handoffs and a measurable cost when it fails.

3. Why Most Organizations Cannot Execute This Path Internally

The binding constraint is the sociotechnical load documented by MIT Sloan and visible in McKinsey's and Gartner's findings. Mapping real workflows, grounding agents in accurate first-party knowledge, designing escalation and approval rules, building reliable connectors, evaluating outputs continuously, and maintaining the system all require sustained expert attention. Most organizations—particularly those operating on limited operational expense budgets—do not have people whose primary role is this work.



G2's October 2026 AI Agents Report of 1,019 B2B decision-makers documented the resulting supervision tax: 31% of organizations had staff spending six or more hours per week overseeing agents, and 45% reported that oversight and rework consumed more than a quarter of expected productivity gains. In the first month of an unmanaged deployment a manager can easily spend 1–2 hours per day catching mistakes before they reach a customer. At fully loaded rates of $75–$90 per hour that becomes $1,500–$3,600 before any durable capacity appears, and the checking often continues. G2 also found that 77% of organizations have already experienced negative fallout from an ungoverned agent.

For a solopreneur or small department this is decisive. The volume of any single routine process rarely justifies hiring a full-time specialist for agent training, evaluation, and governance. The choice becomes continued manual friction or an ungoverned tool that adds cost and risk without removing the original coordination or intake leak. This is the mechanism behind the high cancellation rates. Without the internal capacity to perform the work the research shows is required, projects stall after proof-of-concept or generate more supervision burden than capacity. McKinsey notes that data readiness remains a major obstacle; in banking surveys, 48% of executives identified it as their biggest barrier to AI adoption.

4. A Practical Sequence That Aligns with the Research

The sequence supported by the evidence is deliberate and narrow at the start.

First, perform workflow analysis to locate where volume, unclear ownership, and measurable cost of failure already exist. Typical high-gain areas include inbound call intake that is not answered live, handoffs that stall between steps, routine scheduling or lead qualification, and repeated synthesis or status work that currently requires copy-and-paste. soolisAI's DOSS approach provides one concrete version of this foundation: an adaptable operations layer that fits how the business already runs, improves visibility and clean data, and is established before any agentic capacity is introduced. The objective is sufficient structure that later agents have reliable context and defined handoffs, not a large-scale

system replacement. This matches the research finding that data readiness and workflow redesign are prerequisites rather than downstream problems.

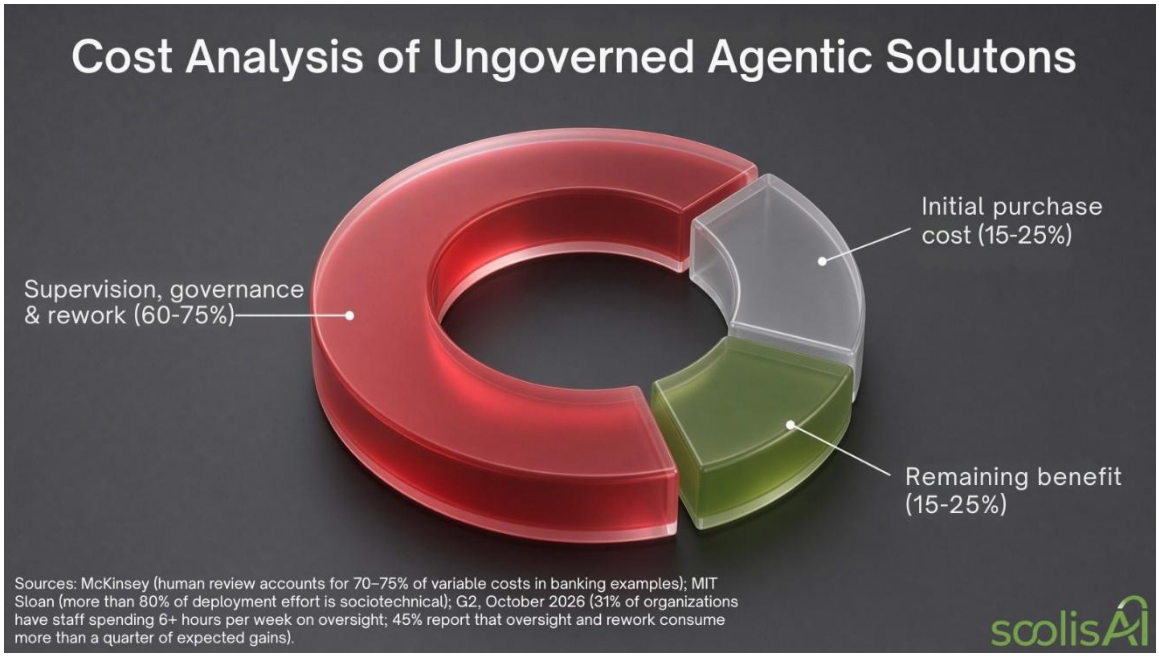
Only after that scoping, apply agentic capacity to the highest-gain areas. These are usually front-office or coordination tasks that consume disproportionate time relative to the judgment they require. Tiffany, soolisAI's smart front-office voice and intake specialist, is built for this layer: it answers on the existing business number, handles structured intake inside a first-party native data perimeter, and routes cleanly so work does not bounce between overloaded staff. Build Your Bench extends the same model to a small, scoped roster of virtual specialists that hand work to one another while remaining inside the organization's data perimeter. Specialists complete repetitive preparation; humans stay strictly above the loop for decisions that require judgment, using mechanisms such as 1-Click Human Approval. This aligns with McKinsey's description of deliberate human oversight and organization-owned enterprise context, and with MIT Sloan's emphasis on domain-expert validation and designed handoffs.

