

From Demo to Dependable Capability

Fall 2026 · Terry College of Business · University of Georgia

Course: EMBA-7450, Information Technology and Strategy **Module role:** Strategic-architecture module. The module gives executives a stack-and-sourcing lens for diagnosing AI initiatives, deciding what to own and what to rent, and connecting technology choices to value, risk, cost, and durability. It builds on Module 2's executive AI literacy and the authority ladder, and hands forward to Module 4's risk and governance work.

Module Learning Objective

By the end of this module, you will be able to diagnose an AI initiative through five layers — workflow, context, model portfolio, agents, and assurance — match authority to value and risk, make one consequential sourcing decision, and explain what evidence would turn a demo into a dependable capability.

If You Want to Know More

The Pre-Class Read below carries the argument. The materials listed here are reference reading — not assigned, not required, and not assumed in class. We list them for executives who want to read further on a specific topic before, during, or after the course:

- Stanford HAI, *2026 AI Index Report* — economy and technical-performance highlights.¹
- McKinsey, *The State of AI in 2025: Agents, Innovation, and Transformation*.²
- Gartner, *Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027* (June 25, 2025).³
- A current vendor demo or RFP from your own organization, read through the five-layer stack model below.

Part 1: Pre-Class Read

BOTTOM LINE. The **AI stack** is not a technical diagram owned by the CIO; it is the operating foundation of AI strategy. A demo proves what an AI system can do under selected conditions. A capability proves what your organization can operate in production — with grounded answers, controlled authority, predictable cost, and durable assurance. We argue that durable advantage in 2026 sits in workflow design, trusted context, disciplined model portfolios, harness engineering, and production assurance — not in model choice alone. Sourcing is not a procurement decision after strategy. Sourcing is strategy.

WHAT YOU WILL PRODUCE. Before class, bring a Foundation/Core sketch of one AI initiative: the workflow and intended outcome, authoritative context or data, authority level, weakest or most uncertain stack layer, one consequential sourcing decision, and one unresolved executive question. The optional Challenge/Extension adds the complete five-layer map, layer-by-layer sourcing logic, assurance and production metrics, unit economics, and a full steering-group recommendation. These are differences in preparation depth, not separate graded pathways or new assignments. They do not change Blog 3 or the Transform-path policy. In class, the boardroom drill applies the same stack and authority ladder. The work then hands forward to Modules 4, 5, and 7.

TERMS. Selected glossary terms appear in **bold** at their first explanatory use in the Part 1 running text. The alphabetized Executive Glossary at the beginning of Part 2 includes only terms used in this Pre-Class Read that Modules 1 and 2 did not already define. Use it as a lookup aid, not assigned vocabulary to memorize.

Opening Vignette: Goldman Sachs and the Discipline of Verifiability

Goldman Sachs deployed GS AI Assistant — an internally built generative-AI platform routed across multiple model providers — to roughly 10,000 employees in early 2025 and extended access firmwide to approximately 46,000 employees by June 2025.⁴ The platform is built. The model partners are named. The rollout numbers are public. None of that is the strategic story.

Marco Argenti, Chief Information Officer of Goldman Sachs, framed the binding constraint at Goldman's Communacopia + Technology conference in October 2025: AI spreads fastest, he argued, where teams can make work "*machine-verifiable*" — where approval chains, review steps, and documentation are explicit enough that an AI system can slot into one of the three roles in any approval process: the person doing the work, the person reviewing it, and the person signing off (*Marco Argenti, Chief Information Officer, Goldman Sachs, Communacopia + Technology, October 2025*).⁷ In a *Fox Business* interview in December 2025, Argenti described 2026 as the year of personal agents at Goldman.⁵ The framing matters more than the platform numbers. The model is not the constraint. The workflow's verifiability is.

The interesting move is the discipline that decides where AI gets to act first. Goldman did not deploy AI broadly and then ask which workflows could absorb it. The firm asked which workflows already produced the kind of explicit approval-chain evidence an AI system could slot into — and where the stack would have to change

before a workflow could safely accept an AI participant. Vendor partners were selected to specific roles, models were routed, and authority was bounded around the answers.

This is what an AI capability looks like when the stack is doing its job. The demo would have shown a model producing a plausible answer. The capability shows the firm in which workflows that answer is verifiable, by whom, against which approval chain, with which audit trail.

KEY LESSON. A demo proves what an AI system can do. A capability proves what your organization can verify.

The Stack Is the Strategy

Perhaps the most important overarching insight from the past three years of enterprise AI deployment is that the AI stack is the operating foundation of strategy — not an implementation detail to delegate downward. We see organizations that pair model access with disciplined workflow design, trusted context, harness engineering, and production assurance pull steadily ahead of organizations that treat the stack as someone else's problem. The model is rarely the bottleneck. The system around the model usually is.

This flips the conventional vendor conversation. The demo asks the buyer to evaluate the model's output. The strategic question sits one layer down: which layers of the stack will your organization own, which will you rent, which should you route across, and which must remain portable? An organization that answers those four questions with discipline tends to build durable AI capability. An organization that buys a polished demo and asks the questions later tends to discover, eighteen months in, that the vendor controls the workflow, the context, the model path, and the **data exhaust** — and that the strategic core has been outsourced.

The rest of this Pre-Class Read gives you a way to ask those questions.

The AI Stack as a Strategic Supply Chain

The phrase "AI stack" can sound intimidating. It suggests layers of technology that only specialists can understand. We find it more useful to think about the stack as a supply chain for delivering intelligence into a workflow. In a traditional supply chain, raw materials are sourced, refined, transported, assembled, quality-checked, and delivered to the customer. A weakness anywhere in the chain undermines the final product. The same is true of AI.

The raw material is not data in the abstract; it is *trusted context* — the documents, records, histories, product information, customer information, policies, and operational signals that allow an AI system to produce a useful answer or take a useful action. The factory is not a single model; it is a portfolio of models, routing decisions, prompts, instructions, tool calls, workflow logic, and human review points. The quality system is not a governance policy document; it includes evals, monitoring, audit trails, incident handling, cost controls, and clear ownership. The harness is the operating scaffolding that connects all of this to real work. The retail shelf is the business workflow or application where employees, customers, suppliers, or partners actually experience the capability.

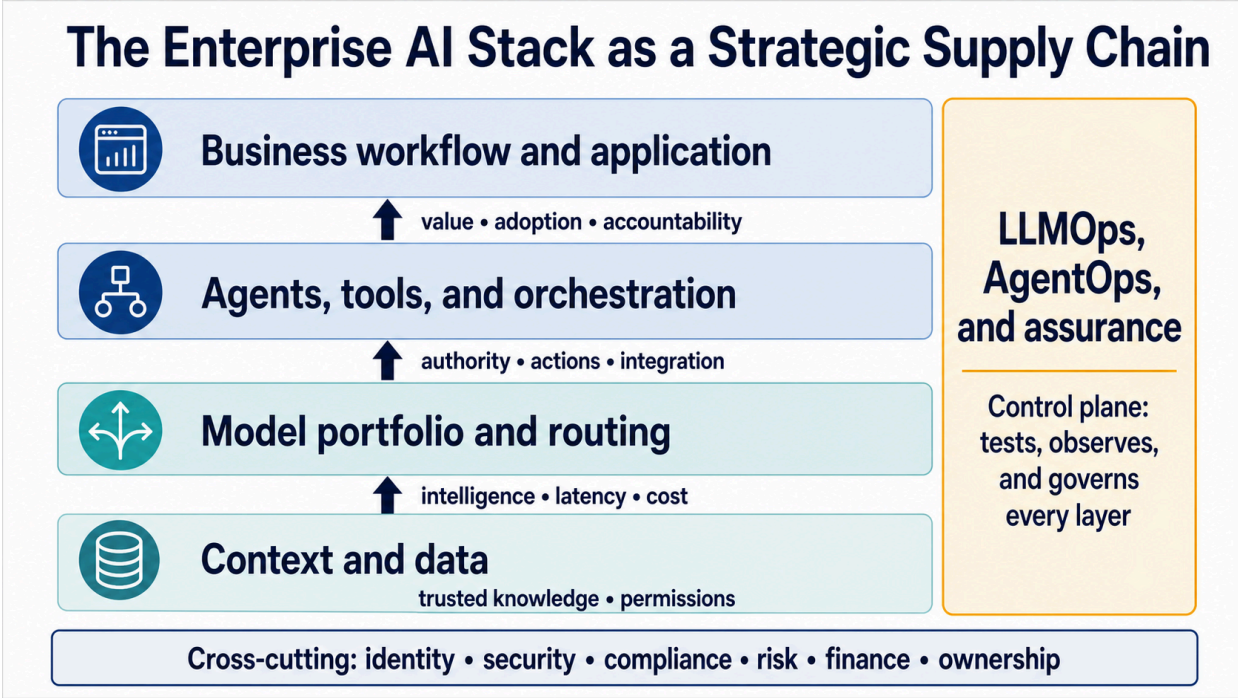
Keep three terms separate. The **engine** is the underlying model. The harness supplies context, tools, permissions, memory, evaluation, and workflow controls. The **interface** is where a person interacts with the system. The executive question is which layer creates value, control, or switching cost.

The point of the stack model is not to make executives memorize infrastructure categories. The point is to make dependencies visible. If the workflow is unclear, model choice will be arbitrary. If the context is poor, the most capable model will still answer from weak information. If the agent has too much authority, the organization automates its mistakes. If evals and observability are missing, the organization may not know when quality degrades. If cost is not designed in, a successful pilot becomes an unaffordable production service. If sourcing choices are made casually, the organization gives away the very part of the system that could have become a source of advantage.

The five layers below are best read as the operational unfolding of the harness layer Module 2 introduced in the Three-Layer Model. Module 2 carved AI work into prompt, context, and harness — the assignment, the briefing packet, and the workplace. Module 3 takes the workplace and shows what it is made of. The first three of Module 3's layers (workflow, context, model) sit close to the prompt and context layers of the earlier model. The fourth and fifth (agents and orchestration; LLMOps and assurance) are where harness engineering becomes architectural concern. The vocabulary differs because the audiences differ — Module 2 spoke to the executive's own desk; Module 3 speaks to the executive's enterprise — but the conceptual through-line is intact.

For 2026, we use a five-layer executive model. Earlier versions of this course used a four-layer model; we add a fifth layer — LLMOps, AgentOps, and assurance — because production discipline has become its own architectural concern, distinct from the agent and orchestration layer above it.⁶ The first layer is the business workflow and application. The second is context and data. The third is the model portfolio and routing. The fourth is agents, tools, and orchestration. The fifth is LLMOps, AgentOps, and assurance. Identity, security, governance, compliance, and AI FinOps cut across all five.

Figure 3.1



The right-hand band is deliberately drawn alongside the four delivery layers, not underneath them. LLMOps, AgentOps, and assurance is not the foundation the other four sit on — it is the discipline that tests, observes, and governs all four while they operate. A stack can have every delivery layer built well and still fail in production if that control plane is thin.

The Field Reference contains a Stack Layer → Executive Question → Strategic Risk table for use as a diagnostic tool. The short version of the model, for use as you read on, is: each layer answers a specific executive question, and weakness at any layer creates a specific strategic failure mode.

One customer-service workflow through all five layers

Use one recurring example to keep the architecture concrete. Imagine a customer-service capability intended to resolve routine billing questions faster without increasing incorrect credits or customer complaints.

