

Modernizing the AI Center of Excellence for Agentic AI

How enterprises must redesign governance, evaluation, security, and cost controls for Agentic AI, copilots, and agentic systems.

WHITEPAPER

Adnan Masood, PhD
Chief AI Architect, UST

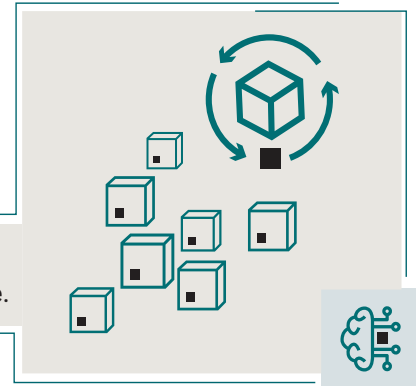
ust.com



■ U
S T

The Enterprise Operating System for AI

Building a modern AI Center of Excellence
From isolated pilots to governed, scalable enterprise value.



Modern AI CoE: Central standards, federated execution, real guardrails, and measurable ROI

A modern AI CoE is the operating system for enterprise AI—not just a data science team or a governance committee. The best model for most organizations is hub-and-spoke: a lean central team sets standards for governance, evals, guardrails, LLMOps, and FinOps, while business units build and scale use cases on shared platforms. In the era of LLMs and agents, the CoE's job is to turn scattered experimentation into safe, repeatable, cost-disciplined business value practices to sustain long-term competitive advantage. By doing so, they can position themselves to thrive in an increasingly digital marketplace.

If it does not operationalize evals, security, agent controls, and cost governance, it is not a true enterprise AI CoE—it is just experimentation with a steering committee.

- **Centralize what must be common:** policy, platform standards, evals, guardrails, cost controls.
- **Federate what must move fast:** business use cases, domain workflows, product ownership.
- **Govern the real risks:** prompt injection, excessive agency, privacy, compliance, and model drift.
- **Manage AI like a product and a portfolio:** not as disconnected pilots.
- **Tie usage to value:** without FinOps, AI enthusiasm becomes runaway spend.

UST helps enterprises build an AI Center of Excellence as a working operating model for AI - not a paper governance body and not a collection of disconnected pilots. UST combines strategy, governance, platform engineering, data and AI architecture, MLOps and LLMOps, and value management to create a CoE that can prioritize the right bets, standardize guardrails, and scale delivery across business units. The result is a hub-and-spoke CoE that centralizes policy, evaluation, security, and cost control while enabling domain teams to move faster and convert AI investment into measurable business outcomes.

Table of contents

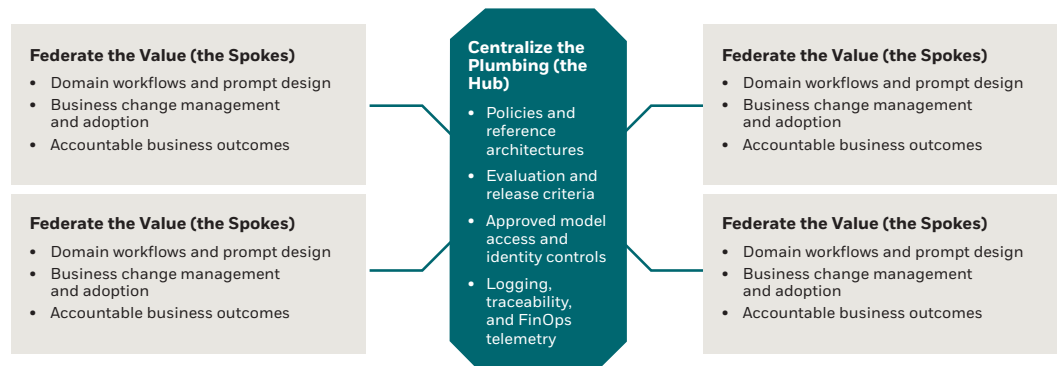
•	Executive summary	3
•	What the CoE is and is not	7
•	Modern enablement stack the CoE must institutionalize	11
•	Evals as a required engineering discipline for AI	12
•	Guardrails and agent controls	13
•	What many CoEs miss in 2026	14
•	Funding, FinOps, and commercial models	15
•	Evidence from case studies and documented examples	17
•	Implementation playbook	20
•	Frequently Asked Questions	23

Executive summary

An enterprise AI Center of Excellence (AI CoE) is a cross-functional operating function that aligns AI to business priorities, industrializes delivery, and reduces risk by standardizing how AI is selected, built, deployed, monitored, and governed. This framing aligns with widely used enterprise descriptions of an AI CoE as a hub for adoption, best practices, enablement, and governance across the organization.^[1]

The strongest “default” model for most large organizations is a federated (hub-and-spoke) operating model: a lean central team defines policy, reference architectures, evaluation and release criteria, and shared platform capabilities, while business units (BUs) and product teams execute use cases using those shared standards. In telecom governance research, the typical maturity path is explicitly described as centralized early, then evolving to hub-and-spoke as use cases scale, with governance supported by cross-functional ethics/oversight and increasing automation/tooling over time.^[2]

The Hub-and-Spoke model balances speed with enterprise control



The center owns common discipline. The business owns business change.

What changed in 2024–2026 (and why it matters for CoE design) is that many organizations are not “just” deploying predictive ML models; they are deploying LLM- and agent-based systems that are:

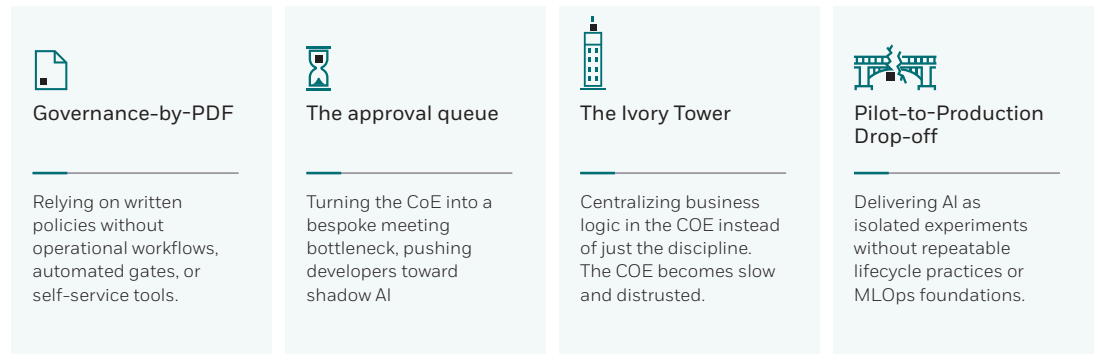
- **Probabilistic and harder to test** like traditional software, creating a need for disciplined evals and regression testing before and after rollout. ^[3]
- **Security-sensitive** in new ways (e.g., prompt injection, insecure plugin/tool design, model denial-of-service, and “excessive agency” where models take actions beyond intended bounds). ^[4]
- **Cost-volatile** because inference spend (tokens/GPU) scales non-linearly with usage, context-length, and architecture choices — creating a need for AI FinOps (showback/chargeback, unit economics, budget guardrails, and optimization disciplines like caching/routing). ^[5]

