RJL SOLUTIONS

ENTERPRISE AI DECISION GUIDE

Choosing an Enterprise Platform for AI Agents

How to choose the employee assistant, the models, the cloud that serves them and the platform that runs your agents.

ASSISTANTS · MODELS · CLOUDS · AGENT PLATFORMS

Richard White
RJL Solutions · rjlsol.com

VERSION 2 · PUBLISHED EDITION
OCTOBER 2026

01 / EXECUTIVE DECISION

Make four decisions, in the right order.

Most AI platform debates stall for one reason. They ask one question where the buyer faces four: which assistant employees use, which models do the work, which contract and cloud serve those models, and which platform runs the agents.

There's no one best agent platform on the market. There's no one best model or one best assistant either. "Should we use ChatGPT or Claude?" combines a workforce decision with a model decision. "Should we use Bedrock, Azure or Google?" combines a model-access decision with an agent-runtime decision. Each has a different owner, a different rate of change and a different cost of reversal. So each deserves its own answer.

DECISION	THE QUESTION	RJL DEFAULT, OCTOBER 2026	COST OF REVERSAL
1 Workforce assistant	Which assistant do employees use every day?	Start with the assistant native to your productivity suite. Add ChatGPT Enterprise or Claude Enterprise for named groups only when a cohort test shows a measurable advantage.	Moderate: retraining, connectors, saved work
2 Models	Which models perform each task?	A small tiered portfolio chosen per task class by your own evaluation, with at least two frontier providers approved. No single-vendor model mandate.	Low if calls are abstracted; high if prompts and tools are tuned to one model
3 Model access and cloud	Through which contract, cloud and data path are models called?	Your strategic cloud's model catalog first. A direct vendor contract only for a model or feature you need that the cloud does not yet offer.	Moderate: commitments, data agreements, quotas
4 Agent and automation platform	Where do agents and workflows run and act?	Automate at the lowest tier that carries the risk. For production agents: one primary runtime, two or three approved native extensions and shared controls.	High: identity, integrations, operations

Decide in that order of governance, and in reverse order of permanence. Models change monthly, so keep Decision 2 reversible. The runtime and its integrations last for years. Decision 4 therefore deserves the hard gates and the production-shaped proof described later in this guide.

STOP DOING Stop three habits: buying an assistant to get a model, choosing a cloud to get one model, and letting each department license its own agent builder. Each one locks in the fastest-changing layer, or fragments the slowest-changing one.

02 / FOUR DECISIONS NOT ONE

Separate the layers before you compare vendors.

The same vendor name can show up in all four layers. Microsoft, Google, OpenAI and Anthropic each sell an assistant, models, model access and an agent runtime. Buy each layer for its own reasons.

DECISION	WHAT YOU CHOOSE AMONG	PACE AND OWNER
1 Workforce assistant	Microsoft 365 Copilot, Gemini Enterprise, ChatGPT Enterprise, Claude Enterprise	**Annual licence cycle** CIO and business
2 Models	GPT, Claude, Gemini and open-weight families, chosen per task class	**Changes monthly** Product, engineering, model risk
3 Model access and cloud	Vendor API, Bedrock, Foundry, Google Cloud or self-hosted	**Contract term** CTO, security, procurement
4 Agent and automation platform	Assistant features, low-code, cloud, native or orchestration runtimes	**Multi-year** Architecture, process owners

SLOWER TO CHANGE, COSTLIER TO REVERSE

Figure 1. The enterprise AI decision stack. Lower layers change more slowly and cost more to reverse.

THE QUESTION PEOPLE ASK	THE DECISION IT ACTUALLY IS
Should we roll out ChatGPT, Claude, Copilot or Gemini?	Decision 1. Choose for the suite, the content employees need, admin controls and adoption, not for benchmark scores.
Is GPT, Claude or Gemini the best model?	Decision 2. There is no single answer. Rank models per task class on your own cases, cost and latency.
Should we use AWS Bedrock, Azure or Google?	Decision 3 for model access and Decision 4 for agent runtime. Your strategic cloud usually answers both, unless a needed model is unavailable there.
Which platform should we automate on?	Decision 4. Choose the build tier first, then the platform closest to the workflow's data, authority and users.