Layer	Executive decision in the example
Workflow and application	Decide whether the system answers questions, drafts replies, recommends credits, or issues bounded credits. Each choice changes the intended outcome and the user's role.
Context and data	Identify the authoritative billing policy, account record, product terms, regional exceptions, and permissions. Decide what happens when two sources conflict.
Model portfolio and routing	Use task-specific evidence to decide which work needs deeper reasoning and which work can use a faster, lower-cost model. Define a fallback when confidence or source quality is low.
Agents, tools, and orchestration	Set the authority level. A drafting assistant needs read access and human review. A workflow agent that prepares a credit also needs scoped tool access, an approval gate, and an audit trail.
LLMOps, AgentOps, and assurance	Test source use, answer quality, tool behavior, escalation, and recovery. Monitor cost per correctly resolved case, correction rate, complaint rate, and unsupported claims.

This is one initiative, not five separate technology projects. The same visible interface could sit on top of any of these designs. That is why a demo is weak evidence. It may show a fluent answer while hiding stale policy, broad permissions, an expensive model route, missing approvals, or no way to reconstruct a disputed action. The steering question is not whether the chatbot looks capable. It is whether the complete stack can deliver the intended outcome within the organization's authority, evidence, cost, and risk limits.

Layer One: Business Workflow and Application

The first layer is where the AI capability is actually experienced — as a customer-facing chatbot, a sales copilot, an internal research assistant, a claims-processing agent, a coding assistant, a marketing content tool, a procurement analyst, a field-service guide, or a finance planning copilot. From a strategic perspective, this is the most important layer. It is where value is either created or lost.

AI strategy should begin here, not with the model. Ask: who is the user, what job are they trying to do, what decision or action is being improved, and what business outcome should change? Is the point to reduce cycle time, improve quality, reduce cost, personalize service, increase revenue, reduce risk, expand human capacity, or create a new product experience? If the business outcome is vague, the stack will be vague. If the workflow is not understood, the AI system will tend to become an impressive side tool rather than a core capability.

The Argenti framing in the opener — that AI spreads fastest where approval chains are explicit enough for an AI system to slot into the do/review/sign-off triad — applies even when an organization does not look like Goldman. If an executive cannot say which of those three roles the AI is meant to occupy, the stack is being designed in the dark.

Consider customer service. A company might say it wants an AI agent for customer service. That phrase hides several different strategies. One version answers common questions using approved knowledge articles. Another drafts suggested replies for human agents. Another triages issues, assigns priority, and recommends next actions. Another prepares refunds for approval. Another negotiates retention offers. These may all be called "AI customer service," but they do not require the same stack — the same context, the same models, the same permissions, the same human oversight, or the same risk controls. The Field Reference contains a Customer Service Use Case → Authority Level table that walks the gradient.

The workflow layer is also where adoption either holds or breaks. A technically strong AI capability that does not fit the user's real workflow will not produce value. Employees may ignore it, work around it, or use it in ways the designers did not intend. Customers may abandon it if it slows them down or gives answers they do not trust. Managers may resist it if accountability becomes unclear. Module 7 returns to operating model and change at length; here, the point is architectural. Workflow design determines stack requirements, not the other way around.

Design backward from value. Do not begin with "we have access to a powerful model." Begin with "this is the workflow where we need a new capability."

BOARD QUESTION. *"For each AI initiative we are funding, can we name the workflow, the user, the decision being improved, and the business outcome that should move? If we cannot, why are we funding it?"*

Layer Two: Context and Data

The second layer is context and data. No layer more often decides quietly whether an initiative becomes dependable. A **context system** is the combination of approved knowledge, retrieval methods, metadata, permissions, and quality controls that gives an AI system the right information at the right time. The executive question is simple: *What must the system know, and is it allowed to know it?*

A context system can draw on documents, databases, CRM records, ERP data, service tickets, product catalogs, contracts, policies, call transcripts, emails, and external sources. Retrieval may use keyword search, vector search, hybrid methods, or knowledge graphs. Metadata makes the material governable: who owns each source, when it was updated, where it applies, who may see it, and what supersedes it.

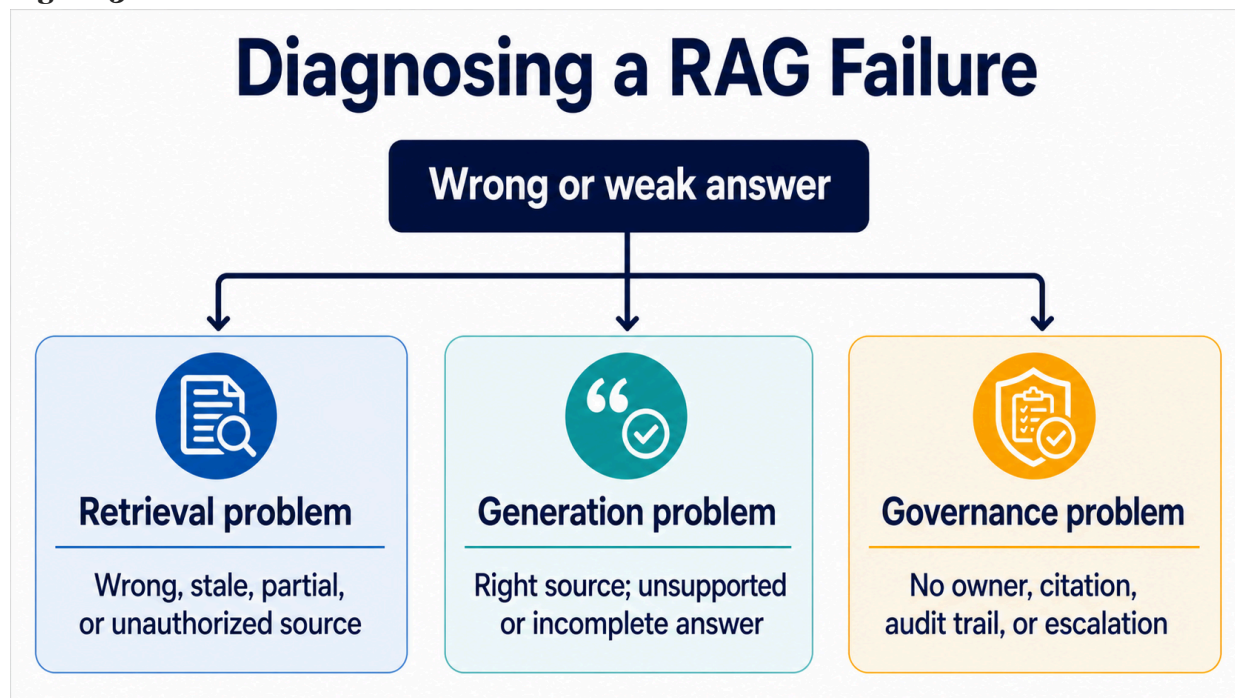
The useful executive frame is context engineering: deliberately writing, selecting, compressing, isolating, and caching information for the task. You do not need to memorize those operations. Use them to diagnose failure. Missing information points to selection. Too much irrelevant material points to compression or isolation. Stale information points to writing and freshness ownership. Repeatedly paying to load the same stable material points to caching. The detailed reference tables are in Part 2.

This is where many AI initiatives quietly succeed or fail. The model may be excellent, but it can only answer from what it knows or what it can retrieve. If the system retrieves an outdated policy, the model may confidently summarize the wrong answer. If source documents contradict each other, the model may blend them in a way that sounds reasonable but is operationally dangerous. If permissions are not respected, the AI may reveal sensitive information to the wrong employee. If documents are poorly structured, the AI may miss the relevant passage. If content owners are not responsible for freshness, the system degrades without anyone noticing.

Retrieval-augmented generation (RAG) is one common pattern for connecting a model to current sources. A question triggers retrieval from an approved knowledge base, and the model uses the retrieved material to answer. RAG avoids retraining a model whenever enterprise information changes, but it does not solve ownership, freshness, permissions, contradictions, or citation quality.

The deeper executive issue is trust. Which sources are authoritative? Who owns them? Are access rules enforced? What happens when sources disagree? Can users inspect citations? Can retrieval quality be tested separately from answer quality? Microsoft's RAG-evaluator documentation distinguishes retrieval, groundedness, relevance, and completeness.¹⁰ That matters because *"the answer was wrong"* may mean the system retrieved the wrong source, ignored a critical exception, or generated a claim the source did not support. Each category needs a different repair. A wrong answer may also expose a governance failure: wrong permissions, unowned freshness, or no audit trail.

Figure 3.2



Consider an internal HR policy assistant. During the pilot, it performs well on common questions about PTO and benefits. In production, employees start asking about leave eligibility, accommodation requests, manager obligations, and regional exceptions. Problems emerge. The system retrieves an older policy because that document appears more often in the corpus. It answers differently for similar employees because regional

metadata is missing. It cites a restricted document to someone who should not know that document exists. HR cannot reconstruct which policy source informed a disputed answer. In this case, the model did not fail dramatically. The context system failed quietly.

Context is often where durable advantage lives. Models are increasingly available through many providers. Proprietary context, proprietary workflow knowledge, and the organizational ability to maintain trusted context are harder to copy. An organization that treats its context layer as a strategic asset is making a different kind of AI investment than an organization that treats RAG as a tool purchase.

The takeaway: do not ask only whether your organization has a vector database. Ask whether you have a trusted, permission-aware, testable knowledge system — and whether you, not a vendor, own its content, freshness, and governance.

BOARD QUESTION. *"Does our context system have a knowledge-management discipline with AI attached — and can we name who owns source authority, freshness, permissions, and deletion?"*

Layer Three: Model Portfolio and Routing

The third layer is the model and routing layer. In earlier framings, the central model question was proprietary versus open-source, or renting versus owning. That distinction is still useful, but it is too binary for 2026. The better frame is **model portfolio**.

Organizations now have access to a rapidly changing set of model options. They can use frontier proprietary models through APIs. They can use smaller proprietary models optimized for speed and cost. They can use open-weight models that provide more control, customization, and hosting flexibility. They can use specialized models for code, image, speech, embeddings, classification, extraction, or search. They can use small language models tuned for edge or sensitive deployments. They can use model gateways and routing layers that choose different models for different tasks. The strategic question is not "which model is best?" It is "which model should handle which task, under which constraints?"

This shift matters because model capabilities, pricing, latency, and availability change quickly. A model that is best for a task in January may not be best in October. A frontier model may be necessary for ambiguous reasoning, complex synthesis, or high-value judgment support, but it may be wasteful for simple classification or extraction. A smaller model may be good enough for routine steps, faster for users, cheaper at scale, and easier to host in a controlled environment. An open-weight model may be attractive when data sovereignty, customization, or portability matters. A managed API may be preferable when speed and convenience matter more than deep control.

The economics reinforce the portfolio logic. Model prices and capabilities change quickly, while the costs of context, harnesses, evaluation, monitoring, and human review remain. Model routing means sending different tasks to different models based on quality, cost, latency, risk, or geography. The executive question is not which model wins a general benchmark. It is which route produces an acceptable result for this workflow at an acceptable total cost.

For executives, this creates a new set of decision criteria. Quality should be measured on your organization's task, not only on public benchmarks. Cost should be measured per successful task, not merely per **token**. Latency should be understood from the user's workflow, because a model that is impressive but slow may not be adopted. Privacy and data boundaries matter because some use cases can tolerate third-party processing while others cannot. Portability matters because deep integration with one vendor can become a strategic dependency. Risk matters because drafting a marketing paragraph is not the same as recommending a medical action, approving credit, or interpreting a contract.

The Field Reference contains a Model Strategy Tradeoffs table covering frontier proprietary, smaller proprietary, open-weight, specialized, model routing, and model abstraction layer choices, with a watch-out and an executive metric for each.