In regulated and multi-region footprints, the CoE’s governance layer must also map to emerging compliance expectations. For example, the EU AI Act is explicitly risk-tiered and calls for controls such as data quality, logging/traceability, documentation, human oversight, and cybersecurity for high risk systems; it also introduces transparency obligations and phased implementation milestones extending through 2026–2027. ^[6]

The biggest failure modes supported by the evidence base are:

- **Pilot-to-production drop-off** and value leakage when AI is delivered as isolated experiments without repeatable lifecycle practices and platform foundations. ^[7]
- **Governance-by-PDF** (policies without operational workflows, gates, tools, and accountability), which industry governance research warns cannot be solved by “just adding people” and requires the right combination of people, process, and automation/tooling. ^[2]
- **Unmanaged agent risk**, especially where models can trigger actions (tool calls, workflow orchestration) without strong least-privilege access, human approvals for high-impact actions, and security guardrails — risks now reflected in security top-10 taxonomies (including “excessive agency” and “insecure plugin design”). ^[4]
- **Runaway cost** without instrumentation that ties usage to business value and budgets, which FinOps guidance identifies as requiring token/GPU granularity plus governance mechanisms like investment councils and incremental funding reviews. ^[8]

The four traps that derail enterprise AI initiatives



Concise maturity-based recommendation (evidence-backed + labeled where inferential):

- **Ad hoc → Piloting:** start with a small CoE focused on standards, risk gating, and 1–3 shared platform primitives (identity/access, logging, evaluation harness). (Inferred consolidation pattern; grounded in early-stage centralization described in governance research and cloud guidance.)^[2]
- **Piloting → Scaling:** move to hub-and-spoke; add LLMOps/MLOps automation, a model/prompt release process, and FinOps showback.^[2]
- **Scaling → Mature:** formalize a management system approach (e.g., ISO-style AIMS), richer measurement/monitoring, third-party governance, and auditable documentation suitable for high-risk AI categories.^[9]

Definition and boundary conditions

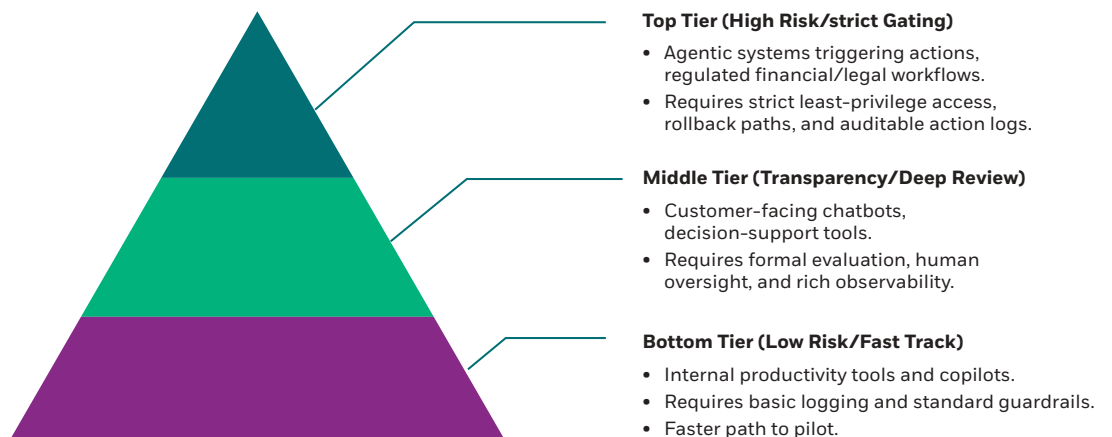
Refined definition

An AI CoE is best defined as an enterprise operating function (often a team plus governance bodies) that runs the way AI is done across the organization: portfolio prioritization, reference architectures, platform enablement, standard operating procedures, risk governance, capability building, and value realization. This aligns with definitions that emphasize adoption + governance + shared resources, and avoids narrowing the CoE to “a data science team.”^[1]

To make the definition operational rather than aspirational, it helps to anchor the CoE to recognized governance structures:

- The AI risk management lifecycle can be modeled using the “govern, map, measure, manage” structure described in the AI RMF Core, with governance as a cross-cutting function infused throughout the other functions and across the AI lifecycle (not a one-time compliance step).^[10]
- A “management system” view (policies, objectives, processes, continuous improvement) is consistent with management system standards like ISO/IEC 42001, which describes an AI Management System (AIMS) and explicitly positions it as an organizational system for responsible development/provision/use of AI.^[9]

Governance must be proportional to business impact



What the CoE is and is not

An AI CoE is not synonymous with:

- A data science delivery team (project execution only)
- An AI platform team (tooling/infra only)
- A Responsible AI or compliance office (policy only)

Instead, it is the integration point where these domains become a coherent operating model—i.e., where the organization decides “how we do AI here,” with repeatable pathways from idea → evaluation → production → monitoring → retirement. This integrated role is explicitly described as a key purpose of an AI CoE: reducing duplicated work, aligning to strategy, enabling shared infrastructure, and improving safety through oversight. ^[1]

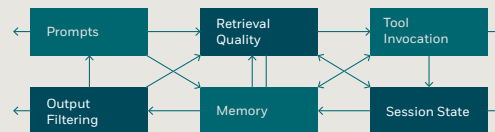
Modern AI fundamentally changes the boundary conditions.

Classic ML



Classic ML Focus: Datasets, model training, and drift. Risk sat mostly inside the model.

Contemporary AI:



Contemporary AI Focus: Prompts, retrieval quality, tool invocation, memory, session state, and output filtering.

THE SHIFT:

Much of the enterprise risk now sits at the system edges. AI risk is highly probabilistic and socio-technical (e.g., prompt injection, sensitive data disclosure, insecure plugin design).

Boundary conditions driven by modern AI

Classic ML CoEs focused on datasets, model training, deployment, and drift.