ONE OWNER PER LAYER Give each decision an accountable owner. The CIO and business leaders own the assistant. Product and engineering, with model risk, own the models. The CTO, with security and procurement, owns access. Enterprise architecture, with process owners, owns the runtime.

Choose the assistant for the suite, the content and the controls.

Four enterprise assistants now compete for every knowledge worker. Each is also turning into an agent platform: it can act across connected applications, run scheduled work and produce files. Treat that convergence as a governance question. It's more than a feature.

ASSISTANT	CHOOSE IT WHEN	DISTINCTIVE STRENGTH (DOCUMENTED OCTOBER 2026)	WATCH CLOSELY
Microsoft 365 Copilot	Microsoft 365, Teams and SharePoint are the system of work and Entra and Purview govern content.	Works inside Office and Teams with tenant permissions. Offers OpenAI and Anthropic models: Claude is selectable in Copilot Chat, Researcher, agentic editing in Excel and PowerPoint, and Copilot Studio.[43]	Claude requests are processed on Anthropic infrastructure outside Azure. Anthropic was enabled by default in commercial tenants from January 2026 and disabled in the EU, EFTA and UK.[43] Decide deliberately.
Gemini Enterprise	Google Workspace, BigQuery and Google Cloud are the center of work.	Deep Research over access-controlled enterprise data, Notebook, a no-code Workflow Builder, Agent Studio and ADK, admin control of the agent gallery, and Agent2Agent interoperability.[31]	Some agents remain preview, for example Data Insights.[31] Confirm edition entitlements; Microsoft-suite customers should test connector depth.
ChatGPT Enterprise, with ChatGPT Work and Codex	Broad general-purpose demand, a mixed or Google-plus-Microsoft estate, and appetite for agentic task completion.	ChatGPT Work, launched 9 July 2026, runs multi-step projects across connected apps and files, with scheduled tasks and desktop computer use. Admins control tools and actions, with a Compliance API and spend limits.[24] Enterprise supports data residency in ten regions.[25]	Overlap with suite licences. Agent actions need the same approval design as any agent. Verify connector permissions and seat plus usage economics.
Claude Enterprise	Document-heavy analysis and writing, regulated knowledge work and engineering teams using Claude Code.	Since 23 October 2026, chat and Cowork are one experience; cloud sessions run in isolated virtual machines with network egress off by default. Admins set model entitlements by role and spend limits by organization, group and member.[29]	Usage-based pricing needs budget controls. Confirm connectors to your suite, retention and residency terms, and which features are GA in your plan.

04 / SELECTING THE ASSISTANT

Follow the suite, then prove the exception.

Your default assistant is the one that already sees your documents, mail, meetings and permissions. Measured work justifies a second assistant. Preference and popularity don't.

The RJL rule

- Microsoft 365 organizations start with Microsoft 365 Copilot; Google Workspace organizations start with Gemini Enterprise. Permissions, content and channels are already governed there.
- Add ChatGPT Enterprise or Claude Enterprise for named populations, such as engineering, research, analytics or legal drafting, when a four-week cohort test shows more accepted work per person.
- Don't buy an assistant to get a model. Copilot already offers OpenAI and Anthropic models, and both independent assistants restrict models by role. Model access for applications is Decision 3.
- Treat unsanctioned personal use as a demand signal. Provide an approved option quickly, then rationalize with evidence.

Controls to require before rollout

CONTROL AREA	EVIDENCE TO OBTAIN
Identity	SSO, SCIM provisioning, role-based access and model entitlements by group.
Data	No training on business data by default, retention settings, residency region, customer-managed keys and abuse-monitoring terms.
Connectors	Which systems connect, whether results respect source permissions, and who approves each connector.
Actions	Which agent actions are allowed, which require approval, and whether a compliance or audit API records them.
Spend	Seat versus usage pricing, organization and group limits, and reporting by team.
Exit	Export of conversations, projects, custom assistants, skills and audit records.