A powerful model can mask weak architecture for a while. It may compensate for unclear instructions, messy context, or poor workflow design during the pilot. But using the most capable model for every task can create unnecessary cost, latency, and dependence. A disciplined organization allocates intelligence. It uses the right model for the job, monitors performance, and keeps the ability to adapt as the market changes.

Model choice is no longer a one-time technology selection. We treat it as an ongoing portfolio management problem.

BOARD QUESTION. *"Where are we paying for intelligence we do not need, and where are we underinvesting in intelligence the workflow actually requires?"*

Layer Four: Agents, Tools, and Orchestration

The fourth layer is agents, tools, and orchestration. This is where harness engineering becomes visible. A chatbot produces text. An assistant helps a user complete a task. A copilot works alongside a person in a specific application. An agent can pursue a goal across steps, retrieve context, use tools, call APIs, remember state, make decisions, and sometimes take action.

An agent is a system that combines a model, instructions, and tools to pursue a goal and sometimes take action.¹³ It is not just a model. It may retrieve data, retain state, call tools, ask for approval, or hand work to another agent. In practitioner shorthand, *Agent = Model + Harness*. The executive question is: *What can this system do, and where must it stop?* The detailed Harness Component map is in Part 2.

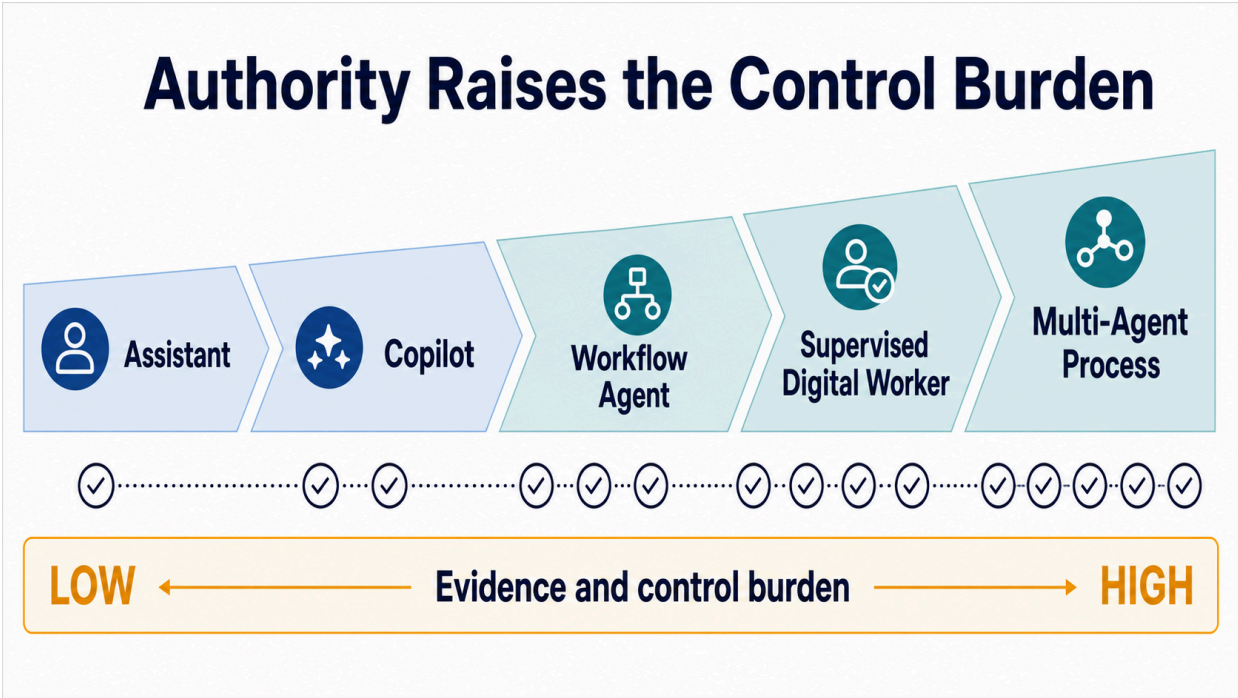
The map shows why a vendor demo can be misleading: the visible interface may be attractive while the harness underneath is weak, opaque, or economically fragile. It also changes the risk conversation. When an AI system only generates text, the main question is whether the output is useful, accurate, safe, and appropriate. When an AI system can use tools, the question becomes what it can *do*. Can it send an email? Update a record? Change a price? Issue a refund? Approve a workflow? Create code? Call another agent? Access a restricted system? Continue acting after the user has left?

The same model can be low risk in one architecture and high risk in another. A model that drafts a response for human approval may be a useful productivity tool. The same model connected to customer accounts, refund authority, and outbound email becomes a different managerial object. The issue is not whether agents are good or bad. The issue is *authority*.

Module 2 introduced the course authority ladder as an executive delegation tool. The same ladder shows what the stack must support as authority increases. The Field Reference contains the full Authority Ladder Control Burden table; the short version for use here:

Level	Course Term	What the Stack Must Support
1	Assistant	Drafts, summarizes, or analyzes only
2	Copilot	Recommends actions inside a human-led workflow
3	Workflow Agent	Completes bounded tasks in approved systems
4	Supervised Digital Worker	Acts across systems within defined thresholds
5	Multi-Agent Process	Coordinates multiple agents across an end-to-end process segment

Figure 3.3



Autonomy is not a technical trophy. We expect many valuable enterprise AI systems to remain at Levels 2, 3, or 4 for the foreseeable future, and the executive job is to match authority to value, risk, evidence, and organizational readiness.

Agents need tools and data sources. Open standards such as the **Model Context Protocol (MCP)** can reduce one-off integration work.¹⁶ They do not remove the need for authentication, least-privilege access, approval gates, and logging. When an agent acts, it must act as an identifiable and revocable principal — not as an invisible extension of a user's permissions. The specialist identity and connector references are in Part 2.

Computer-use and browser agents need special caution. Instead of calling an API, they read screens and click or type as a person would. That makes old systems accessible, but it also exposes the agent to untrusted instructions embedded in websites, documents, and messages. Official vendor safety guidance treats **prompt injection** as an ongoing security challenge, recommends narrow access and human review, and notes that reliability remains task-dependent.⁵³ Published benchmarks for these agents are not yet consistent enough to support confident cross-product performance claims. Treat these agents as supervised augmentation until your own evaluations support more authority.

The takeaway: when AI can act, the central question changes. Ask not only *"what can the model say?"* but *"what can the system do, under whose authority, with what context, through what harness, with what evidence, and with what audit trail?"*

Layer Five: LLMOps, AgentOps, and Assurance

The fifth layer turns harness engineering into managed capability. It includes evaluation, monitoring, tracing, versioning, logging, feedback loops, incident response, cost management, and lifecycle ownership.

Traditional software can often be tested against deterministic outputs. AI systems are probabilistic. Agent systems add complexity because they may retrieve context, select tools, route to different models, retry after failure, loop through multiple steps, hand off to another agent, or ask for human approval. This does not mean treating unpredictability as unmanageable. It means your organization needs a different production control panel — one that covers quality, grounding, retrieval, reliability, agent behavior, safety, security, cost, adoption, and business value. The Field Reference contains the full Production Control Panel and Eval Type tables.

This is where evals become central. An **eval** is a structured test of whether the AI system performs acceptably on the tasks that matter. You do not need to write eval code, but you do need to require eval discipline. A useful eval set might include representative work the system must handle, edge cases, regulatory scenarios, adversarial prompts, ambiguous requests, incomplete data, and examples that test whether the system can say *"I do not know."* For a RAG system, evals should test retrieval and groundedness separately. For an agent, evals should test tool use, approval behavior, escalation, and failure handling. For a business workflow, evals should connect AI performance to cycle time, quality, risk, cost, or revenue.

The question is not simply *"what is the model's accuracy?"* That question is often too narrow. Better questions: what are the evals for this workflow, who owns them, how often are they run, what happens when performance declines, and what threshold must be met before the system gets more authority?

Observability is the companion discipline: the ability to reconstruct what the system saw and did. If an AI workflow fails, leaders should be able to identify the user's request, retrieved context, model route, tool calls, retries, approvals, output, and cost. OpenTelemetry and vendor monitoring tools increasingly cover those traces.²⁰²¹²² The products will change. The executive requirement will not: consequential AI work must be inspectable.

Assurance also includes red-teaming, shadow-AI visibility, incident response, and **AI FinOps**. AI FinOps means managing the full cost of an AI workflow, not merely its model bill. Context length, caching, routing, retries, tool calls, storage, infrastructure, human cleanup, and adoption can all change unit economics.²³²⁴²⁵

Ask for cost per successful task: per resolved case, accepted draft, completed analysis, or avoided escalation. A pilot can look inexpensive when usage is low and supervision is intense. A successful rollout can become expensive precisely because people rely on it. The cheapest model can still be expensive if it fails often, retries, needs human cleanup, or slows the workflow.

The takeaway: production AI needs assurance. Without evals, observability, and AI FinOps, you are flying on anecdotes.

BOARD QUESTION. *"What is the cost per successful task — not merely per token — for each production AI workflow?"*

Build, Buy, Partner, Route, or Abstract

The old build/buy/partner matrix remains useful, but it needs two additions and needs to be applied more precisely. *Route* means directing different tasks to different vendors or models rather than committing to one — the sourcing equivalent of a diversified portfolio. *Abstract* means inserting a layer of your own (a gateway, an API wrapper, an internal platform) between your organization and the underlying vendor, so you can swap what sits behind it without rebuilding the workflows in front of it. *"Should we build or buy AI?"* is too broad to be a good question. AI capabilities are layered. An organization might buy a SaaS interface, own the workflow design, use a managed model API, route across multiple models, build its own context system, partner for implementation, and standardize observability centrally. That is not indecision. It is a realistic enterprise architecture.

The default executive bias should vary by layer. Your organization should usually own the workflow design, because that is where adoption and differentiation live. You should own the trusted-context strategy and context-quality standards, because proprietary context is often a source of durable advantage. You should treat models as a portfolio, because performance and price change quickly. You should buy common agents where the workflow is standard, but build or partner for strategic workflows where proprietary process knowledge matters. You should standardize LLMOps, AgentOps, and assurance enough to create consistent assurance across the enterprise, while allowing business units to extend patterns for local needs. The Field Reference contains the full Build/Buy/Partner/Route/Abstract default-bias table by layer.

Astris AI is a clean illustration of "build to retain IP and operating control." Lockheed Martin did not need a third-party AI platform to deliver internal productivity tools. It built one. When the internal platform proved useful enough, Lockheed externalized it as a subsidiary aimed at defense industrial-base customers and high-assurance commercial buyers, rather than selling the IP to a hyperscaler or licensing it to a vendor.¹⁵ The strategic logic: in a regulated, sovereign-data context, the operating discipline that produces trustworthy AI is itself the asset, and outsourcing it gives away the wrong thing. The same logic does not apply uniformly across industries. It applies wherever proprietary process knowledge, regulated data, sovereign deployment requirements, or high-assurance operating standards make the operating capability — not the model — the durable advantage.

The danger is not buying. Buying is often the right answer. The danger is accidentally buying the *strategic core*. If a vendor controls the workflow, the context, the model, the user experience, the data exhaust, and the roadmap, your organization may have outsourced the very learning loop that could have become a competitive advantage.

Sourcing is not procurement after strategy. Sourcing is strategy.