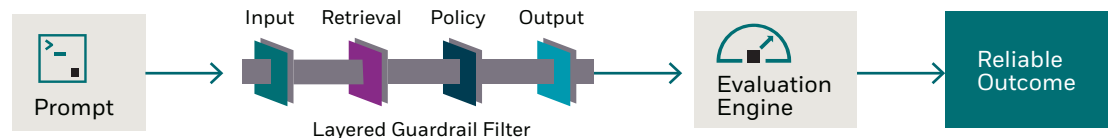
A contemporary enterprise AI CoE must additionally cover:

- **LLM and agent security risks**, including prompt injection, insecure output handling, training data poisoning, model denial-of-service, supply chain vulnerabilities, sensitive information disclosure, insecure plugin/tool design, excessive agency, overreliance, and model theft—now commonly cataloged for LLM applications. ^[4]
- **Generative AI-specific risk management**, reflected in the existence of a dedicated generative AI profile companion to AI RMF (NIST-AI-600-1), explicitly created to help organizations identify and manage risks unique to or exacerbated by generative AI. ^[11]
- **Regulatory differences by risk tier and use case type**. The EU AI Act describes a risk-based model (including prohibited practices; high-risk controls like logging, documentation, human oversight, robustness/cybersecurity; and transparency obligations for certain systems like chatbots and generative AI outputs), with phased applicability between 2025–2027. ^[6]

These boundary conditions imply a practical design rule:

Rule (inferred, grounded in sources): If an AI system can (a) materially impact customers/employees, (b) make or shape decisions, or (c) take actions via tools/workflows, the CoE's scope must include evaluation discipline, governance gates, identity/access controls, and incident response—not just “model building.”^[4]

Converting persuasive demoware into accountable systems



Institutionalize Evals

Measure task success, hallucination tolerance, refusal behavior, latency, and cost. Re-run metrics whenever prompts, models, or data sources change.

Layered Guardrails

Think of guardrails as a system spanning input filtering, retrieval controls, policy checks, tool permissions, and human intervention points.

The Goal

Mitigate the OWASP Top 10 for LLM applications before design failures become security incidents.

Operating model and governance design

Operating model options and when each works

The practical choice is not “centralized vs decentralized” as a binary; it’s “what do we centralize, and what do we federate?” Evidence across governance research, cloud guidance, and FinOps guidance supports a maturity pathway where centralization is common early, but scale requires federation with standardization.^[2]

Below is a decision-useful comparison matrix. Also watch our video that explains AI CoE organizational structures at www.ust.com/en/ust-ai-labs.

Operating model comparison matrix

Decision-oriented synthesis for enterprise AI Center of Excellence design

Model	Where it fits best	Primary advantages	Primary risks	GenAI/agents implication
Centralized build	Early maturity; few domains; scarce senior AI talent	Fast standardization; fewer tools/vendors; consistent controls	Bottlenecking; real domain ownership; “ticket queue” dynamics	Safer early rollout but can throttle demand; hard to keep up with “AI everywhere” pressure ⁽²⁾
Federated hub-and-spoke	Most large enterprises; multiple BUs; broad AI demand	Combines speed + control: central standards & platform; local execution	Requires clear decision rights and automation; risk of “shadow standards” if enforcement is weak	Recommended pattern as use cases scale; governance needs tooling, not just people ⁽²⁾
Embedded (BU-led) with central policy	Digital-native product organization; strong engineering discipline; high autonomy	Maximum domain ownership; fastest local iteration	Tool sprawl; inconsistent risk posture; weak cross-BU reuse	Risk accelerates for agents unless guardrails are centralized (identity/access, eval release gates) ⁽⁴⁾
Platform-led shared services	Enterprises with strong platform engineering; many AI builders	High reuse; self-service; reduces duplicated infrastructure	Can become “platform for platform’s sake” without value governance	Work well for LLMops/MLOps, eval harnesses, routing, and observability ⁽⁷⁾
Hybrid “AI product line” model	Large firms with repeatable domains (claims, onboarding, service)	Combine CoE standards with reusable domain accelerators	Require strong product management and funding discipline	Often best for scaling agentic workflows safely by domain

Governance architecture that scales

A contemporary CoE should be designed as a system of governance bodies + workflows, not only a reporting line. The most repeatable structures align to three layers:

- **Strategic oversight:** a steering committee that prioritizes the AI portfolio and aligns it to enterprise strategy (common in CoE descriptions as a way to prioritize high-impact use cases and align initiatives).^[1]
- **Risk governance:** an AI governance function aligned to “govern/map/measure/manage,” emphasizing continuous risk management and integration across the lifecycle (not just pre-release review).^[10]
- **Investment and cost governance:** an “AI Investment Council” concept is explicitly recommended in FinOps guidance as a mechanism to align rapid AI investments with strategic goals and ethical standards, with incremental “fail-fast” review cycles and ownership clarity.^[8]

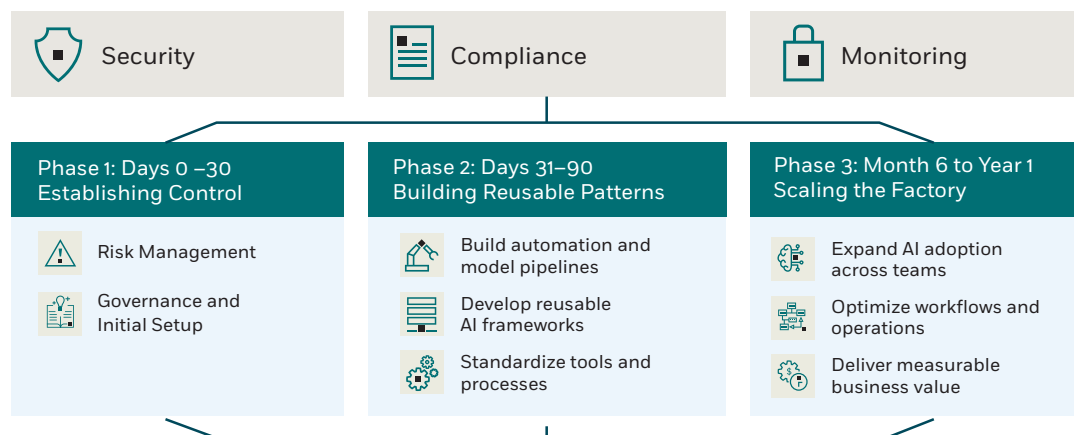
Regulatory-anchored gating

A practical way to prevent the CoE from becoming either toothless or a bottleneck is to implement risk-tiered gates:

- For jurisdictions impacted by the EU AI Act, the Act’s risk model provides a ready-made segmentation: prohibited practices (banned), high-risk systems (strict requirements including logging, documentation, human oversight, robustness/cybersecurity), transparency obligations, and minimal-risk categories—with an explicit rollout timeline.^[6]
- For financial services, model governance expectations long predate GenAI; SR 11 7 describes model risk management as including robust development/implementation/use, effective validation, and sound governance, with emphasis on model inventories and “effective challenge.” Even where GenAI is not explicitly named, the governance pattern maps closely to AI lifecycle controls a CoE would institutionalize.^[12]
- For insurance, NAIC’s AI governance topics describe expectations that insurer decisions supported by AI comply with law, and that governance of AI use is expected; in practice, survey evidence shows organizations using AI review processes and routing vendor AI use through governance protocols involving procurement and AI governance committees.^[13]