Throughout, private or first-party grounding is maintained. SOPs, pricing, client records, and institutional knowledge stay inside the protected boundary and are never exposed to public model training loops. Human oversight is designed from the outset rather than added after problems appear. Access checks and escalation rules are defined according to risk, consistent with the foundations McKinsey and Gartner describe as necessary for safe autonomy.

The Start Simple model operationalizes the sequence for organizations with limited budgets. A team can begin with a single specialist node for the single highest-friction role (often Tiffany for voice intake), prove the reduction in manual recovery and stalled handoffs, and expand the roster only after that first constraint is demonstrably lighter. Expert-led setup, training on the organization's own data, integration support, security monitoring, and ongoing performance optimization are included, so the organization does not have to staff the sociotechnical work itself. This directly addresses the readiness gaps Gartner documents and the 80% sociotechnical effort MIT Sloan identifies.

5. How Expert-Led AIaaS Makes the Validated Path Attainable on Limited Budgets

Because the majority of effort and ongoing cost sits in data grounding, validation, governance, and supervision rather than in the model call itself, an expert-led service changes the economics. The organization supplies its real workflows and first-party knowledge. The service performs the scoping analysis (DOSS-style), trains specialists on that private context, designs the escalation and approval rules, maintains connectors, and provides continuous evaluation. Client-side humans remain in the decision roles they already own. This mirrors the hybrid human-agent operating model McKinsey is building internally and the "managers of agents" shift MIT Sloan describes, without requiring the organization to hire and retain the specialists who perform the setup and oversight work.



For a solopreneur or 5- to 30-person team this converts the hybrid capacity pattern into practical daily relief without a six-figure custom engineering project or a new internal hire. Capacity is applied to work the business already has—unanswered calls, stalled handoffs, repeated status checks—so the benefit appears against existing leakage and time rather than against a speculative initiative. Organizations that follow the sequence also reduce their exposure to the cancellation rates Gartner projects, because scope, data perimeter, and oversight are established before volume is added. The same principles that support McKinsey's 10-month banking payback examples and internal hour savings become accessible at smaller scale when the sociotechnical load is absorbed by the expert-led layer.

Start Simple packaging further lowers the barrier. A first specialist can be stood up with a defined setup and modest monthly upkeep fee that covers ongoing skill and task-configuration maintenance, with usage billed separately. Expansion to a coordinated small roster occurs only after results are visible. This matches the research recommendation to apply agents where the greatest gains are likely first and to expand only after validation.