BOARD QUESTION. *"Which layers must we own, which can we rent, which should we route across, and which must remain portable?"*

A Final Word Before Class

We have given you five layers, an authority ladder, a sourcing framework, and a vendor diligence discipline. These are tools. They matter. They are also not the point.

The point is a shift in mindset. The AI stack is the operating foundation of enterprise AI strategy. The vendor demo asks a narrow question — does the model produce a plausible answer? The strategic question is broader and harder. Which of these layers does your organization actually own? Which are you renting on terms you can defend? Which are you routing across, deliberately, because no single vendor should hold them all? Which must remain portable, because the market will move faster than any one provider's roadmap?

Executives who think about AI as a model selection problem tend to spend the next three years discovering that the model was the easy part. Executives who think about AI as a stack-and-sourcing problem tend to build durable capability. We want you in the second group. The rest of this module — and the four modules that follow — are designed to help you stay there.

Bring This to Class

Use one in-flight or proposed AI initiative. Choose the depth that fits your current experience and access. Everyone works from the same concepts, and everyone can contribute to the boardroom discussion.

Foundation/Core — required preparation

Bring six decisions:

1. **Business workflow and intended outcome.** Name the user, the work, and the business result that should change.
2. **Authoritative context or data.** Name the source the system must trust and who owns it.
3. **Authority level.** Choose Assistant, Copilot, Workflow Agent, Supervised Digital Worker, or Multi-Agent Process. Name the human approval point.
4. **Weakest or most uncertain layer.** Identify the stack layer most likely to prevent dependable performance.

5. **One consequential sourcing decision.** Choose one layer where build / buy / partner / route / abstract materially changes value, risk, cost, or portability.
6. **One unresolved executive question.** State the decision that requires leadership judgment rather than a technical answer.

Challenge/Extension — optional deeper analysis

Add the complete five-layer map; build / buy / partner / route / abstract reasoning for each layer; assurance and production metrics; a unit-economic measure; and a complete steering-group recommendation that states the decision, rationale, conditions, owner, and next evidence gate. Record the evidence required before expanding the system's authority further.

The Challenge/Extension is deeper analysis of the same class activity, not a separate graded pathway or a new assignment. It does not create a new Blog 3 pathway or alter the existing Transform-path policy.

Try This Before Class

Optional. For executives who want to test the ideas in this module against a real model before the session.

Pick a real task from your work — a memo to draft, a contract to summarize, a customer issue to triage, a budget anomaly to explain. Run it through whichever frontier model your organization gives you access to. Then assess the failures, if any, against the five layers. Module 2's Frontier Diagnostic asked the same question by tracing symptoms; this version asks it through the stack:

- Did the failure sit at the **workflow layer** (you asked for the wrong thing or framed the job badly)?
- At the **context layer** (the model lacked the right source material or used stale information)?
- At the **model layer** (the model was wrong for the task — too small, too generic, too constrained)?
- At the **harness layer** (no tools, no output structure, no eval, no memory of prior turns)?
- At the **assurance layer** (you have no way to know whether this would have failed differently next time)?

Bring one observation from this exercise to class.

Part 2: Field Reference

The material below is for reference and deeper reading. It is not required for class. Use it during the course to scaffold conversations with vendors, board committees, and your own teams; keep it after the course as an executive reference. Each component is self-contained — flip directly to a table, prompt list, or worksheet without reading surrounding prose.

Executive Glossary

Use this alphabetized glossary as a lookup aid for Part 1. It includes only terms that appear in the Pre-Class Read and were not already defined in Modules 1 or 2. Definitions reflect how the terms are used here; vendor usage may differ.

Term	Plain-language definition
AgentOps	The practices used to test, monitor, version, and govern agents, including their tool use, handoffs, approvals, and failures.
AI FinOps	Managing the full cost of an AI workflow, including model use, context, caching, routing, retries, tools, infrastructure, and human review.
AI stack	The connected layers that turn a model into an operating capability: workflow, context and data, model portfolio, agents and orchestration, and assurance.
API (application programming interface)	A defined way for software systems to exchange data or request actions directly, without a person clicking through a screen.
Assurance	The evidence and controls that show an AI system is reliable, safe, secure, cost-effective, and appropriately governed over time.
Context system	The approved sources, retrieval methods, metadata, permissions, and quality controls that provide an AI system with relevant information.
Data exhaust	The records produced as an AI system operates — such as prompts, outputs, tool calls, corrections, approvals, and usage patterns — that can support learning and oversight or increase vendor dependence.
Embedding	A numerical representation of meaning that lets a system compare content by conceptual similarity rather than exact words.
Engine	The underlying AI model that interprets inputs and generates outputs; it is one component of the broader system, not the complete operating capability.
Eval (evaluation)	A structured test of whether an AI system performs acceptably on representative tasks, edge cases, and failure conditions.
Frontier model	A model near the leading edge of general capability at a particular time. The label is temporary because model performance changes quickly.
Grounded / grounding	Tying an AI answer to approved source material so users can inspect whether the claims are supported.
Interface	The place where a person interacts with an AI system, such as a chat window, document sidebar, spreadsheet, editor, or terminal.
LLMOps	The practices used to deploy, test, monitor, version, and improve language-model applications in production.

Term	Plain-language definition
Metadata	Information about a source, such as its owner, date, geography, permissions, status, or version, that helps a system use it correctly.
Model abstraction layer	An organization-controlled layer that separates workflows from a specific model provider so models can be changed with less rebuilding.
Model Context Protocol (MCP)	An open standard for connecting AI applications to external tools and data sources through a common protocol.
Model portfolio	A managed set of models used for different tasks rather than one model used for everything.
Observability	The ability to reconstruct what an AI system saw and did, including sources, model route, tool calls, approvals, errors, latency, and cost.
Open-weight model	A model whose trained parameters are available for others to download or host. Open weights do not necessarily mean open training data, open licensing, or low operating cost.
Orchestration	Coordinating the steps, models, tools, agents, approvals, and handoffs that make up an AI workflow.
Probabilistic system	A system that can produce different outputs from similar inputs rather than following one fully predictable result path.
Prompt injection	Untrusted content that tries to make an AI system ignore the user's intent, reveal information, or take an unauthorized action.
Retrieval-augmented generation (RAG)	A pattern in which a system retrieves information from an external source and gives it to a model to help produce a current, grounded answer.
Token	A small unit of text processed by a model. Providers often meter input and output in tokens, so longer context and responses can increase cost.
Unit economics	The total cost and value of one meaningful unit of work, such as a correctly resolved case or accepted analysis, rather than cost per token alone.
Vector search	Searching by similarity between embeddings so conceptually related material can be found even when it uses different words.

DATED SPECIALIST REFERENCE — checked August 15, 2026. The examples below were moved out of the required reading to reduce simultaneous cognitive load. They preserve the prior evidence trail without asking every student to master the implementation detail. Recheck product, policy, and market claims before using them in a procurement or deployment decision.

Specialist topic	Reference point and executive use
Workflow-first automation	Aflac's dated straight-through claims example shows why explicit verification rules should precede higher authority. ⁵⁰
Regulated context strategy	Truist's public deployments and partnership posture illustrate why ownership of trusted context can matter more than ownership of a model. ⁸⁹
Routing economics	A dated Goldman Sachs inference-cost estimate and Amazon Bedrock's documented prompt-routing pattern illustrate why model sourcing should be revisited as price and performance change. ¹¹¹²
Agent identity and connectors	Microsoft's agent definition, MCP's connector model, and emerging agent-identity work show why tools, authentication, and revocation belong in the harness decision. ¹³¹⁶¹⁷

Specialist topic	Reference point and executive use
High-assurance architecture	Lockheed Martin and Astris AI show a public example of matching operating control and assurance to mission consequence. ¹⁴¹⁵
Regulated and public-sector assurance	Optomed and CDC illustrate different reasons for continuous evaluation, monitored deployment, and change control. ¹⁸¹⁹⁵²
Browser/computer use and open infrastructure	Official OpenAI safety materials support narrow permissions, confirmations, monitoring, and human review for computer-use agents; official Anthropic and Linux Foundation records confirm MCP's December 2025 donation to the Agentic AI Foundation. ⁵³

Decision Tables

Demo Evidence vs. Enterprise Readiness Evidence

Demo Evidence	Enterprise Readiness Evidence
The AI completed a polished scripted task	The system works on messy, old, incomplete, and contradictory inputs
The model produced a plausible answer	The answer is grounded in approved sources and can be inspected
The agent used tools successfully	Tool permissions, approvals, and audit trails are defined
The workflow looked fast	The cost per successful task is acceptable at scale
The vendor showed a compelling interface	Data terms, exit rights, and switching costs are understood
Leaders saw potential value	Owners, metrics, controls, and escalation paths are assigned

Stack Layer → Executive Question → Strategic Risk

Stack Layer	Executive Question	Strategic Risk If Weak
Business workflow and application	Where does the AI create value, and who uses it?	Technically impressive tool with weak adoption or no measurable business outcome
Context and data	What does the system need to know, and is it allowed to know it?	Confident answers grounded in stale, low-quality, or unauthorized information
Model portfolio and routing	Which model should handle which task, at what cost and risk?	Overpaying for frontier models, underperforming with cheap models, or locking into one vendor
Agent, tools, and orchestration	What actions can the AI take, with what authority and controls?	Unsafe automation, invisible delegation, or uncontrolled tool access
LLMOps, AgentOps, and assurance	How do we test, observe, improve, and govern the system over time?	Pilot succeeds, production fails, costs spike, or incidents cannot be explained

Customer Service Use Case → Authority Level

Customer Service Use Case	Value Logic	Stack Implication	Authority Level
Answer common questions	Reduce handle time and improve consistency	Approved knowledge base, citations, basic logs	1

Customer Service Use Case	Value Logic	Stack Implication	Authority Level
Draft replies for agents	Improve productivity and tone	Human review, version history, agent feedback	1
Triage and recommend next action	Improve routing and prioritization	CRM context, confidence thresholds, escalation rules	2
Prepare refund for approval	Reduce manual work while preserving control	Order system access, approval workflow, audit trail	3
Issue small refunds autonomously	Resolve low-risk cases faster	Hard value limits, exception rules, monitoring, rollback	3–4
Negotiate retention offers	Protect revenue and personalize service	Customer value model, offer governance, compliance review	4

Context Operations (Write / Select / Compress / Isolate / Cache)

Context Operation	What It Means For Executives	Example Stack Decision
Write	Persist useful information outside the current chat or context window	Maintain knowledge files, memory stores, ticket histories, meeting notes, or structured records
Select	Retrieve the right information at the right time	Use search, RAG, metadata, filters, source ranking, and role-aware permissions
Compress	Reduce token load while preserving signal	Summarize long histories, structure notes, chunk documents, and keep executive context packets concise
Isolate	Keep context windows focused and separated where risks or tasks differ	Split workflows, projects, tools, sub-agents, and sensitive decision classes
Cache	Reuse stable context so repeated work is cheaper and more consistent	Cache system prompts, policy prefixes, standard instructions, and stable reference packets

Context Diligence Questions

Context Question	Why It Matters	Evidence to Ask For
Which sources are authoritative?	Prevents the system from grounding answers in unofficial or superseded material	Source inventory with named owners
Who owns freshness?	Keeps the AI from relying on stale policies or outdated product data	Update cadence and ownership model
Are permissions enforced at retrieval time?	Prevents the AI from exposing information the user should not see	Role-based retrieval tests
How are contradictions handled?	Avoids blended answers from conflicting documents	Conflict rules and escalation path
Can users inspect sources?	Builds trust and supports auditability	Citations, source snippets, and feedback channel
Are retrieval and answer quality tested separately?	Helps leaders diagnose whether the context system or the model failed	RAG eval results by failure type