The Guardrail layer

Managing modern risk



Modern enablement stack the CoE must institutionalize

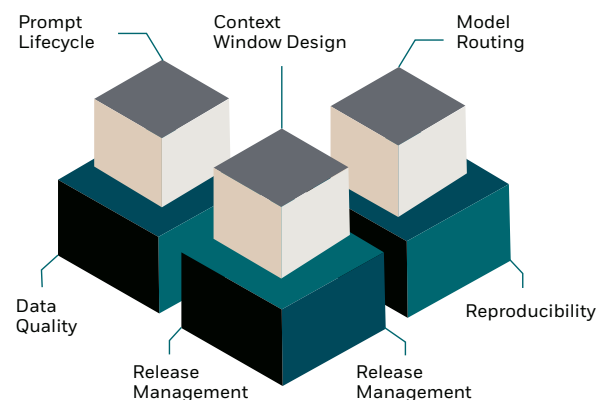
MLOps and LLMOps as “enterprise plumbing”

A modern AI CoE should treat MLOps/LLMOps as core enterprise capabilities for reliability and scale:

- MLOps is commonly described as practices that unify ML development and operations, bridging Dev/Ops for ML and enabling consistent development, testing, deployment, monitoring, and retraining. ^[14]
- MLOps automation typically includes CI/CD/CT (continuous training), and it is explicitly broader than “deploy a model”: it includes data validation, model validation, monitoring, and pipeline automation to mitigate the complexity and pitfalls of real-world ML systems. ^[7]
- LLMOps extends these concerns to LLM-specific issues, and guidance frames it as including monitoring, security audits, and disaster recovery planning—explicitly linking operations discipline to risk reduction. ^[15]

Watch the author explain how MLOps & LLMOps are core to AI reliability and scale at www.ust.com/en/ust-ai-labs.

The modern enablement stack expands from MLOps to LLMOps



The Foundation Remains

Data quality, release management, and reproducibility still matter. Do not ignore foundational MLOps.

The LLMOps Expansion

- Prompt lifecycle management.
- Retrieval behavior and context window design.
- Model routing and failure analysis across probabilistic workflows.

Incident Readiness

The AI CoE must standardize traces, logs, and prompt sampling. A print-ready policy document is not a control. A monitored, testable production system is the necessary control.

A practical CoE implication: the CoE should own (or strongly influence) the standards and shared services that make MLOps/LLMOps real—registries, evaluation harnesses, release gates, observability, and incident response—not only “templates.” ^[10]

Evals as a required engineering discipline for AI

For LLM applications, “it seems good in a demo” is not adequate. Authoritative guidance treats evals as essential for reliability:

- OpenAI’s developer guidance describes evals as tests to ensure model outputs meet specified criteria, and explicitly frames evals as essential when upgrading or trying new models in order to build reliable applications. ^[3]
- OpenAI’s open-source Evals repository positions evals as a framework and registry for evaluating LLMs and LLM systems, reinforcing that evaluation is an operational discipline, not a one-off QA step. ^[16]
- Academic evaluation work (e.g., HELM) argues for multi-metric evaluation beyond accuracy (including robustness, bias, toxicity, efficiency) to expose tradeoffs systematically and improve transparency of language models. ^[17]

CoE design implication (inferred, grounded): every enterprise should have at least a minimum eval stack that includes gold datasets, regression tests for high-risk behaviors, cost/performance measurements, and release criteria tied to risk tiers. ^[3]



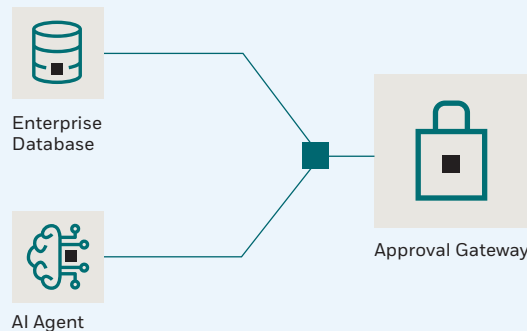
Guardrails and agent controls

Guardrails are not a single tool; they are a layered system spanning input, retrieval, model behavior, tool use, and output. Security taxonomies for LLM apps explicitly identify risks that map directly to guardrail requirements, including prompt injection, insecure output handling, sensitive information disclosure, insecure plugin design, and excessive agency. ^[4]

A CoE should translate these risks into enforceable enterprise controls:

- **Model access + identity and least privilege:** Agentic systems are specifically called out as requiring rule-setting and least-privilege access in practical CoE guidance, because agents often connect to enterprise systems and can take actions. ^[18]
- **Human-in-the-loop for high impact:** High-risk regulatory framing emphasizes “appropriate human oversight” alongside traceability/logging and robustness/cybersecurity; agentic systems should inherit this posture, particularly where actions affect customers, money movement, or legal outcomes. ^[6]
- **Resilience and incident readiness:** LLMOps guidance explicitly links risk reduction to monitoring systems, disaster recovery plans, and security audits — capabilities the CoE should standardize and operationalize. ^[15]

Agentic AI shifts the focus from answer quality to operational authority.



The Pivot point

The moment an AI system can call tools, modify records, or orchestrate other systems, governance maturity must radically increase

Required Agentic Controls

- Clear permission and least-privilege access
- Auditable action logs and rollback paths
- Approval thresholds (human-in-the-loop) for high-impact actions
- Strict boundaries between what agent may recommend and what it may do

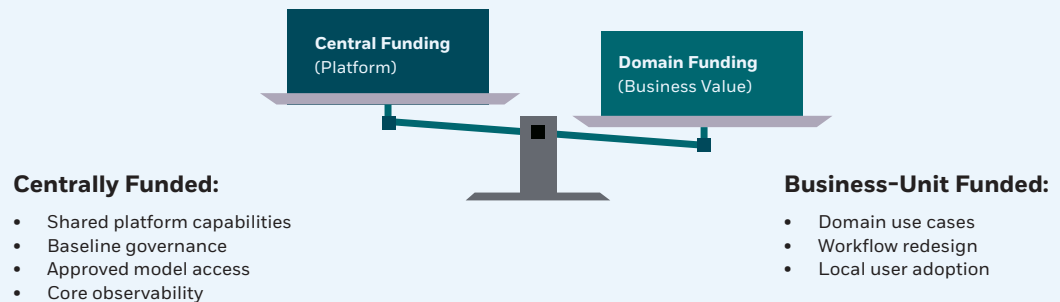
What many CoEs miss

Based on current guidance, the “missed but critical” contemporary topics a CoE should include often are:

- Agent lifecycle governance (registration, approvals, tool permissioning, audit trails for actions), reflecting both “excessive agency” risk and regulatory expectations for traceability/human oversight. ^[4]
- AI FinOps integration (cost per token, cost per inference, forecasting, and optimization patterns like prompt caching/routing), because cost management cannot be an afterthought for enterprise-scale AI. ^[19]
- Third-party model/supply chain governance, explicitly raised both in AI RMF governance categories (third-party/supply chain risk) and LLM security taxonomies (supply chain vulnerabilities). ^[10]

Funding, FinOps, and commercial models

A hybrid funding model drives accountability without throttling adoption



Showback before chargeback: Implement cost visibility (showback) early to build cultural accountability, transitioning to hard chargeback only when budgets and telemetry are mature.