6. Realistic Expectations and Next Steps

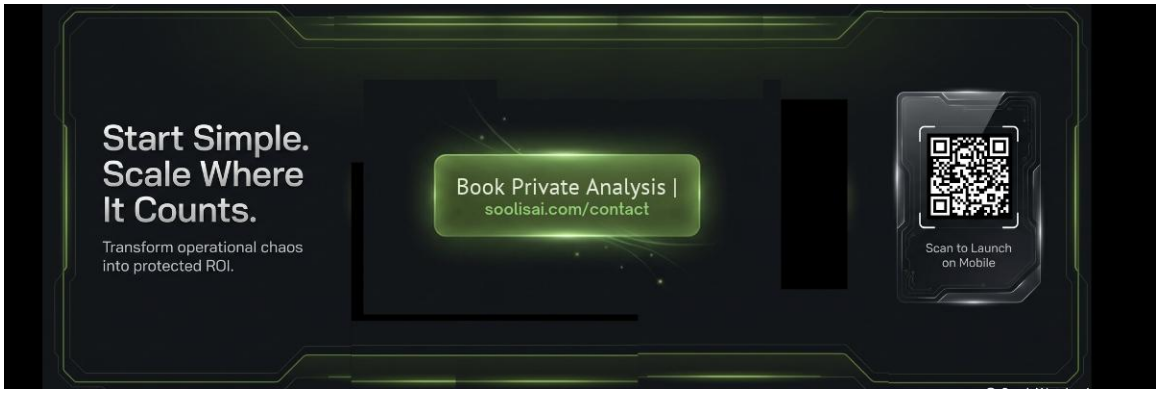
Full autonomous operation remains years away for most organizations, consistent with Gartner and Deloitte assessments of multi-year readiness requirements. Only a small minority currently have the governance and workforce redesign plans in place. The near-term, evidence-supported outcome is reliable expansion of capacity on routine work under explicit human control, grounded in private data, with clean handoffs and measurable reduction in recovery time. McKinsey's internal results and banking examples show what is possible once the foundations exist; the Gartner and MIT Sloan findings show why most organizations have not yet reached that point.

The practical starting point is a focused workflow analysis. That analysis identifies whether a single specialist (structured voice intake via Tiffany, for example) or a small coordinated roster is the highest-gain entry point, confirms the private-data and approval boundaries, and establishes the baseline against which results can be judged. Teams that complete this

step before applying agentic capacity are following the sequence the independent research associates with durable results rather than with the supervision taxes and cancellations that continue to characterize unmanaged experiments.

Organizations of any size can execute the path. Those without internal staff dedicated to agent training, evaluation, and governance simply require a partner that already performs that sociotechnical work inside a governed, first-party perimeter. That is the practical distinction between the approach validated by McKinsey's agentic coworker model and internal operations, Gartner's readiness data, and MIT Sloan's implementation research, and the experiments that continue to feed the 40% cancellation projection.

Next Step for Operational Leaders
If missed calls, stalled handoffs, or coordination overhead are already visible in a typical week, the next step is a scoped workflow analysis. Book a quick review at soolisAI to determine where the highest-gain starting point sits and how a Start Simple specialist or small bench returns that capacity—and the control—to the people who already own the operation.



Master Research Sources & EEAT Grounding Archive

McKinsey (2026). The rise of the agentic coworker: A banking CIO's guide to the next wave of AI; Building the foundations for agentic AI at scale; AI is changing work. Now it has to change the organization; Stacking the odds: A blueprint for successfully scaling agentic AI; banking analyses across 52 agentic use cases (approx. 10-month payback, ~2.9× over three years, contingent on redesign, data quality, and governance; human review 70–75% of variable costs in examples); internal operations (~40,000 humans + 20,000–25,000 AI agents; ~1.5 million hours saved on routine search/synthesis; ~2.5 million charts generated in six months; pairing goal within 18 months). CEO statements on combined workforce of ~60,000.

Gartner (2026). Hype Cycle for Agentic AI; CIO and Technology Executive surveys (~2,000 respondents: ~71% expect deployment within 12 months, ~21% with governance capabilities, ~4% with human-agent workforce redesign plans; only ~17% had deployed agents); projections that more than 40% of agentic AI projects will be cancelled by end of 2027 due to cost, value, and risk-control gaps; guidance on runtime governance, guardian agents, agent identities, and escalation.

MIT Sloan (2026). What effective AI leaders already know; 5 'heavy lifts' of deploying AI agents (>80% of effort sociotechnical; ~4 hours implementation per hour of model work);

The Emerging Agentic Enterprise (with BCG; survey of >2,000; more than one-third already deploying, 44% planning; manage as coworkers with dynamic risk-based oversight); related work on domain-expert validation and parallel management of agents (“managers of agents, not of people”).

Supporting operational benchmarks: Invoca (2026) lead conversion benchmarks (~70 million calls; ~56% live answer rate overall, 52% in home services); Quo (2026) missed-call analysis (~16.7 million calls; ~69% no callback within 48 hours; ~15% voicemail leave rate); McKinsey coordination estimates (35–60% of work time); 2026 manager survey (16.5 hours per week).

G2 (October 2026). AI Agents Report (1,019 B2B decision-makers; 31% spend 6+ hours weekly on oversight; 45% lose >25% of gains to review/rework; 77% experienced negative fallout from ungoverned agents).

soolisAI Research Group (2026). Operational synthesis aligning independent findings with Start Simple, DOSS workflow scoping, Tiffany front-office intake, Build Your Bench virtual roster, 1st-party native data perimeter, and 1-Click Human Approval.