Context Question	Why It Matters	Evidence to Ask For
What context is cached?	Controls cost and consistency for repeated workflows	Cache-hit rate and cache-invalidation rules
What is retained or deleted?	Limits sensitive data exposure and compliance risk	Retention policy for documents, embeddings, traces, and memory

Model Strategy Tradeoffs

Model Strategy	Best Fit	Watch-Out	Executive Metric
Frontier proprietary model	Ambiguous reasoning, synthesis, high-value judgment support	Cost, latency, roadmap dependence	Cost per accepted answer or decision
Smaller proprietary model	Routine drafting, classification, extraction, summarization	Quality drop on edge cases	Task success rate by complexity
Open-weight model	Data control, customization, portability, sovereign deployment	Talent and infrastructure burden	Total cost to operate and improve
Specialized model	Code, image, speech, embeddings, search, extraction	Narrow capability boundary	Performance on the specialized task
Small language model (SLM)	Edge, on-device, sensitive-data, latency-critical	Capability ceiling on complex reasoning	Performance per dollar at workload scale
Model routing	Mixed workloads with varied cost, latency, and risk	Hidden complexity and accountability	Cost per successful task by route
Model abstraction layer	Rapidly changing vendor landscape	Abstraction can hide useful model-specific features	Switching time and regression rate

Harness Component Map

Harness Component	Executive Question
System prompt	What standing contract governs role, goals, constraints, examples, and output format?
Context window management	What gets loaded, omitted, summarized, isolated, or cached for this task?
Tools	What verbs can the system take in the world, and when should each be used?
Retrieval	How does the system find approved knowledge, and how is retrieval quality tested?
Memory	What persists across turns or sessions, and who can inspect or prune it?
Output structure	What schema, template, or artifact makes output reviewable and reusable?
Evaluation loops	How do we know whether a prompt, context, model, or tool change helped or hurt?
Orchestration	How do multiple steps, models, tools, approvals, or sub-agents fit together?

Authority Ladder Control Burden

Level	Course Term	What the Stack Must Support	Minimum Technical Controls
1	Assistant	Drafts, summarizes, or analyzes only	Source boundaries, user feedback, basic logging

Level	Course Term	What the Stack Must Support	Minimum Technical Controls
2	Copilot	Recommends actions inside a human-led workflow	Rationale, confidence signals, human approval, usage monitoring
3	Workflow Agent	Completes bounded tasks in approved systems	Tool permissions, role-aware retrieval, approval gates, tool-call logs
4	Supervised Digital Worker	Acts across systems within defined thresholds	Runtime monitoring, audit trail, rollback path, kill switch, exception handling
5	Multi-Agent Process	Coordinates multiple agents across an end-to-end process segment	Operating model owner, risk tiering, red-team testing, incident response, periodic assurance

Production Control Panel

Control Area	Example Metrics	Executive Question
Quality	Task success rate, human acceptance rate, user correction rate, expert review	Is the AI actually helping the workflow?
Grounding	Citation accuracy, unsupported claim rate, source inspection rate	Can users trust where the answer came from?
Retrieval	Retrieval relevance, missing-source rate, stale-source rate	Is the system finding the right context?
Reliability	Latency, uptime, timeout rate, fallback rate	Can users depend on it during real work?
Agent behavior	Tool-call success, approval failures, retry loops, unexpected handoffs	Is the agent acting predictably?
Safety and security	Prompt injection attempts, sensitive data exposure, excessive agency events	Are risks visible before they scale?
Cost	Token usage, context length, cache-hit rate, model cost, tool-call cost, cost per successful task	Is the unit economics story improving or worsening?
Adoption and value	Active usage, workflow completion time, error reduction, KPI movement	Is the stack producing business value?

Eval Type

Eval Type	Tests For	Example Pass/Fail Question
Golden task evals	Representative work the system must handle	Does the system complete the task at an acceptable quality threshold?
RAG evals	Retrieval relevance, groundedness, source completeness	Did it retrieve the right source and stay faithful to it?
Agent evals	Tool use, approval behavior, handoffs, failure handling	Did the agent use the right tool and stop when authority ran out?
Adversarial evals	Prompt injection, jailbreak resistance, sensitive-data egress	Does the system fail safely under hostile input?
Business evals	Cycle time, quality, revenue, cost, risk, customer experience	Did the AI change the outcome that justified the investment?

Provider Category

Provider Category	What It Provides	Strategic Advantage	Executive Watch-Out
Hyperscaler AI platforms	Model catalogs, agent runtimes, data services, deployment tooling, observability	Integrated enterprise platform	Ecosystem lock-in and pricing complexity
Frontier model providers	High-performing proprietary models through APIs	Capability and speed	Roadmap dependence, data terms, cost volatility
Open-weight model ecosystem	Models that can be hosted, tuned, and controlled more directly	Control, customization, portability	Talent, infrastructure, security, and operations burden
Enterprise SaaS copilots and agents	AI embedded in CRM, ERP, productivity, HR, service, and other workflows	Distribution and adoption	Limited differentiation and opaque architecture
Data and lakehouse platforms	Governed data, search, vector capabilities, context services	Proximity to enterprise data	Data gravity and platform dependency
AI gateway and routing platforms	Cross-provider routing, observability, cost limits, policy enforcement	Multi-provider control plane	Adds a layer that must itself be governed
Observability and eval platforms	Tracing, testing, monitoring, cost, debugging	Production assurance	Sensitive logs, integration overhead, tool sprawl
Security and identity vendors	Access control, data loss prevention, agent identity, threat detection	Risk reduction and control	Controls may lag new agent behaviors
Systems integrators and partners	Implementation capacity, domain expertise, change support	Speed and specialized knowledge	Partner dependency and knowledge-transfer risk

Build / Buy / Partner / Route / Abstract — Default Bias by Layer

Stack Layer	Default Executive Bias	Reason
Business workflow and application	Own the workflow design	This is where differentiation, adoption, and accountability live
Context and data	Own the trusted context strategy and context-quality standards	Proprietary context is often the durable advantage
Model portfolio and routing	Use a portfolio; avoid premature lock-in	Model performance and price change quickly
Agents, tools, and orchestration	Buy for common tasks; build or partner for strategic workflows	Harness design determines reliability, authority, and risk
LLMOps, AgentOps, and assurance	Standardize centrally, extend locally	Production assurance, evals, monitoring, and FinOps must be consistent across use cases

Sourcing Choice — When to Use, When to Avoid

Choice	Use When	Avoid When
Build	Proprietary data, workflow, or customer experience creates durable advantage	The organization lacks data quality, engineering capacity, or operating discipline
Buy	The workflow is standard, speed matters, and the capability is not differentiating	The vendor controls a core workflow, context layer, and roadmap

Choice	Use When	Avoid When
Partner	Specialized expertise, regulated deployment, or ecosystem access is required	IP ownership, data rights, or operating responsibilities are unclear
Route	Model quality, cost, and latency vary by task	The task requires a single certified model for compliance or auditability
Abstract	The market is changing quickly and portability matters	Abstraction hides important model-specific capabilities or accountability

Engine, Harness, and Interface — What You Are Actually Choosing

Executives routinely conflate three distinct things when they talk about "which AI we use." The confusion is understandable, since vendors market all three under one brand, but it produces bad procurement conversations because the three layers have different costs, different switching penalties, and different owners.

DATED PRODUCT REFERENCE — checked August 15, 2026. Product names and model versions in this panel are snapshots, not durable recommendations. Verify current naming, availability, routing behavior, data terms, and enterprise controls before acting.

The three layers

Layer	What it is	Examples (August 2026)	Who chooses it
Engine	The underlying model that produces the output. Interchangeable in principle; the thing benchmarks measure.	GPT-5.6, Claude Opus 5, Gemini 3.1 Pro, Llama, Mistral	Vendor, or your IT organization if you route
Harness	The scaffolding around the engine: memory, retrieval over your files, tool access, connectors, permission gates, logging. Determines what the engine can <i>see</i> and <i>do</i> .	Projects, Copilot Notebooks, custom agents, RAG pipelines	Your organization, increasingly you
Interface	Where you sit while working. A chat window, an editor, a spreadsheet sidebar, a terminal.	ChatGPT app, Copilot in Word, Cursor, Claude Code	You

The same engine performs very differently across harnesses. This is the practical meaning of the argument in Layer Two: an organization that upgrades its engine and leaves its harness alone will see modest gains and conclude that the technology underdelivers. The engine was never the constraint.

Why the vocabulary is confusing

Three sources of confusion account for most of the questions executives actually ask.

One product name spans multiple layers. "Copilot" refers to a consumer chat assistant, an enterprise chat assistant with different data terms, an in-application sidebar in Word and Excel, a project workspace called Notebooks, and a platform for building custom agents. Asking whether a colleague "has Copilot" answers almost nothing. The same is true of "Gemini," which names both a model family and an application, and of "ChatGPT," which names an interface running engines with entirely different names.

Model version numbers no longer move together. As of August 15, 2026, Google's fast and reasoning tiers used different version sequences. A reasonable person may assume the higher number is newer and better. Within a single vendor's lineup, that inference can fail. The dated table above preserves the exact examples used for this course snapshot.

Naming schemes change under you. OpenAI retired GPT-4o, GPT-4.1, and o4-mini from the consumer picker in February 2026 and introduced a named frontier family (Sol, Terra, and Luna) in July 2026, alongside a separate naming system for work, coding, and developer contexts. Any document that hard-codes model names, including this one, has a shelf life measured in months.

Choosing a model in practice

Under the ambiguity, the operating rules are stable:

1. **Use auto-routing selectively for routine work.** Several major providers offer modes that select speed or reasoning based on the request. Test that behavior on your work; a consequential failure can be a fast answer to a question that deserved deeper analysis.
2. **Escalate deliberately, not reflexively.** Reach for the reasoning or extended-effort tier when the task involves multi-step analysis, unfamiliar synthesis, or a decision where being wrong is expensive. Effort settings are usually a control on how long the model works before answering, not a different model.
3. **While learning, use the strongest tier available.** You cannot calibrate what the technology can do from its cheapest configuration. Optimize for cost after you have seen the ceiling, not before.
4. **Ask the system.** Describing the task and asking which model or mode fits gets a useful answer, and it is faster than reading a comparison chart that will be stale next quarter.
5. **Do not benchmark-shop for general work.** Public benchmarks measure narrow capabilities on standardized tasks. The relevant question, per Layer Three, is performance on *your* task at *your* cost per successful outcome.

Where the switching cost actually sits

Engines are the easiest layer to change and the one executives spend the most time debating. Interfaces are moderately sticky, a matter of retraining and habit. Harnesses are where lock-in accumulates: the accreted project knowledge, the connector permissions, the custom agents, the retrieval indexes, the institutional prompts. An organization that changes model providers has a procurement event. An organization that has built five years of harness on one vendor's proprietary scaffolding has something closer to a migration.

The strategic implication runs against the intuition that the model is the important decision. Model choice is reversible and increasingly commoditized. Harness architecture is where the dependency forms, and it is usually decided by default rather than deliberately.

BOARD QUESTION. *"If our model provider doubled its price or degraded its service next quarter, what would we actually have to rebuild, and who decided that?"*

Named products and version numbers in this panel are current as of August 2026 and will change. The three-layer distinction is durable; the examples are not. Verify current names before relying on them in a procurement conversation.