Why AI funding is different in the LLM/agent era

Traditional analytics programs could often fund platform and delivery using stable infrastructure cost models. By contrast, agentic AI introduces cost drivers that are highly usage-dependent and architecture-dependent:

- FinOps guidance for AI highlights non-linear dynamics (usage growth and inference cost multiplication) and emphasizes optimization decisions that affect both performance and budget sustainability. ^[5]
- FinOps guidance also stresses that “per-token price” is often misleading and that real costs include operational nuances (e.g., context window growth) and mitigation techniques like caching and batch processing. ^[20]
- FinOps working group guidance explicitly links AI optimization and accountability to granular metrics and decision frameworks that connect cost, performance, and business impact. ^[5]

AI FinOps

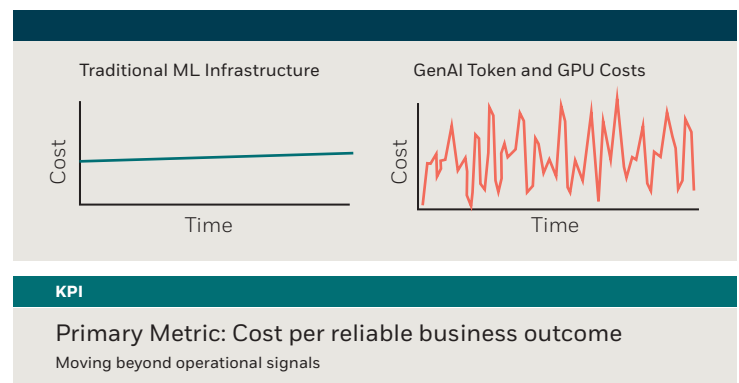
Measuring the cost of reliable business outcomes

The Economic Shifts

AI costs fluctuate violently with token volumes, context growth, and agent chains.

The True Cost

Cost per token and cost per inference are just operational signals. The real metric is the cost to produce a reliable business outcome.



Comparative commercial models for an AI CoE

The following is a decision-oriented model comparison.

Funding options and tradeoffs enterprise AI delivery			
Funding model	What it optimizes for	Typical failure mode	When it works best
Central enterprise fund	Speed of initial capability build; consistent controls	“Free AI” drives demand spikes and unclear ROI	Early maturity, platform foundations, baseline governance and shared tooling ⁽²¹⁾
BU-funded use cases	Clear product ownership and value accountability	Tool sprawl and duplicated patterns without central standards	Mature organizations with strong platform standards and enforcement ⁽¹⁾
Hybrid: central platform + BU use cases	Balanced incentives; scalable reuse	Requires strong chargeback/showback design	Most large enterprises: aligns with hub-and-spoke adoption patterns ⁽²²⁾
Showback (cost visibility)	Culture accountability without hard billing	“Awareness without behaviour change”	Transition phase before chargeback; when politics make chargeback hard ⁽²²⁾
Chargeback / internal rate cards	Cost discipline and unit economics	Teams avoid using the shared platform and drive shadow AI	High scale, strong governance, mature tagging/telemetry and budgeting ⁽²²⁾
Subscriptions/ shared service	Predictability for platform services	Mispricing demand: underfunded innovation	Stable, reusable capabilities (model gate way, eval systems, observability) ⁽²³⁾
Managed service / partner-run	Speed and staffing relief	Lock-in and weak internal capability transfer	Short-term acceleration with an explicit capability transfer plan ⁽²⁴⁾

Practical AI FinOps controls the CoE should institutionalize

FinOps materials provide concrete metrics and implementation approaches that map well to CoE responsibilities:

- Define standard KPIs like cost per token, cost per inference, cost per API call, and anomaly detection for spend. ^[19]
- Implement a cost and usage tracker that supports centralized or hub-and-spoke tracking, with token-level attribution and integration into existing FinOps practices. ^[22]
- Use an investment-governance model (AI Investment Council) to align rapid AI investments with strategy and ethics, and to support incremental funding with frequent “fail-fast” reviews. ^[8]
- Treat optimization as an engineering-finance partnership problem, focusing on cost/performance/value tradeoffs and patterns like prompt routing, RAG optimization, and utilization improvements. ^[5]

Evidence from case studies and documented examples

This section uses documented examples where “AI CoE” is explicit, plus “analogous” operating models when the label is used differently. The purpose is to extract what is repeatable: operating model choices, enablement primitives, governance mechanisms, and measurable outcomes.

Measurable operating impact across the enterprise

Intel	Banco Ciudad	OTP Bank
Built an internal AI CoE tied to repeatable productionization reporting \$1 billion in business value in a single year.	Centralized strategy via an AI CoE to deploy Copilot and Azure-based agents. Developed 10+ AI agents in six months with measured call center savings.	Leveraged a CoE to enforce a centralized data foundation and repeatable workflows, providing that architecture and data discipline do the heavy lifting in regulated environments.

Intel ^[10] IT’s AI CoE as an internal value engine

A first-party Intel IT case document describes its AI group as “quickly becoming an AI center of excellence,” reporting USD 1B in business value in 2020, a team of 200+ AI professionals, and 500+ AI solutions in production—illustrating a mature internal CoE pattern: scale of delivery, productionization discipline, and measured outcomes. ^[25]

The same document also cites an industry benchmark claim that many AI projects do not reach deployment (pilot-to-production failure), reinforcing why a CoE focuses on repeatable delivery and operationalization rather than experimentation alone. ^[25]

BANCO CIUDAD ^[8] AS A COE-LED ENTERPRISE COPILOT AND AGENT PROGRAM

A Microsoft customer story documents the bank establishing an AI CoE to centralize AI strategy and scale adoption, reporting outcomes such as 10 AI agents developed in six months, 90% internal approval for deployed agents, hours redirected, and projected financial impact—an example of a CoE explicitly driving agentic AI enablement with measurable adoption and value metrics. ^[26]

The case also demonstrates a practical organizational pattern: start with a governed pilot (licenses/users), build a multidisciplinary team, then scale as demand grows—matching the “centralize early, federate later” adoption dynamic described in governance and cloud guidance. ^[26]