MEASURE THE COHORT, NOT THE SURVEY Run each candidate with matched groups doing real work for four weeks. Compare tasks completed and accepted, rework, time to first useful output and incidents. Satisfaction scores alone reward whoever gives the most fluent demo.

05 / DECISION 2 MODELS

Choose models per task class, not per vendor.

Frontier models leapfrog one another within months. The decision that lasts is a method. Classify the work, shortlist candidates, test them on your own cases and keep the choice reversible.

TASK CLASS	WHAT DECIDES QUALITY	TIER	SHORTLIST TO TEST, OCTOBER 2026
Long-running agentic work, coding and multi-step tool use	Tool-call accuracy, recovery from errors, instruction persistence and long-context fidelity	Frontier	Claude Opus 5.5; GPT-6 Astra or GPT-6.1 Sol; Gemini 4 Argon[26,23,30]
Reasoning over documents and data	Grounding, citation accuracy and refusal to assert unsupported facts	Frontier or mid	Claude Opus 5.5 or Sonnet 5.5; GPT-6.1 Sol; Gemini 4 Argon
High-volume extraction, classification and routing	Schema adherence, consistency, latency and unit cost	Fast	Claude Haiku 5.5; GPT-6 Luna; Gemini 3.8 Flash or 3.5 Flash-Lite; a tested open-weight model
Real-time voice	Latency, interruption handling and tool use while speaking	Realtime	GPT-Live 1 or GPT-Realtime-2.1; Gemini 3.8 Live[23,30]
Visual reasoning and computer use	Accuracy on screenshots and documents, safe action on interfaces	Mid or frontier	GPT-6.1 Terra; Claude Opus 5.5; Gemini 3.8 Flash
Sovereign, air-gapped or on-premises	Where the weights run, licence terms and support	Open-weight or on-premises	gpt-oss, Llama, Mistral or Qwen self-hosted; Gemini on Google Distributed Cloud, air-gapped[37,32]

The shortlists reflect each vendor's documented positioning: Anthropic positions Opus 5.5 for long-running agentic coding and knowledge work and Haiku 5.5 for classification, extraction and routing; OpenAI positions GPT-6 Luna for focused high-volume tasks and GPT-6.1 Terra for visual reasoning and computer use.[26,23] They are candidates to test, not results.

RJL STARTING DEFAULT

Approve one frontier model from each of two providers your strategic cloud serves, plus one fast model for volume work. Route by task class. Re-run the evaluation each quarter and on every major release. Two providers give you a tested fallback and real bargaining power.

06 / MODEL LANDSCAPE

Know who sells which model, and where.

Availability broadened sharply in 2026. OpenAI models reached Amazon Bedrock and, in preview, Google Cloud. Claude runs on all three hyperscalers. Check this table again before you sign anything; it describes 29 October 2026.

PROVIDER	CURRENT FAMILIES	WHERE YOU CAN CALL THEM
OpenAI	GPT-6 Astra (flagship), GPT-6.1 Sol, GPT-6 Luna and GPT-6.1 Terra, all with about one million tokens of context; GPT-Live and GPT-Realtime for voice; gpt-oss open weights.[23]	OpenAI API. Microsoft Foundry: GPT-6 Astra GA 3 September, GPT-6 Luna 22 September and GPT-6.1 Sol 29 September 2026.[40] Amazon Bedrock: GA since 1 June 2026, with GPT-6 Astra GA 8 September 2026.[34,35] Google Cloud: private preview from 21 October 2026.[33]
Anthropic	Claude Opus 5.5, Sonnet 5.5 and Haiku 5.5, plus Fable 5.1 for demanding long-horizon work; one million tokens of context across the four.[26]	Claude API, Amazon Bedrock, Claude Platform on AWS, Microsoft Foundry, and Google Cloud in Google-managed or Anthropic-managed form.[27] In Foundry, Fable is preview, and some versions run only on Anthropic infrastructure.[39]
Google	Gemini 4 Argon, GA 13 October 2026; Gemini 3.8 Flash, GA 2 September 2026; Flash-Lite, Live and speech models; Gemma open weights.[30]	Gemini API, Google Cloud's Gemini Enterprise Agent Platform, and Google Distributed Cloud, where the air-gapped offering is GA.[32] Gemini is sold through Google's own services.
Open-weight families	gpt-oss, Llama, Mistral, Qwen, DeepSeek and Gemma, among others.	Hyperscaler catalogs, including Bedrock in AWS GovCloud for gpt-oss,[37] or your own infrastructure.

