STRATEGY & TECHNOLOGY

The Obelisk's Missing Layer

The infrastructure required to power smaller, AI-enabled consulting teams

By Garrett Amaru & Dennis Kubaile | Granulytix

Harvard Business Review recently described the consulting pyramid giving way to a leaner "obelisk" model — smaller teams, more AI enabled analysis, and human judgment at a premium. That framing gets the destination right. But it leaves a harder question unanswered: what operating infrastructure is needed to power this new consulting model?

THE PYRAMID IS BECOMING AN OBELISK — BUT NOT THE WAY FIRMS THINK

For decades, consulting ran on a simple economic engine: a wide base of junior analysts performing research, modeling, and synthesis; a narrow apex of senior partners selling work and guiding strategy; and billable leverage in between. The pyramid worked because analytical labor was expensive and scarce. AI collapses the economics that supported it.

Duncan, Anderson and Saviano's obelisk framing captures the structural consequence: as AI automates work that once justified thousands of junior hours, the pyramid's base collapses. Firms need fewer analysts, more judgment, and entirely new team structures. That much is right.

While the obelisk framing focuses on organizational implications, it does not adequately explore the systematized execution infrastructure required to operationalize the model at scale. A leaner team without encoded execution infrastructure is not an obelisk — it is simply a smaller pyramid with less leverage. The shape changes only when the engine underneath changes first.

That engine is what this article is about. The uncomfortable reality is that while most consulting firms advise clients on AI transformation, much of the industry still approaches AI internally as an augmentation layer rather than re-architecting the underlying execution infrastructure.

AI ACCESS IS TABLE STAKES — AND ALREADY BROADLY DEPLOYED

The numbers are striking. More than $12 billion has been committed to AI across the largest consulting firms since 2023, largely toward client-facing acceleration, AI service offerings, acquisitions, and broad employee enablement. But widespread deployment has not consistently translated into measurable enterprise value. McKinsey's 2025 State of AI survey — covering nearly 2,000 organizations — found 88% function-level AI adoption, yet only 6% reported material EBIT impact. A parallel Morgan Stanley/RSM survey found that 79% of enterprises had deployed Microsoft Copilot — and 74% could show no tangible business value from it.

> ***"A seat license is a utility, not a capability. Nothing accumulates. Every engagement starts cold."***

The likely culprit? Workflow redesign. In the same McKinsey survey, redesigning workflows was the strongest correlate of AI-driven EBIT impact, yet only 21% of GenAI-enabled organizations had redesigned their workflows. While correlation is not causation, the implication is difficult to ignore: AI access alone does not drive results. Operating-model change does.

Most firms have built what might be called the knowledge layer: AI-assisted research tools, internal document retrieval systems, and drafting assistants. McKinsey's Lilli — built on 100,000 documents and reporting broad firmwide adoption — is the best

public example. These systems are real, valuable, and increasingly table stakes.

But knowledge layers are not executable delivery systems. A prompt is consumable. An output is produced, the interaction ends, and little accumulates beyond the immediate task. The next engagement starts cold.

Competitive advantage requires something built on top of foundational AI access — governed workflows that retain context, enforce methodology, validate outputs, and improve through repeated use.

THE OPERATING LAYER: WHAT ACTUALLY POWERS THE OBELISK

The operating layer is workflow-execution infrastructure: AI embedded into the firm's analytical methodology end-to-end, with the firm's IP codified, quality standards enforced, and outputs auditable. Orchestration ties these components into a single executable workflow. This is what allows smaller teams to produce work that previously required much larger analytical support structures.

Four components define it:

Methodology encoding. Most consulting IP lives in the heads of senior practitioners and often leaves with them. An operating layer encodes methodology directly into execution logic — required analyses, sequencing logic, drilling criteria, validation checks, and escalation rules. The firm's analytical approach becomes a durable asset rather than retained institutional memory.

Deterministic computation. AI is exceptionally strong at synthesis, reasoning, and interpretation. But a language model alone is not a calculation engine. In client work, arithmetic cannot be approximate. A workflow layer routes quantitative calculations through coded computation engines while reserving AI for interpretation, prioritization, and narrative synthesis.

The distinction matters for both accuracy and liability.

Validation gates. Work product cannot advance when defined checks fail — numerical reconciliation, citation integrity, or editorial criteria. This is not a nice-to-have. Stanford's HAI found that leading legal AI tools hallucinate in roughly one in six queries. Without architectural validation, those failures can reach client deliverables.