OTP BANK [1] FOCUSING THE COE ON DATA FOUNDATIONS AND GOVERNED SCALE

A Cloudera customer story documents that the bank established an AI CoE to fuel adoption and enable teams to explore analytics and AI use cases, while emphasizing security and governance requirements due to regulatory constraints. The same account describes a centralized data foundation and “repeatable workflows,” reinforcing a common CoE success factor: shared data/platform primitives that enable BU execution. ^[27]

This case is also a reminder that in regulated domains, CoE value is frequently unlocked first through data governance and platform standardization before advanced GenAI features, aligning with both AI RMF governance expectations and EU high-risk requirements for traceability and documentation. ^[27]

U.S. GENERAL SERVICES ADMINISTRATION ^[2] DEMONSTRATING COE AS ENABLEMENT + ACQUISITION GOVERNANCE

A U.S. federal example documents an AI CoE used to share tools and infrastructure support, create an “AI Guide for Government,” build a community of practice, and provide acquisition guidance for generative AI and specialized computing—illustrating that a CoE can function as an enterprise enablement layer that reduces overhead for experimentation while standardizing responsible adoption and procurement practices. ^[28]

For enterprise CoEs, this is directly analogous to “internal service catalogs” and “approved patterns” for vendor/model procurement and platform consumption. ^[28]

ALTMAN SOLON ^[12] ON TELECOM GENAI GOVERNANCE: HUB-AND-SPOKE PLUS ETHICS COMMITTEES

A telecom governance report describes a maturity trajectory (centralized early, hub-and-spoke at scale), and includes an example where a telecom provider first built a 10-person AI CoE to work with data governance teams, then integrated generative AI capabilities and planned to extend expertise into business units. It also emphasizes that governance scales through tooling and automation, not simply adding people. ^[2]

This example is particularly relevant to agentic AI because it ties governance to process and technology (metrics, third-party model protocols, monitoring, audit readiness)—controls aligned with LLM security risks and broader AI RMF governance expectations. ^[2]

SHELL [10] ILLUSTRATING LONG-RUN COE EVOLUTION FROM COORDINATION TO STANDARDS AND PLATFORMS

A CIO feature describes Shell launching an early AI/data science center of excellence to coordinate analytics projects across business lines, then expanding the role toward common standards, platforms, and assurance processes as AI scaled—directly reflecting the “industrialize and diffuse” path that many enterprises follow. ^[30]

While this is a secondary source and older, it remains a useful longitudinal pattern: start with coordination + expertise concentration, then broaden to platform, governance, and skills diffusion, which matches the AI RMF emphasis on continuous governance and lifecycle integration. ^[30]

Standing up the AI CoE



Watch the author explain how to stand up the AI CoE at www.ust.com//en/ust-ai-labs.

Implementation playbook

Assumptions for this plan (explicit assumptions, not claims): mid-to-large enterprise (5,000+ employees), multiple business units, some ML in production, accelerated AI demand, and at least moderate regulatory exposure (e.g., EU footprint or regulated products).

FIRST MONTH

Set the CoE up as an operating system, not a project:

- Publish a charter that defines decision rights, intake mechanism, risk tiers, and the initial service catalog (common CoE responsibilities: alignment, prioritization, shared resources, and governance).^[1]
- Choose a governance baseline: AI RMF (govern/map/measure/manage) plus a small “playbook” of required artifacts (inventory, documentation templates, evaluation checklists).^[10]
- Stand up minimum LLM governance: approved model access paths, logging/traceability, and initial guardrails mapped to the OWASP LLM Top 10 categories (especially prompt injection, sensitive information disclosure, insecure plugin/tool design, excessive agency).^[4]
- Establish AI cost telemetry foundations (showback-ready tagging, token tracking, per-inference/per-token metrics) using FinOps patterns.^[22]

Deliverables by day 30 (inferred synthesis, grounded in sources): charter, intake rubric, initial standards (data classification + model access), “minimum eval harness,” and cost tracking instrumentation.^[18]

QUARTER ONE

Build the “minimum viable AI factory” primitives:

- Implement or standardize MLOps/LLMOps patterns so teams can deploy with repeatability (CI/CD/CT concepts for ML; monitoring and DR plans for LLMs).^[7]
- Define release criteria tied to risk tiers (low-risk internal copilots vs customer-facing vs regulated decision systems), using risk-based regulation constructs and governance expectations.^[6]
- Operationalize vendor/third-party governance: procurement questions for AI use, AI acceptable-use controls in contracts, and third-party risk assessment integration—patterns seen in insurance governance practice evidence.^[31]
- Pilot 2–3 CoE-backed “lighthouse” use cases that demonstrate repeatable patterns (e.g., RAG knowledge assistant, document intelligence workflow, agentic automation with human approvals). The manufacturing and banking case studies show why pilots must be paired with governance and automation to scale.^[21]

FIRST YEAR

Scale safely through federation:

- Transition to hub-and-spoke distribution: embed “AI champions” or domain AI leads into business units and standardize shared platforms and review gates, consistent with governance maturity descriptions.^[2]
- Establish an AI Investment Council and adopt incremental funding reviews, unit economics, and optimization routines (prompt caching, routing, utilization) as standard governance, consistent with FinOps guidance.^[8]
- Expand to management-system rigor where needed (e.g., ISO/IEC 42001 style AIMS practices) if your risk or regulatory environment requires stronger assurance.^[9]

Enterprise AI capability will either be intentional or incidental

Without a Center of Excellence, AI capability fragments across tools and vendors. Governance becomes reactive, costs become opaque, and the organization confuses activity with progress.

The Leadership Imperative: The point of the CoE is not control for its own sake. It is structure that makes sustained, safe AI ambition possible.



UST's value in building an AI CoE is that we do not stop at designing the structure; we help operationalize the machinery that makes enterprise AI repeatable, governable, and commercially viable. UST works with executive, business, technology, data, and risk leaders to define the CoE charter, decision rights, intake and prioritization model, risk tiers, funding approach, KPI framework, and governance cadence so the enterprise is clear on what should be centralized, what should be federated, and how AI decisions get made. We then help stand up the shared foundation the CoE needs to function at scale: data and AI reference architectures, approved model and data access patterns, secure integration methods, evaluation harnesses, guardrail and red-teaming patterns for LLM and agentic systems, human approval checkpoints for high-impact actions, observability, auditability, and lifecycle workflows that take teams from experimentation to production.

Beyond the technology stack, UST helps define the operating rhythm of the CoE itself - intake reviews, architecture checkpoints, model and prompt release criteria, vendor governance, risk escalation paths, portfolio reviews, and executive reporting - so governance becomes part of delivery rather than a separate after-the-fact exercise. Because UST brings together cloud, data engineering, application modernization, AI/ML, MLOps and LLMOps, and responsible AI capabilities, the CoE is built as an execution engine rather than a policy-only forum. In parallel, UST helps launch lighthouse use cases that prove value quickly while producing reusable assets - prompts, connectors, retrieval patterns, templates, dashboards, and deployment pipelines - that other business units can adopt without reinventing the stack. As adoption grows, UST helps evolve the CoE into a hub-and-spoke model with shared platforms, embedded domain ownership, standardized release gates, vendor and model governance, workforce enablement, and change-management playbooks that accelerate adoption without fragmenting standards.