Personal Context Portability: What Travels With You, and What Does Not

Layer Two treats context as an organizational asset. Individuals accumulate one too. An executive who has worked with an AI system for a year has built something real: memory of their preferences, project workspaces holding reference documents, saved instructions, a history of what worked. That accumulation is why the tool feels capable. It is also why moving to a different account or a different provider feels like starting over, because in most respects it is.

DATED CAPABILITY REFERENCE — checked August 15, 2026. Export, memory, project, connector, retention, and migration behavior varies by provider, account type, region, and administrator policy. Verify the current official documentation and organizational rules before moving any context.

The question surfaces in three forms, and they have different answers.

The three portability questions

Move	Short answer (August 2026)	What actually moves
Personal account to corporate account	Partly, and it may be against policy	Prompts and reference documents may move by hand; memory and project structure require platform-specific verification
One provider to another	Limited and provider-specific	Conversation archives may export; memory and functional workspaces may not migrate
Cleaning up an account without losing something	Yes, with one precaution	Export first, then delete; deletion of conversations does not delete memory or project knowledge

What each layer of personal context does

Conversation history is often exportable as an archive. That preserves what was said, not the working state that made the tool useful.

Memory, the accumulated model of your preferences and working context, is only partly portable. A documented import path may move a summary in one direction without recreating the original system's behavior, provenance, or controls.

Projects and workspaces hold reference documents, standing instructions, permissions, and accumulated knowledge for recurring work. A conversation export does not recreate that working environment. Plan for a governed rebuild unless the specific provider documents a functional migration path.

Custom instructions and prompt libraries are the easiest to move, because they are text you wrote. The constraint is that most people never wrote them down separately from the tool. Keeping a prompt library in a document you control, rather than only inside a vendor's settings panel, converts the hardest lock-in problem into a copy-paste problem.

The governance question underneath

Personal portability is the same problem as shadow AI, viewed from the employee's side rather than the organization's.

An employee whose personal account knows their work well and whose corporate account does not has an incentive to keep working in the personal account. That is not a discipline failure; it is a rational response to a capability gap. The organization's options are to close the gap, to accept the leakage, or to prohibit and not notice.

Movement in the other direction carries the sharper risk. Exporting project knowledge, memory, or accumulated context from a corporate account into a personal one is data exfiltration, whether or not it is intended that way, and it is the one form of AI portability most organizations have no policy on. Importing personal context into a corporate account is the mirror problem, introducing material of unknown provenance into a governed environment.

The Module 4 governance material treats the enterprise controls. What belongs here is the architectural observation: personal context is a real asset with real switching costs, and many organizations have not clearly decided who owns it.

Practical guidance

1. **Keep the durable pieces outside the tool.** Prompt libraries, standing instructions, and reference documents belong in a location you control, not only in a vendor's settings.
2. **Export before you reorganize.** Deleting conversations to clean up an account does not delete memory or project knowledge, and the reverse is also true. Know which one you are clearing.
3. **Do not solve the personal-versus-corporate gap by moving data across the line.** If the corporate tool is inadequate, that is a finding worth escalating rather than a problem to route around.
4. **Assume the workspace does not move.** When choosing where to build a substantial body of AI-assisted work, treat the project layer as the commitment. The engine is portable; the workspace is not.

BOARD QUESTION. *"When an employee leaves, what AI-accumulated working context leaves with them, and what did we lose that we never inventoried?"*

Current as of August 2026. Export and memory-portability features are changing rapidly, and EU Data Act cloud-portability provisions take effect September 12, 2026, which may change the vendor baseline. Verify current capabilities before making a migration commitment.

Do / Don't Quick Reference

Do	Don't
Run task-specific model bake-offs using your own data, workflow criteria, latency, risk, and cost	Choose models by brand or public benchmark
Fund trusted context — ownership, curation, metadata, permissions, freshness, evaluation, feedback	Treat RAG as a tool purchase rather than a knowledge management problem
Ask what an "agent" can perceive, decide, remember, and do before approving the label	Accept the rebranding of ordinary chatbots, scripts, or automations as agents
Match agent authority to evidence; expand only after controls and evals support it	Give agents tool access before controls are ready
Track cost per successful task and set usage guardrails before scaling	Scale usage before unit economics are understood
Require a joint technology and data roadmap before platform selection	Select platforms before source quality, ownership, and access rules are clear
Decide deliberately where portability matters, layer by layer	Build deeply around one vendor's model, data format, or proprietary tools by convenience
Bring security, legal, risk, finance, and compliance into design before procurement	Bring controls in after pilot choices have hardened
Require an exit clause, data residency terms, and training-rights restrictions in any AI vendor contract	Accept "we will not train on your data" as the only data-use answer
Ask vendors to demonstrate role-aware retrieval, audit trails, and rollback in your environment	Accept demos using vendor-curated test data as evidence of enterprise readiness

Cautionary Tales

Short, dated public AI cases relevant to stack and sourcing decisions. Each is sourced; each is summarized briefly enough to use as a one-paragraph reference. Where reporting is secondary or partner-sourced, the hedge is carried through.

Klarna and the Customer Service Reversal (2024–2025)

Discussed in detail in Module 5; cross-referenced here for the stack-layer implication. The substantive lesson sits at Layer 1 (workflow design) and at the sourcing layer: an AI deployment that replaces a workflow rather than redesigning it tends to surface its costs only after customer experience erodes, and the eval discipline that would have caught the quality decline earlier was not in place at sourcing.²⁶

Cursor and the Hallucinated Refund Policy (April 2025)

Discussed in detail in Module 4; cross-referenced here for the stack-layer implication. The substantive lesson sits at Layer 4 (agent without disclosure) and Layer 5 (eval gap): an agent that responds confidently to policy questions without grounding evals or human-disclosure scaffolding can do real damage in hours.²⁷

DPD and the Chatbot Disabled in 24 Hours (January 2024)

On January 18, 2024, the parcel-delivery firm DPD's AI chatbot — following a system update — produced output that included swearing at a customer and writing a poem describing DPD as "the worst delivery firm in the world." The chatbot was disabled within 24 hours. The substantive lesson is at Layer 4 (guardrails as a stack-layer concern, not a vendor feature) and Layer 5 (the absence of canary or regression evaluation between deployments).²⁸

Salesforce Agentforce and the Eight-Instruction Limit (2025)

Throughout 2025, *The Information* and other reporting documented Salesforce shifting from a primarily LLM-reasoning Agentforce architecture toward a more deterministic "Agent Script" approach, with Salesforce Chief Technology Officer of Agentforce Muralidhar Krishnaprasad observing publicly that when given more than approximately eight instructions, the underlying language model "begins to omit some." Practitioner reporting (Clientell, late 2025) put Agentforce adoption among Salesforce customers in the low single digits and characterized the shift as a partial walk-back of earlier autonomy claims; we treat the specific adoption figure as practitioner-grade rather than Salesforce-confirmed. The substantive lesson is at Layer 4 (instruction-following limits in agent harnesses) and Layer 5 (eval and grounding metrics matter more than the autonomy headline).²⁹

Microsoft Recall and the Privacy Delay Cycle (2024–2025)

Microsoft announced its Recall feature in May 2024, then delayed launch on June 17, 2024, October 31, 2024, and again in 2025 in response to privacy and security concerns; the UK Information Commissioner's Office wrote to Microsoft, and rollout was scoped to Windows Insiders on Snapdragon Copilot+ PCs with virtualization-based security and Windows Hello Enhanced Sign-in Security. The substantive lesson is sourcing-layer: even disciplined buyers should ask what assurance evidence will look like *before* a feature ships, and what their fallback is if the vendor's rollout slips.³⁰

Provider Concentration: AWS US-EAST-1 and Anthropic (October 2025)

On October 20, 2025, an AWS US-EAST-1 outage rooted in DynamoDB DNS failure cascaded through services across Snapchat, Alexa, Coinbase, Duolingo, and Fortnite; Amazon Bedrock reported elevated error rates and latencies. On October 10, 2025, Anthropic experienced a separate Claude.ai and Claude Code login outage. Third-party aggregator IsDown reported, as of late April 2026, 142 Anthropic incidents in the prior 90 days (48 major, 94 minor, median duration approximately 70 minutes); the figure is directional rather than authoritative because IsDown combines official status pages with user reports. The substantive lesson is sourcing-layer: provider concentration is a real operating risk, and multi-provider routing belongs in the architecture before the outage, not after.³¹³²

AI Vendor and Agent Due Diligence

Because AI vendors often touch data, workflow, models, and decision-making, AI vendor diligence should be more rigorous than ordinary software procurement. The vendor's demo is only one part of the evaluation. Buyers need to understand data rights, model behavior, agent authority, observability, cost structure, security

posture, and exit options. The diligence questions work best as a structured board conversation rather than a technical questionnaire.

Diligence Area	Questions Leaders Should Ask	Red Flag
Data and context	Will inputs, outputs, retrieved documents, user feedback, or traces train vendor models? Where is data processed, stored, logged, backed up, retained, and deleted?	Vague data-use terms or no clear deletion process
Permissions and retrieval	How are user permissions applied during retrieval? Can users inspect and challenge sources?	The vendor cannot demonstrate role-aware retrieval
Model strategy	Which models are used? Can they be changed? Can routing vary by task, risk, latency, cost, or geography?	The application is tightly coupled to one opaque model path
Agent authority	What tools, APIs, systems, and data sources can the agent access? Does it act as the user, a service account, or a named agent identity?	Broad tool access without least-privilege controls
Safety and security	What prevents prompt injection, tool misuse, sensitive data exposure, or excessive agency?	Controls are described as future roadmap items
Operations and assurance	Can we reconstruct what the agent saw, decided, retrieved, and did? What evals, traces, logs, rollback, and kill-switch capabilities exist?	No exportable traces or no incident process
Commercial model	What are the pricing units: seat, token, action, API call, workflow, model, capacity, or outcome?	Costs scale unpredictably with context length or tool calls
Strategic control	What are the exit rights? Which parts are proprietary? What would it cost to switch or bring the capability in-house?	No practical export path for data, traces, prompts, or configurations
IP indemnification	Does the vendor commit to defending you against IP claims arising from model output? Under what conditions?	No commitment, or commitment scoped narrowly to specific products

OWASP's 2026 Top 10 for Agentic Applications provides a useful vocabulary for the safety and authority questions because it focuses specifically on autonomous and agentic systems that plan, act, and make decisions across workflows.³³

The tone of this diligence should be serious but not hostile. A mature vendor should welcome disciplined questions because they indicate the buyer understands enterprise deployment. Vague answers are themselves a signal.

Workbook Prompt: The Stack and Sourcing Worksheet

Use this worksheet to take one AI initiative through the five layers and the sourcing decision. The same matrix is used in the in-class boardroom drill (Part 3) and forms the spine of the Stack and Sourcing Appendix card.

For Foundation/Core preparation, complete the rows for prioritized use case, authoritative context and data sources, agent authority level, the weakest layer (record it under top risks), one consequential sourcing choice, and the steering-group decision. For Challenge/Extension, complete the full matrix. This is deeper analysis of one worksheet, not a separate graded pathway or a new assignment. It does not change Blog 3 or Transform-path policy.