THIS TABLE WILL BE WRONG WITHIN A QUARTER

Keep an approved-model register with model and version, endpoint, region, data terms, release stage, retirement date, owner and the evaluation that justified approval. Your teams should treat the register as the source of truth. This page isn't.

07 / HOW TO PICK A MODEL

Run your own evaluation; treat leaderboards as a shortlist.

Public benchmarks measure someone else's task. A two-week bake-off on your own cases settles most model decisions. It also leaves behind a test harness you'll reuse at every release.

STEP	WHAT TO DO
1 Define acceptance	For each task class, write what a correct, usable output is and who decides. Name the critical errors that disqualify a model.
2 Build a golden set	Collect 100 to 300 real, de-identified cases per task class, including edge cases, adversarial inputs, permission tests and cases where the right answer is to refuse or escalate.
3 Grade consistently	Use a rubric calibrated with human reviewers. Introduce model-based grading only after its agreement with humans is measured.
4 Test on the real endpoint	Run candidates through the cloud, region and tools you will use in production. Feature support can differ by platform for the same model.[27]
5 Decide and pin	Choose a primary and a fallback per task class, pin versions, record the evidence and set the next review date.

Illustrative model scorecard

CRITERION	WEIGHT	MEASURE
Accepted-output rate	30%	Share of outputs accepted without material edit
Critical error rate	20%	Disqualifying errors per hundred cases, including unsupported claims and unsafe actions
Cost per accepted output	15%	All tokens, retries and review time divided by accepted outputs
Tool and format reliability	15%	Correct tool calls, valid schemas, recovery after a failed call
Latency	10%	95th percentile time to a complete answer at expected load
Operational fit	10%	Availability in your cloud and region, data terms, release stage and retirement date

LIFECYCLE IS A SELECTION CRITERION

Providers retire versions on published schedules; Anthropic lists retirement for Opus 5.5 as not sooner than 22 September 2027, and Microsoft publishes a Foundry retirement schedule.[27,41] Budget a re-evaluation before every forced migration.

08 / ROUTING AND OPEN WEIGHTS

Reserve frontier models for the work that needs them.

Most enterprise volume is narrow, repetitive work that a fast model handles well. Routing by task class usually cuts model cost more than any price negotiation.

Managed routers can choose among models within a family, for example Amazon Bedrock Intelligent Prompt Routing and the Microsoft Foundry model router.[38,40] They're useful. Your own routing by task class gives clearer evidence, though: every request should record the task class, model, version and endpoint used.

When open-weight or self-hosted models are justified

DRIVER	OPEN-WEIGHT FITS WHEN	A MANAGED PROPRIETARY API FITS WHEN
Sovereignty	Data must stay in your facility or an air-gapped network.	A cloud region and contract terms meet the requirement.
Cost at scale	Volume is very high, steady and narrow.	Volume is variable, modest or mixed.
Customization	Fine-tuning on proprietary data materially improves results.	Prompting and retrieval reach the acceptance bar.
Capability	A smaller model passes your evaluation for the task.	The task needs frontier reasoning or long agentic work.
Operations	Your team can run GPU capacity, patching, evaluation and safety controls.	You prefer the provider to operate and secure inference.

ABSTRACTION IS CHEAP INSURANCE Call models through a gateway, or a thin internal interface that records model and version on every request. Keep prompts, tool definitions and evaluation sets in version control. Adopt a vendor-specific feature on purpose, with a named owner. Don't let it creep into every workflow by accident.

09 / DECISION 3 MODEL ACCESS

The same model can arrive through four different contracts.

Where you call a model changes who processes your data, which terms apply, which features you get, how quotas work and where the bill lands. That holds even when the model name is identical.

