

The AI capability layer

Four progressive capabilities sit in the AI layer of the maturity model, and they are not one thing. Each has its own value case, its own performance bar, and its own failure modes — which is why treating them as a single “AI capability” is the most common reason AI investment gets mis-prioritized or under-governed.

This is a working reference, not a summary. The targets below are illustrative starting points; the questions are the useful part. Take the one capability you are actively investing in, answer its questions honestly, and fill in its current-state column. That filled-in gap — not the illustrative target — is what should drive your next prioritization conversation.

THE FOUR QUESTIONS BEHIND EVERY ROW

What is it? The outcome the enterprise can produce or protect — not the tools used to produce it.

How well must we do it? The performance expectation, and the gap between current and desired state.

Why does it matter? Which lever it pulls: revenue, cost, service, compliance, or future positioning.

What could go wrong? In performing it — or in failing to perform it at all.

The four capabilities at a glance

CAPABILITY	THE OUTCOME IT PRODUCES	WHERE ITS VALUE CONCENTRATES	THE QUESTION THAT EXPOSES YOUR GAP
1 · NLP & Computer Vision Basics	Unstructured content — text, images, audio, documents — reliably turned into structured, usable signal.	Enabling. Cost avoidance, and a force multiplier for everything above it.	If a better foundation model lands next quarter, how long until we are using it?
2 · GenAI Applications	Answers grounded in our own current, authoritative data, delivered inside the tool people already use.	Near-term revenue and service. The visible proof point.	How many of our proofs-of-concept reached governed production?
3 · Multi-Agent & Orchestrated AI	Multi-step work delegated to agents that plan, act, use tools, and check in only when it matters.	Cost and cycle time on coordination-heavy work.	Which task classes are we actually willing to delegate — and which must stay human-led?
4 · Autonomous Systems	AI and agentic systems running safely, observably and cost-effectively at scale with minimal supervision.	Capacity to scale and de-risk everything below it.	Can we monitor, secure and cost-govern AI as fast as we can build it?

CAPABILITY 1

NLP & Computer Vision Basics

The ability to understand and extract meaning from unstructured content using foundation models and computer vision — on a model-access architecture that isn't locked to a single provider. It is an outcome, not a mechanism: at capability level it does not matter which provider runs it.

Read this as an *enabling* capability. Its value shows up as cost avoided and as headroom for the three capabilities above it — underinvest here and you cap all of them.

Reduce cost	Replaces manual document review, transcription and classification labor with automated extraction.
Improve service	Faster turnaround on document-, image- and audio-dependent processes: claims intake, support tickets, compliance review.
Position for future	Every downstream AI capability depends on this layer's data pipelines and model-access architecture.
Achieve compliance	Reliable extraction and PII-aware processing are prerequisites for regulated document workflows, not enhancements.

PERFORMANCE DIMENSION	ILLUSTRATIVE TARGET	OUR CURRENT STATE
Extraction / classification accuracy	≥ 95% on defined content types	—
Model swap time	A configuration change, not a re-architecture; under one sprint	—
Pipeline reliability	99%+ successful runs, malformed inputs auto-quarantined	—
Time to onboard a new content type	Set locally — depends on current pipeline flexibility	—

ASK BEFORE YOU SET A TARGET

- Which content types do we still process manually, or with brittle point solutions?
- Is model selection configuration-driven today, or hardcoded into application code? A long answer to the swap-time question signals a large gap.
- Who owns data-validation standards across modalities — or is each team applying its own?

CAPABILITY 2

GenAI Applications

The ability to ground generative AI in our own data and deliver it inside real products and workflows — retrieval, governed prompts, and integration that meets the latency and reliability needs of the use case. RAG architecture and vector-store choice are implementation; the capability is the grounded answer arriving where the work happens.

Protect revenue	Product-embedded search, summarization and drafting can differentiate and retain.
Reduce cost	Deflects support volume, accelerates drafting, cuts time spent hunting internal knowledge.
Improve service	Fuller answers for customers and employees, grounded in current data rather than a stale FAQ.
Position for future	This is what most stakeholders mean by “we want AI” — and the proof point that decides whether Capabilities 3 and 4 get funded.

PERFORMANCE DIMENSION	ILLUSTRATIVE TARGET	OUR CURRENT STATE
Answer groundedness	≥ 90% grounded in retrieved source, verified by an evaluation framework	—
Retrieval relevance	Top-k precision meets a threshold defined per use case	—
Time to production (POC → live)	Set locally	—
User-facing latency	Streaming for chat; a defined SLA for batch and report use cases	—

ASK BEFORE YOU SET A TARGET

- Not “can we build a chatbot” but: how many of our GenAI POCs reached production *with* prompt governance, guardrails and monitoring? A low conversion rate is the clearest signal here.
- How would we know if the retrieval layer drifted out of sync with its sources — before a user did?
- Are prompts versioned, reviewed and audited, or treated as disposable code?