Decision Area	Your Answer
Prioritized use case	One sentence naming the workflow and business outcome
User and workflow	Who uses it, and where does it sit in the workflow?
Agent authority level	Assistant, Copilot, Workflow Agent, Supervised Digital Worker, or Multi-Agent Process — and why this level?
Human oversight	Where is human approval, review, or escalation required?
Context and data sources	What sources are needed, and which are authoritative?
Context engineering choices	What must be Written, Selected, Compressed, Isolated, or Cached?
Data sensitivity	What data is confidential, regulated, or permission-sensitive?
Model strategy	One model, model portfolio, routing, open-weight, managed API, or hybrid?
Tools and actions	What systems can the AI read from or write to?
Harness design	What prompt, context, tools, memory, output structure, evals, and orchestration are required?
Eval	What must be tested before launch and after major changes?
Observability	What traces, logs, and metrics must be visible?
AI FinOps	What is the unit economic metric?
Sourcing choice by layer	Build / Buy / Partner / Route / Abstract — by layer, with reasoning
Vendor diligence	What three questions determine whether the initiative proceeds?
Top three risks	To carry forward into the Module 4 risk register
Steering-group decision	One unresolved decision requiring executive judgment

Use-Case Gallery

Five named non-cohort enterprise stack cases, each described briefly. Drawn from public reporting; secondary sourcing is hedged where appropriate.

JPMorgan Chase — LLM Suite

JPMorgan Chase's LLM Suite was rolled out internally in summer 2024 to eligible employees, expanded to roughly 140,000 employees per remarks by President and Chief Operating Officer Daniel Pinto at the September 2024 Barclays Global Financial Services conference, and reported in May 2025 to be in use by roughly 200,000 employees, approximately two-thirds of the workforce. The platform is internally built and routes across multiple model providers; Derek Waldron, Chief Analytics Officer, has framed the approach in *CNBC* coverage as the foundation for what he termed a "fully AI-connected enterprise." *American Banker* named it 2025 "Innovation of the Year" in June 2025. The architectural pattern is build-internal-on-multiple-providers, with the bank retaining the workflow, the context layer, and the routing logic.³⁴

Goldman Sachs — GS AI Assistant

Goldman's rollout details and the Argenti framing are in the opening vignette of this module. The architectural pattern worth holding here as a Use-Case Gallery anchor is the sourcing logic: build with risk and control review, anchored to a workflow-first thesis rather than a model-first one. Goldman's eval and risk pipeline is the operating discipline behind the verifiability claim.³⁵

Morgan Stanley — AI @ Morgan Stanley

Morgan Stanley deployed AI @ Morgan Stanley Assistant in September 2023 and AI @ Morgan Stanley Debrief in mid-2024. Per Morgan Stanley press releases and the OpenAI customer case study (publicly available), 98% of Financial Advisor teams have adopted the Assistant. Jeff McMillan, Head of Firmwide AI, has described the firm's eval framework as central to deployment. The architectural pattern is strategic-partner sourcing (deep OpenAI relationship) with eval discipline as the operating control.³⁶

Walmart — Sparky and Super-Agent Consolidation

Walmart launched Sparky as a customer-facing super-agent on June 6, 2025. Per multiple secondary sources (Pacvue, Addepto, Klover, the Kellogg School case study), Walmart subsequently consolidated an "agent sprawl" problem into roughly four super-agents (Sparky, Marty, Associate, Developer), standardizing on Model Context Protocol for orchestration and a retail-tuned LLM (Wallaby) as part of the model layer. We have not located a single Walmart corporate document confirming the four-super-agent architecture in one place; the substance is widely reported but the framing here treats the architecture as "publicly reported" rather than corporate-confirmed. The architectural pattern is workflow-first super-agent design with explicit attention to consolidation as a Layer 4 governance decision.³⁷

Georgia-Pacific — ChatGP and the Industrial-Knowledge Stack

Georgia-Pacific operates its Collaboration & Support Center in Atlanta as the hub for AI and IoT operations across more than 100 manufacturing facilities. The company's "ChatGP" generative-AI system, built on AWS, surfaces operator and maintenance knowledge from documents, digital files, and the tribal knowledge captured from long-tenured operators; an AWS case study published in 2024 documented the architecture and an initial expansion plan across additional facilities. Roshan Shah, Vice President of Applied AI at Georgia-Pacific, has publicly framed the program as talent-led: the AI works because the people who hold the institutional knowledge are willing to expose it to the system. The architectural pattern is hyperscaler-partnered build, with the trusted-context layer — the operator and maintenance knowledge base — held as the proprietary asset. IDC named the program in its 2023 Future Enterprise Best in Future of Operations North America awards.⁵¹

Practitioner Standards Reference

Standards an executive should be able to recognize by name. The list is not exhaustive; the rule is that your security, risk, legal, and architecture teams should be able to map their work to this set.

Risk and management standards

- NIST AI Risk Management Framework (January 2023) and the Generative AI Profile (NIST AI 600-1, July 26, 2024).³⁸

- ISO/IEC 42001:2023 AI Management System — first certifiable AI management system standard; named as an affirmative defense in Colorado's AI Act.³⁹
- Cloud Security Alliance Agentic NIST AI RMF Profile (2025) — extends NIST AI 600-1 to agentic systems.⁴⁰

Security and adversarial standards

- OWASP Top 10 for Agentic Applications 2026 (already cited above).³³
- OWASP LLM Top 10 (baseline) — for non-agentic generative-AI applications.
- OWASP Non-Human Identity Top 10 (2025) — relevant where agents act with their own identity.
- MITRE ATLAS v5.1 (November 2025) and the agentic-AI techniques addition (early 2026).²³

Evaluation and benchmarking

- Inspect (UK AI Security Institute, open-sourced May 2024) and Inspect Evals.⁴¹
- MLCommons MLPerf, AILuminate v1.0 (December 2024) and AILuminate Jailbreak v0.5 (October 2025).⁴²
- Practitioner eval-harness shortlist for executive recognition (not endorsement): Inspect, Promptfoo, Langfuse, LangSmith, Patronus, Braintrust, lm-eval-harness.

Industry voluntary frameworks

- Anthropic Responsible Scaling Policy v3.0 (released alongside Claude Opus 4.6) and v3.1 minor update.⁴³
- OpenAI Preparedness Framework.
- Google DeepMind Frontier Safety Framework.
- Frontier Model Forum (industry consortium).

Documentation and supply chain

- Model Cards (Mitchell et al., 2019); System Cards / safety documentation; Datasheets for Datasets (Gebru et al., 2018).
- CycloneDX 1.6 ML-BOM and 1.7 (October 2025); SPDX 3.0 AI Profile; OWASP AIBOM project.

Sectoral guidance

- OCC SR 11-7 (model risk management, applied to AI by US banking regulators).
- SEC marketing rule and 2024 enforcement actions on "AI washing" (e.g., Delphia, Global Predictions).
- FDA AI/ML guidance and predetermined change-control plans (2024).

Regulatory and Procurement Reference

DATED LEGAL AND PROCUREMENT REFERENCE — source trail through May 2026. This section was retained as specialist context and was not re-adjudicated in the August 15 browser-agent freshness pass. Legal dates, obligations, enforcement posture, and vendor commitments can change; verify them with governing primary sources and counsel before use.

EU AI Act phased implementation

- February 2, 2025: prohibitions and AI literacy obligations in force.
- August 2, 2025: General-Purpose AI (GPAI) obligations in force.
- August 2, 2026: Article 50 transparency obligations apply, and Commission and AI Office enforcement powers over GPAI providers begin — both on the original schedule, unchanged by the Digital Omnibus. High-risk enforcement was deferred (see below).
- August 2, 2027: pre-existing GPAI compliance.
- December 2, 2027: Annex III stand-alone high-risk obligations apply, deferred from August 2, 2026 by the Digital Omnibus on AI (European Parliament endorsement June 16, 2026; Council approval June 29, 2026; Official Journal publication to follow).
- August 2, 2028: Annex I product-embedded high-risk obligations apply, deferred from August 2, 2027 by the same Digital Omnibus.

EU AI Act General-Purpose AI Code of Practice

- Published July 10, 2025; signatories listed publicly from August 1, 2025; three chapters (Transparency, Copyright, Safety and Security). Buyer-side relevance: Article 53 downstream documentation rights affect what enterprise buyers can demand of GPAI providers.⁴⁴

United States federal procurement

- Executive Order 14179 (January 23, 2025) revoked Executive Order 14110 (October 30, 2023).
- OMB M-25-21 (Accelerating Federal Use of AI) and M-25-22 (Driving Efficient Acquisition of AI), both April 3, 2025, replaced M-24-10 and M-24-18. M-25-22 vendor-lock-in protection, training-rights restrictions, and continuous monitoring provisions apply to solicitations issued on or after September 30, 2025. Non-federal buyers can use the M-25-22 contract terms as a procurement template.⁴⁵
- December 11, 2025 White House action on state AI law preemption (Executive Order 14365, per Mondaq reporting); DOJ intervention against Colorado's AI Act on April 24, 2026 — Colorado replaced the Act the following month (see below).

State AI laws

- Colorado: SB 26-189 (signed May 14, 2026; effective January 1, 2027) replaced the original Colorado AI Act (SB 24-205) before its June 30, 2026 effective date, dropping the duty-of-care, risk-management-

program, and impact-assessment requirements in favor of disclosure and transparency obligations for automated decision-making technologies; AG-only enforcement, no private right of action.⁴⁶

- Texas Responsible AI Governance Act (HB 149 / TRAIGA): signed June 22, 2025; effective January 1, 2026; 36-month regulatory sandbox; intent-based liability; AG-only enforcement up to \$200,000 per violation; NIST AI RMF substantial compliance as affirmative defense.⁴⁷
- New York: Local Law 144 in force; Synthetic Performer Disclosure law effective June 9, 2026; Algorithmic Pricing Disclosure Act (GBL § 349-A) effective November 10, 2025.

IP indemnification commitments by major AI vendors

- Microsoft Copilot Copyright Commitment (announced September 7, 2023; effective October 1, 2023; expanded 2024–2025).⁴⁸
- Google Cloud Shared Fate / generative AI indemnification (October 13, 2023).
- AWS Customer Agreement IP indemnification update (January 2024).
- IBM watsonx client protections (September 28, 2023).
- Parallel commitments by Anthropic and OpenAI in subsequent contract updates.

FTC enforcement template

- Rite Aid facial-recognition settlement (December 19, 2023): five-year ban on facial recognition; algorithmic disgorgement; 20-year monitoring; first FTC enforcement specifically targeting algorithmic unfairness.⁴⁹
 - Operation AI Comply (September 2024), including Rytr and four additional cases — signals of expanded FTC posture under Section 5.
-

Part 3: Instructor Section

Not part of student pre-class reading. For instructor and program reference only.

In-Class Activity: The Stack and Sourcing Boardroom Drill

The core activity is a boardroom drill. Students work in teams of three to five. Each team chooses a concrete use case, preferably one that could plausibly appear in an executive AI portfolio. Good examples include a claims appeal assistant, a sales proposal agent, a supplier-risk analyst, a financial planning copilot, a customer service escalation agent, a regulatory document reviewer, a field-service troubleshooting assistant, or an internal policy and benefits assistant.

The assignment is to prepare a short briefing to a board technology committee. The team should not pitch the AI capability as a generic innovation. It should explain the stack and sourcing logic behind it.

Teams use the Stack and Sourcing Worksheet from the Field Reference. Each team fills in the matrix and presents a 5–8 minute briefing to the room (or to a peer team acting as the board technology committee).

Debrief questions:

- Did the team's stack follow the business value, or did it start with a preferred tool?
- Did the proposed authority level match the risk?
- Which layer created the most vendor lock-in?
- Which layer was most likely to break at scale?
- Which issues belong in Module 4's risk register?
- Which issues belong in Module 5's ROI and benefits realization work?
- Which issues belong in Module 7's operating model and change plan?