UST also helps strengthen the human side of the CoE by defining roles, building practitioner playbooks, and enabling product, engineering, operations, legal, compliance, and business teams to work inside a shared AI operating model rather than around it. Just as importantly, UST helps institutionalize AI FinOps and value realization so enthusiasm does not become uncontrolled spend: token and infrastructure telemetry, showback or chargeback models, performance-quality-cost trade-off monitoring, and portfolio reviews tied to business outcomes. This combination of strategy, architecture, governance, platform engineering, and scaled delivery is what turns an AI CoE from a well-intended initiative into a durable enterprise capability - one that can govern responsibly, enable teams confidently, and convert AI ambition into measurable ROI.

Reach out to us today to see how you can get started with your AI COE.

Watch our experts discuss how to build an AI CoE at www.ust.com/en/ust-ai-labs.

Frequently Asked Questions

What is the single best “one sentence” definition of an AI CoE? A cross-functional enterprise structure that accelerates AI adoption and delivery while standardizing governance, best practices, and shared platforms, aligned to strategic goals and measurable value. ^[1]

Why do enterprises create AI CoEs instead of letting teams innovate independently? Because independent teams often duplicate efforts, choose inconsistent tools, and struggle to scale pilots into production; a CoE exists to coordinate shared resources and governance to reduce inefficiency and risk. ^[1]

What is the best default operating model? A hub-and-spoke model: central standards, governance, and platform enablement with federated execution across business units, often as a deliberate evolution from early centralization. ^[2]

How do you prevent the CoE from becoming a bottleneck? By moving review and enforcement into automated standards (templates, eval harnesses, policy-as-code, gated pipelines) and pushing delivery to BUs while the center focuses on shared platform and governance. ^[2]

Where should the CoE sit in the org? Evidence suggests the reporting line matters less than having clear decision rights and executive sponsorship, with strong interfaces to IT/platform, business leadership, and risk governance; examples show CoEs operating effectively as enterprise enablers regardless of exact placement. ^[1]

What’s the difference between the AI CoE and the AI platform team? The platform team builds and runs shared infrastructure; the CoE sets operating standards, governance gates, portfolio prioritization, and value tracking, often coordinating the platform team’s priorities as part of enterprise AI strategy. ^[1]

What’s the difference between the AI CoE and “Responsible AI”? Responsible AI often focuses on principles and compliance; the CoE operationalizes responsible AI through workflows, measurement, release gates, and platform controls consistent with lifecycle governance frameworks. ^[10]

What must be centralized in most enterprises? At minimum: ⁽¹⁾ model and data access pathways, ⁽²⁾ evaluation and release criteria, ⁽³⁾ logging/traceability, and ⁽⁴⁾ guardrails for LLM/agent risks, because these are prerequisites for scalable governance and cost control. ^[10]

What should often be federated? Use-case development, domain data preparation, and product ownership should usually sit in BUs/product teams, once shared platform and governance standards exist. ^[2]

How do LLMs and agentic AI change the CoE compared to classic ML? They introduce new risk classes (prompt injection, insecure tool/plugin design, excessive agency, sensitive data leakage) and require heavier emphasis on evals, guardrails, and usage-based cost governance. ^[4]

What is LLMOps, and who should own it? LLMOps is a set of practices for managing LLM lifecycle performance, scalability, and risk reduction through monitoring, security audits, and operational planning; ownership is typically shared between the CoE (standards/governance) and platform engineering (implementation/operations). ^[15]

What is MLOps, and why does it matter for a CoE? MLOps unifies ML development and operations through automation and monitoring across the ML lifecycle; it enables repeatable delivery and reduces pitfalls that arise when ML systems are treated as “just model code.” ^[7]

What is the role of evals? Evals are structured tests of model/system outputs

Frequently Asked Questions

against defined criteria, used to ensure reliability and manage risk when changing prompts, data, or models; guidance frames evals as essential for reliable LLM applications.^[3]

How do you define “production-ready” for an AI system? A practical definition includes: ⁽¹⁾ passing eval thresholds, ⁽²⁾ threat modeling aligned to LLM risk taxonomies, ⁽³⁾ logging/traceability for outputs and actions, ⁽⁴⁾ human oversight for high-impact actions, and ⁽⁵⁾ cost guardrails and monitoring.^[4]

How do you govern AI agents differently than chatbots? Agents can take actions via tools; security taxonomies explicitly list “excessive agency” and “insecure plugin design,” implying governance must include tool permissioning, least privilege, approvals, and auditable action logs.^[4]

What are guardrails in practice? Guardrails are layered controls that prevent or detect unsafe inputs, unsafe retrieval, unsafe tool use, or unsafe outputs, mapped to known LLM risk categories (prompt injection, sensitive disclosure, etc.).^[4]

How do you manage shadow AI (employees using unapproved tools)? A common approach is to provide sanctioned paths (approved models/tools, training, and policies) and enforce governance through enterprise controls (identity/access, data classification, procurement policy), consistent with AIMS and risk governance principles.^[9]

What is AI FinOps, and why is it mandatory now? FinOps for AI extends cloud financial management to AI workloads, emphasizing token/GPU-level visibility, forecasting, optimization, and governance because AI costs can escalate rapidly and non-linearly.^[5]

What is the minimum viable AI cost governance? Token/call/inference-level telemetry, cost anomaly detection,

showback dashboards, and a lightweight cost review cadence (often via a cross-functional council).^[19]

How do you choose between open vs closed models? FinOps guidance frames this as a cost/performance/control tradeoff decision, balancing immediate capability and support against customization, privacy/control, and long-run cost structure; the CoE should standardize selection criteria and document tradeoffs.^[23]

How do procurement and vendor risk fit into the CoE? Evidence from insurance governance practice shows AI usage questions being added to procurement, vendor AI acceptable-use documentation being requested, and vendor AI use cases being reviewed through governance committees; the CoE typically defines these standards and workflows.^[31]

How do regulated industries integrate model risk management? Banking supervision guidance describes model risk management as including development/use discipline, validation, and governance with model inventories and effective challenge; CoE governance commonly operationalizes these controls for AI systems.^[12]

What metrics should the CoE own? A blended set: scaled-to-production rate, eval pass rates, incident rates, time-to-first-prompt/time-to-production, cost per token/inference, cost anomalies, and business value realization — metrics directly reflected in FinOps AI guidance and in CoE measurement guidance.^[19]