Validation also has to be embedded directly into the execution architecture itself. It cannot be cleanly retrofitted later.

Persistent state and auditability. Today, most AI-assisted consulting work remains trapped inside individual chat threads — context lost and institutional knowledge trapped in one consultant's browser history.

An execution layer changes that dynamic.

For example, a workflow may identify pricing variance as a likely margin driver during an initial scan. Deterministic computation handles the pricing and margin calculations, validation gates confirm reconciliation and source integrity, and a human reviewer approves the finding before it advances. The supporting evidence, rejected hypotheses, and validated conclusions persist as institutional context within the engagement so another team member can extend the analysis into elasticity modeling or commercial actions without restarting the investigation from scratch.

Critically, the workflow also produces traceability by design. Inputs used, methodology applied, validation steps passed, and intermediate outputs are recorded as part of the workflow execution itself rather than reconstructed after the fact.

Taken together, these elements are what allow consulting delivery to scale and make the obelisk model operationally viable.

THE IRONY: WHY FIRMS SELLING TRANSFORMATION HAVEN'T TRANSFORMED

If this workflow infrastructure is so valuable, why have most firms not built it? The answer is a capability constraint — and a structural tension within the consulting model itself.

Building execution infrastructure requires three disciplines that rarely coexist: engineering depth — the ability to build and operate production-grade systems; AI architecture experience — knowing when to use models, deterministic computation, validation, and orchestration; and methodology context — the ability to translate the firm's analytical approach into executable logic.

But the challenge is not simply talent availability. The infrastructure required to make the obelisk work also compresses the leverage economics that historically powered the pyramid. Workflow-execution systems institutionalize capabilities that once depended on layers of junior analytical labor and apprenticeship-based knowledge transfer.

Incumbents rarely embrace changes that disrupt the economics that made them successful. This is the irony the industry increasingly faces. Consulting firms correctly advise clients that AI access alone does not create transformation —operating-model change does. Yet few have applied that same logic internally.

THE WINDOW IS CLOSING

Several signals suggest this transition may move faster than many firms suspect.

Phil Fersht of HFS Research characterized early 2026 as "the last 18 months of labor-intensive services." Gartner forecasts the cost-to-value gap for process-centric service contracts will be reduced by at least 50% by 2027 through agentic AI reinvention. The EU AI Act also signals where enterprise expectations are moving: toward greater auditability, governance, validation, and traceability for AI-enabled work.

The strategic implication is straightforward: workflow infrastructure compounds. Every engagement improves the system for the next one. First movers widen the advantage over time, forcing followers into perpetual catch-up.

WHAT TO DO: FOUR PRINCIPLES

1. Build above the foundation. Agent Skills and related platform primitives are making parts of the AI foundation layer increasingly portable across providers. Firms should not rebuild commodity infrastructure. Buy it. The real source of differentiation is above the platform layer: firm-specific methodology, workflow design, validation logic, and consulting-specific execution. Build it.

2. Create specialized workflows. General-purpose AI alone is not the solution. Consulting work is highly structured, with phases, methodology, review patterns, and validation criteria that vary across firms and engagement types. The workflow layer operationalizes that structure, institutionalizing firm IP – the one layer no vendor can supply – to deliver consistent outputs regardless of team capabilities.

3. Assemble the three-skill team. The challenge is finding the small group capable of combining all three disciplines: engineering depth, AI architecture, and consulting methodology. Generic AI hiring alone does not create the necessary integrated capability.

4. Validate from day one. Validation gates cannot be cleanly bolted onto workflows as an afterthought. Numerical reconciliation, citation checks, reviewer patterns, and auditability have to be designed into the architecture from the beginning. Building validation in from the start is far cheaper than retrofitting it later, and both costs are small compared with the reputational damage of letting unverified work reach client deliverables.

The obelisk is not simply a metaphor for doing more with less. It describes what consulting delivery looks like when methodology becomes executable, validation becomes architectural, and institutional capability compounds across engagements. Know-how transfers from individuals into a true firm asset — precisely the kind of organizational disruption consultants routinely advise clients to navigate.

HBR described the destination. This is the road.

About the authors

Garrett Amaru and Dennis Kubaile are former strategy consultants and co-founders of Granulytix, which develops AI-enabled analytical workflows that accelerate insight and improve consulting delivery.