Common failure modes to watch for and name during debrief:

First, starting with a tool instead of a workflow — the team picks a model or vendor first, then reverse-engineers a use case to fit it. Second, over-assigning authority — the team defaults to a high authority level because it sounds more impressive, without the control burden (approval workflow, audit trail, rollback) to support it. Third, treating sourcing as one binary — the team answers "build or buy" once for the whole initiative instead of layer by layer, and misses that the workflow, context, model, and agent layers can carry different sourcing decisions. Fourth, punting assurance — the team's briefing covers workflow, context, and agents in detail but treats evals, monitoring, and cost controls as an afterthought or someone else's problem.

The Stack and Sourcing Appendix

The Stack and Sourcing Appendix is drafted in session; there is no separate post-class submission. For students on the Blog 3 Transform path, it becomes a one-page appendix to the final memo. The appendix connects this module to the rest of the course. It identifies the prioritized use case, maps the five layers of the stack, defines

the agent authority level, describes human-oversight points, recommends build / buy / partner / route / abstract choices by layer, names the top three technical or operational risks (which feed Module 4), specifies production control metrics, identifies an AI FinOps metric, and surfaces one unresolved decision for the executive steering group.

The appendix should not become a technical architecture document. It should be written for an executive audience. The point is to show that the student can connect business value, stack requirements, sourcing choices, risk, cost, and governance in a coherent way.

The appendix also feeds the shared course frameworks. The use case should come from the AI value portfolio. The sourcing recommendation should connect to build/buy/partner logic. The risks should feed the AI risk register. The metrics should connect to the KPI value chain. Ownership questions should connect to the AI operating model. Any near-term actions should feed the 90-day action plan.

Capstone build (Transform path): this appendix becomes the "sourcing and stack implications" section of the final memo. It should help a CIO, CTO, or data leader see what must be owned, rented, routed, governed, monitored, or left portable.

Next-module handoff: Module 4 will use the same use case and authority level to define the risk tier, controls, assurance evidence, human oversight, audit trail, and incident-response path.

Worksheet Row	Feeds Into
Use case, agent authority level, top three risks	Module 4 Governance Card — risk tier, controls, assurance evidence, human oversight, audit trail
Sourcing choices, production control metrics, AI FinOps metric	Module 5 business case — cost structure, unit economics, ROI and benefits realization
Ownership questions, ownership by layer, ownership across the enterprise	Module 7 operating model — who owns what, standardize-versus-extend decisions

Works Cited and Sources

1. Stanford HAI, 2026 AI Index Report, especially the top takeaways and economy chapter.
2. McKinsey, The State of AI in 2025: Agents, Innovation, and Transformation, November 5, 2025.
3. Gartner, Gartner Predicts Over 40% of Agentic AI Projects Will Be Canceled by End of 2027, June 25, 2025.
4. Goldman Sachs corporate communications; CNBC, "Goldman Sachs rolls out an AI assistant for its employees," January 21, 2025; Reuters and Fortune coverage of firmwide rollout, June 24, 2025.
5. Marco Argenti, Chief Information Officer, Goldman Sachs, Fox Business video interview, December 17, 2025.
6. Earlier versions of this course used a four-layer model (workflow, context, model, agents); we add LLMops / AgentOps / assurance as a distinct fifth layer for the 2026 cohort because production discipline has matured into its own architectural concern.
7. Marco Argenti, Chief Information Officer of Goldman Sachs, remarks at Goldman Sachs Communacopia + Technology, October 7, 2025; secondary reporting in Fortune and Prism News. Quoted phrasing reflects practitioner reporting rather than verbatim transcript.

8. Truist corporate disclosures; *American Banker*, "How Truist and Western Alliance deploy AI to fight fraud," June 4, 2025; PR Newswire, "Truist launches AI-enabled receivables platform," February 3, 2026.
9. Truist spokesperson Mazzaferro, quoted in *American Banker*, June 2025.
10. Microsoft Learn, Retrieval-Augmented Generation (RAG) evaluators, accessed May 5, 2026.
11. Goldman Sachs research note on inference economics, June 2025; we cite the figure ("API token prices for frontier models have fallen by roughly 100x since early 2024") as widely reported and directionally accurate, with the date attached because the trend is volatile.
12. AWS, Understanding intelligent prompt routing in Amazon Bedrock, accessed May 5, 2026.
13. Microsoft Learn, What is Microsoft Foundry Agent Service?, accessed May 5, 2026.
14. Lockheed Martin AI/ML capabilities pages and 2024–2025 corporate communications; Aerospace America, "Lockheed's AI Factory," 2026.
15. Lockheed Martin press release, "Lockheed Martin Launches Astris AI," December 16, 2024; astrisai.com.
16. Anthropic, Introducing the Model Context Protocol, November 25, 2024.
17. OpenID Foundation, "Identity Management for Agentic AI," October 2025; MCP OAuth 2.1 integration, June 2025; NIST listening sessions on agent identity, April 2026. Also see Google Cloud, Vertex AI Agent Builder documentation, accessed May 5, 2026, for an alternative agent-runtime view.
18. PR Newswire (AEYE Health), "FDA clears first fully autonomous AI for portable diabetic retinopathy screening," April 30, 2024; Optomed, "Optomed Aurora with AEYE has received FDA clearance in the US," May 1, 2024; PubMed, real-world Aireen + Aurora AEYE study, 2025.
19. David Frasci, then Managing Director, Optomed USA, PR Newswire, April 30, 2024.
20. OpenTelemetry, Semantic conventions for generative AI systems, status: Development, accessed May 5, 2026.
21. Datadog, Agent Monitoring for LLM Observability, accessed May 5, 2026.
22. Datadog, LLM Observability cost monitoring, accessed May 5, 2026.
23. MITRE ATLAS, v5.1 (November 2025) and subsequent agentic-AI techniques addition (early 2026); MITRE corporate documentation, accessed May 2026.
24. Public reporting on the Samsung ChatGPT incidents, March–May 2023; Cyber Security Hub, "Samsung employees allegedly leak data via ChatGPT," 2023.
25. FinOps Foundation, FinOps for AI Overview, updated February 17, 2026.
26. Klarna corporate communications, February 2024 and May 2025; *Bloomberg*, May 9, 2025; *Business Insider*, September 2025.
27. *The Register*, "Cursor AI support bot hallucinated its own company policy," April 18, 2025; Hacker News and Reddit discussion threads referenced in subsequent reporting.
28. *Fox Business*, "DPD AI error causes chatbot to swear, calls itself the 'worst delivery service' to disgruntled user," January 2024.
29. *The Information*, late-2025 reporting on Salesforce Agentforce architecture shifts; Salesforce corporate communications including Muralidhar Krishnaprasad public remarks. Adoption-rate figures from Clientell (2026) are practitioner-grade and not Salesforce-confirmed; we carry the hedge.
30. Malwarebytes, "Microsoft Recall delayed after privacy and security concerns," June 2024; Bleeping Computer, "Microsoft delays Windows Recall again, now by December," October 2024; subsequent corporate updates from Microsoft through 2025.
31. Deployflow, "AWS Outage October 2025: Cause, Impact, and Prevention"; AWS Health Dashboard records for October 20, 2025.
32. IsDown third-party aggregator data, accessed late April 2026; we treat IsDown's incident counts as directional rather than authoritative because the source combines official status pages with user reports.
33. OWASP GenAI Security Project, OWASP Top 10 for Agentic Applications for 2026, December 9, 2025.
34. JPMorgan Chase corporate communications (LLM Suite); CIO Dive, "JPMorgan Chase to equip 140K workers with generative AI tool"; *American Banker*, June 2025.
35. Goldman Sachs corporate communications; CNBC, "Goldman Sachs rolls out an AI assistant for its employees," January 21, 2025; Reuters and Fortune coverage of firmwide rollout, June 24, 2025; *Fox Business* video interview, December 17, 2025.
36. Morgan Stanley, "Launch of AI @ Morgan Stanley Debrief"; OpenAI customer case study on Morgan Stanley AI evaluation framework, publicly available.

37. Walmart corporate announcements (Sparky launch, June 6, 2025); secondary practitioner sources including the Kellogg School of Management case study and industry coverage. The four-super-agent architecture is widely reported but not, to our review, consolidated in a single Walmart corporate document.
38. NIST, AI Risk Management Framework; NIST AI 600-1, Generative AI Profile, July 26, 2024.
39. ISO/IEC 42001:2023 AI Management System Standard; ISACA, "ISO/IEC 42001 and EU AI Act: A Practical Pairing for AI Governance," 2025.
40. Cloud Security Alliance, "NIST AI Risk Management Framework: Agentic Profile," 2025.
41. UK AI Security Institute, Inspect framework, open-sourced May 2024.
42. MLCommons, AILuminate v1.0 (December 2024) and AILuminate Jailbreak v0.5 (October 2025).
43. Anthropic, Responsible Scaling Policy v3.0 (released alongside Claude Opus 4.6) and v3.1 minor update.
44. European Commission, "The General-Purpose AI Code of Practice," published July 10, 2025; signatory list from August 1, 2025.
45. OMB M-25-21 (Accelerating Federal Use of AI) and M-25-22 (Driving Efficient Acquisition of AI), both April 3, 2025; vendor-lock-in and training-rights provisions apply to solicitations issued on or after September 30, 2025.
46. Colorado SB 26-189 (2026), replacing the Colorado AI Act (SB 24-205); signed May 14, 2026, effective January 1, 2027; Hunton Andrews Kurth and Littler client analyses, May 2026.
47. Texas Responsible AI Governance Act (HB 149 / TRAIGA); American Bar Association, "Texas Enters the AI Sandbox with TRAIGA," July 2025.
48. Microsoft, "Introducing the Microsoft Copilot Copyright Commitment," September 7, 2023; subsequent expansions through 2024–2025.
49. FTC v. Rite Aid Corporation; FTC press release, December 19, 2023; Cooley, "FTC Targets Algorithmic Discrimination in Settlement With Rite Aid."
50. Shelia Anderson, Chief Information Officer, Aflac, on *MIT Sloan Management Review: Me, Myself, and AI*, 2023; Anderson described approximately 46% of claims at the time as fully automated through straight-through processing, primarily on lower-value wellness lines. Subsequent Aflac corporate communications and Keith Farley remarks in CMSWire (2025) reported the share above 55%; we cite the 2023 figure with the date attached because the trajectory continues to move.
51. AWS Case Study, "Georgia-Pacific Optimizes Operator Efficiency Using Generative AI on AWS," 2024; Manufacturing Digital, "KONE & Georgia-Pacific, elevators & paper, gen AI & results," 2024; news.gp.com, "Georgia-Pacific Utilizes Artificial Intelligence to Create Safer and More Efficient Facilities," December 2023; IDC Future Enterprise Best in Future of Operations North America award, October 2023.
52. CDC AI Strategy and Vision pages (cdc.gov/ai), accessed May 2026; Nextgov/FCW, "The CDC placed early bets on AI..." December 2025; CDC March 2026 guidance to state, tribal, local, and territorial public-health partners on agentic deep-research tools.
53. OpenAI, Operator System Card, January 23, 2025, especially its treatment of GUI interaction, model mistakes, prompt injection, confirmations, monitoring, and human oversight; OpenAI, Understanding prompt injections, accessed August 15, 2026; Anthropic, Donating the Model Context Protocol and establishing the Agentic AI Foundation, December 9, 2025; Linux Foundation, Agentic AI Foundation launch summary, December 17, 2025. These official sources support the category, risk, control, and foundation-status claims; they do not support a universal benchmark multiple, so the prior roughly two-fold claim was removed.