CAPABILITY 3

Multi-Agent & Orchestrated AI

The ability to have multiple agents reason, use tools and collaborate toward a shared goal — with human checkpoints on high-stakes decisions and enforced boundaries on what agents may do. This is not a bigger chatbot. It is autonomy over a sequence of decisions and actions, which is a materially different thing to own.

Reduce cost	Automates multi-step workflows — research, draft, route, follow up — that consumed skilled hours per instance.
Improve service	Faster end-to-end resolution of requests that span several systems or data sources.
Reduce cost (2°)	Cuts coordination overhead by routing sub-tasks to the right specialized model automatically.
Position for future	Builds the operational muscle — tool integration, state management, human-in-the-loop — without which autonomy cannot be trusted at all.

PERFORMANCE DIMENSION	ILLUSTRATIVE TARGET	OUR CURRENT STATE
Task completion rate	≥ 85% of delegated multi-step tasks completed without human rescue	—
Tool-call reliability	A defined failure-rate ceiling, with automatic retry and fallback	—
Human-in-the-loop escalation rate	Tracked, trending to the rate that fits the task's risk tier — not zero; zero usually means loose guardrails	—
Runaway loops / cost overruns	Zero tolerance, with hard stopping conditions enforced	—

ASK BEFORE YOU SET A TARGET

- Which task classes are candidates for agentic delegation, and which stay human-led regardless of what is technically possible? This gap is as much a governance decision as a technical one.
- Internal report generation and a customer-facing or financial action should not share a target. Have we set them separately?
- When an agent does something wrong, who owns it: the tool-integration team, the model team, or the business owner of the workflow?

Autonomous Systems

Orchestration and governance at scale: the enterprise integration, safety, cost-management and monitoring layer that makes it responsible to grant AI real autonomy. This is connective tissue. Its outcome isn't a new feature — it is the confidence that every other AI capability can be trusted to run unattended, and it is judged less by what it produces than by what it prevents.

Reduce cost	Token optimization, caching, model-tiering and capacity planning keep autonomy from becoming an unbounded line item.
Achieve compliance	Continuous monitoring for misuse, drift and policy violations, plus audit logging — usually the specific control an auditor asks to see evidenced.
Avoid risk cost	Prevents the cost of a governance failure, which is typically far larger than the governance program itself.
Position for future	Lets the organization say yes to the next autonomy request with confidence, instead of relitigating risk every time.

PERFORMANCE DIMENSION	ILLUSTRATIVE TARGET	OUR CURRENT STATE
Observability coverage	100% of production AI and agentic workloads instrumented for tracing, cost and quality	—
Mean time to detect drift or misuse	Set locally, against risk tolerance	—
Cost-anomaly response time	Automated alerting with a defined response SLA	—
Continuous evaluation coverage	Automated quality gates blocking deployment on regression, across 100% of prompt and model updates	—

ASK BEFORE YOU SET A TARGET

- Is our observability instrumented at the AI-quality level — hallucination rate, drift — or only at the infrastructure level, uptime and latency?
- Is this capability being built proactively, or will it be built after an incident? The honest answer for most organizations: *we can build AI capabilities faster than we can monitor, secure and cost-govern them.*
- Does governance sit with a separate team from the engineers building the systems? If so, expect controls bolted on late and enforced unevenly.

Three things that hold across all four

Value concentrates unevenly, so a revenue-only filter misprices two of the four.

Capability 1 is almost entirely enabling value; Capability 2 carries most near-term revenue and service value; Capabilities 3 and 4 trade near-term value for the capacity to scale and de-risk everything beneath them. Prioritize on “biggest revenue case” alone and you will systematically underfund 1 and 4 — then be surprised when 2 and 3 turn unreliable or ungovernable at scale.

Risk doesn't scale with maturity — it changes in kind.

Capability 1's risks are mostly about data quality. By Capability 4 they are about accountability, regulatory exposure and systemic failure modes. One governance framework applied uniformly across all four will over-engineer the low-risk capabilities and under-govern the high-risk ones.

A gap is only as good as the current-state number behind it.

Every illustrative target here needs a real current-state figure before it becomes actionable. The fastest route to usefulness is to fill in the current-state column for the one or two capabilities under active investment this planning cycle — not to complete all four at once.

DO THIS NEXT

Pick the single capability under active investment or debate right now. Complete its performance table with real current-state numbers from whatever monitoring exists today — even informal monitoring. That filled-in gap, not the illustrative targets above, is what should drive the next round of prioritization and investment conversations.