How long does it take to stand up a credible CoE? Case evidence shows meaningful outcomes in months when the CoE is structured as a pipeline + governance model (e.g., dozens of ideas, multiple POCs, and initial deployments in ~120 days for an AI CoE). The exact timeline varies with maturity and constraints, but a quarter is often enough to establish foundations and first lighthouse wins.^[24]

Frequently Asked Questions

When is an AI CoE not necessary? If AI usage is narrow (one or two contained projects), the organization may not need a dedicated enterprise operating structure; however, once AI spans multiple teams and higher-risk contexts, governance and shared enablement become central. This inference is grounded in the risk-based framing of AI governance and the organizational scaling dynamics described in CoE and governance sources.^[10]

What does “evidence vs consultant claim” look like here? The evidence strongly supports that scalable governance requires (a) operational workflows, (b) automation/tooling, and (c) cross-functional accountability—not just policy documents. This is consistent across AI RMF lifecycle governance, telecom governance findings, and FinOps guidance emphasizing councils, ownership clarity, and granular telemetry.^[10]



References and further readings



1. IBM, "AI center of excellence." [Online]. Available: <https://www.ibm.com/think/topics/ai-center-of-excellence>
2. Altman Solon, "Generative AI Governance for Telecommunications." [Online]. Available: <https://www.altmansolon.com/hubfs/Generative-AI-Governance-Telcos/Generative-AI-Governance-for-Telecommunications.pdf>
3. OpenAI, "Evals." [Online]. Available: <https://developers.openai.com/api/docs/guides/evals/>
4. OWASP, "OWASP Top 10 for Large Language Model Applications." [Online]. Available: <https://owasp.org/www-project-top-10-for-large-language-model-applications/>
5. FinOps Foundation, "Optimizing GenAI Usage." [Online]. Available: <https://www.finops.org/wg/optimizing-genai-usage/>
6. European Commission, "Regulatory framework AI." [Online]. Available: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
7. Google Cloud, "MLOps: Continuous delivery and automation pipelines in machine learning." [Online]. Available: <https://docs.cloud.google.com/architecture/mlops-continuous-delivery-and-automation-pipelines-in-machine-learning>
8. FinOps Foundation, "Managing AI Value." [Online]. Available: <https://www.finops.org/wg/managing-ai-value-finops-practice-operations/>
9. ISO, "ISO/IEC 42001:2023." [Online]. Available: <https://www.iso.org/standard/42001>
10. NIST, "AI Risk Management Framework: 5-SEC-CORE." [Online]. Available: <https://airc.nist.gov/airmf-resources/airmf/5-sec-core/>
11. NIST, "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence." [Online]. Available: <https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence>
12. Federal Reserve, "SR 11-7: Guidance on Model Risk Management." [Online]. Available: <https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm>
13. NAIC, "Artificial Intelligence." [Online]. Available: <https://content.naic.org/insurance-topics/artificial-intelligence>
14. Google Cloud, "What is MLOps?" [Online]. Available: <https://cloud.google.com/discover/what-is-mlops>
15. Google Cloud, "What is LLMOps?" [Online]. Available: <https://cloud.google.com/discover/what-is-llmops>
16. OpenAI, "Evals repository." [Online]. Available: <https://github.com/openai/evals>
17. P. Liang et al., "Holistic Evaluation of Language Models," arXiv:2211.09110. [Online]. Available: <https://arxiv.org/abs/2211.09110>
18. Oracle, "AI Center of Excellence." [Online]. Available: <https://www.oracle.com/artificial-intelligence/ai-center-excellence/>
19. FinOps Foundation, "FinOps for AI Overview." [Online]. Available: <https://www.finops.org/wg/finops-for-ai-overview/>

References and further readings

20. FinOps Foundation, “GenAI FinOps: How Token Pricing Really Works.” [Online]. Available: <https://www.finops.org/wg/genai-finops-how-token-pricing-really-works/>
21. AWS, “Empowering manufacturing innovation: How AI/GenAI Centers of Excellence can drive modernization.” [Online]. Available: <https://aws.amazon.com/blogs/mt/empowering-manufacturing-innovation-how-ai-genai-centers-of-excellence-can-drive-modernization/>
22. FinOps Foundation, “How to build a Generative AI cost and usage tracker.” [Online]. Available: <https://www.finops.org/wg/how-to-build-a-generative-ai-cost-and-usage-tracker/>
23. FinOps Foundation, “How to forecast AI services costs in cloud.” [Online]. Available: <https://www.finops.org/wg/how-to-forecast-ai-services-costs-in-cloud/>
24. SEI, “Generative AI Center of Excellence.” [Online]. Available: <https://www.sei.com/case-study/generative-ai-center-of-excellence-gen-ai-coe-in-qsr-industry/>
25. Intel, “Building an AI Center of Excellence.” [Online]. Available: <https://cdrdv2-public.intel.com/760460/Building%20and%20AI%20Center%20of%20Excellence.pdf>
26. Microsoft, “Banco de la Ciudad de Buenos Aires — Microsoft 365 Copilot.” [Online]. Available: <https://www.microsoft.com/en/customers/story/24281-banco-de-la-ciudad-de-buenos-aires-microsoft-365-copilot>
27. Cloudera, “OTP Bank.” [Online]. Available: <https://www.cloudera.com/customers/otp-bank.html>
28. GSA, “Artificial intelligence delivers real results through GSA.” [Online]. Available: <https://www.gsa.gov/blog/2025/01/10/artificial-intelligence-delivers-real-results-through-gsa>
29. Cognizant, “Global bank processes documents 98% faster with generative AI.” [Online]. Available: <https://www.cognizant.com/us/en/case-studies/global-bank-processes-documents-98-percentage-faster-with-generative-ai>
30. CIO, “AI center of excellence: A new engine for driving business transformation.” [Online]. Available: <https://www.cio.com/article/194085/ai-center-of-excellence-a-new-engine-for-driving-business-transformation.html>
31. NAIC, “Health Survey Report.” [Online]. Available: <https://content.naic.org/sites/default/files/inline-files/Health%20Survey%20Report%20-%20FINAL%205.9.25.pdf>



Together, we build for boundless impact

Since 1999, UST has worked side by side with the world's best companies to make a powerful impact through transformation. Powered by technology, driven by AI, inspired by people, and led by our purpose, we partner with our clients from design to operation. Our AI-driven digital solutions, proprietary platforms, engineering, R&D, products, and innovation ecosystem turn core challenges into impactful, disruptive business outcomes. With deep industry knowledge and a future-ready mindset, we infuse expertise, innovation, and agility into our clients' organizations—delivering measurable value and positive lasting change for them, their customers, and communities around the world. Together, with 30,000+ employees in 35+ countries, we build for boundless impact—touching billions of lives in the process.

ust.com

© 2026 UST Global Inc.

Version 0202-20260701

U ■
S T