ACCESS PATH	WHO OPERATES INFERENCE	WHAT CHANGES	FITS WHEN
Direct from the model vendor (OpenAI API, Claude API, Gemini API)	The model vendor	Newest models and features first; the vendor's data terms; separate procurement and security review.	You need a capability that is only available directly, or the vendor's own agent services.
Hyperscaler-operated (Bedrock; Foundry Azure-hosted; Claude on Google Cloud, Google-managed)	The cloud provider	Cloud identity, private networking, regional endpoints and consolidated billing. AWS states that Bedrock model providers have no access to customer prompts and completions.[36]	Your strategic cloud must govern data, network and operations.
Vendor-operated, bought through the cloud (Foundry Anthropic-hosted; Claude on Google Cloud, Anthropic-managed; Claude Platform on AWS)	The model vendor	Cloud procurement and billing, but processing on vendor infrastructure. Microsoft states Anthropic-hosted Claude runs outside Azure.[39] The Anthropic-managed Google offering uses a global location.[28]	You want the vendor's full feature set with cloud procurement.
Self-hosted open weights	Your organization	Full control and full responsibility for capacity, security, evaluation and support.	Sovereignty, air gaps or very high steady volume.

READ THE ENDPOINT, NOT THE LOGO

For each approved model, record who processes the data, the region, retention and abuse-monitoring terms, the features that endpoint supports and the rate limits. The same model name can carry different answers on each path.

Bedrock, Foundry or Google: let the strategic cloud lead.

For model access, the three hyperscalers have converged on multi-vendor catalogs. What still differs is which models each serves, how they're hosted and how well the catalog fits your identity, data and operations.

CLOUD	MODEL CATALOG, OCTOBER 2026	CHOOSE IT WHEN	WATCH CLOSELY
Amazon Bedrock	Claude; OpenAI GPT-5 and GPT-6 families (GA); open-weight models including gpt-oss.[27,35,37]	AWS is the strategic cloud and you want several frontier vendors behind one IAM, network and billing boundary.	Feature timing versus vendor direct, regional and cross-region inference settings, quotas, and the separate AgentCore runtime decision.
Microsoft Foundry	OpenAI GPT-6 family (GA); Claude in Azure-hosted or Anthropic-hosted form; partner and open models.[40,39]	Azure, Entra and Microsoft 365 are the estate, and Copilot or Teams is the main channel.	The hosting option decides whether data leaves Azure. Data Zone deployment covers only listed Azure-hosted Claude models, and Foundry adds no built-in content filter for Claude at deployment.[39]
Google Cloud, Gemini Enterprise Agent Platform	Gemini; Claude in Google-managed or Anthropic-managed form; OpenAI GPT-6 in private preview; open models including Gemma.[30,28,33]	Workspace and BigQuery are central, Gemini multimodal capability matters, or air-gapped Gemini is required.	OpenAI models are not yet GA. Product names changed in 2026, and regional endpoint options differ by hosting form.[28]

Don't set up a second strategic cloud just for one model or agent feature. The exception is when the differentiated value survives the cost of identity federation, networking, data movement, security operations, procurement, support and a second engineering skill base. You can often separate model access from the runtime decision.

A MODEL IS RARELY A REASON FOR A SECOND CLOUD

As of 29 October 2026, Claude is GA in all three hyperscalers and OpenAI frontier models are GA in two, with Google in preview. Gemini remains the main model reason to work in Google Cloud, and Gemma open weights travel elsewhere. Confirm whether model spend, including marketplace purchases, counts toward your existing cloud commitments.

Automate at the lowest tier that can carry the risk.

Automation doesn't start with an agent platform. An assistant feature, or a deterministic workflow with one model step, serves many workflows better. Step up a tier only when the work requires it.

TIER	USE IT FOR	EXAMPLES, OCTOBER 2026	STEP UP WHEN
0 Assistant features	Personal and team productivity, reusable instructions and light scheduled work.	ChatGPT projects and Work scheduled tasks;[24] Claude projects, skills and connectors; agents in Microsoft 365 Copilot; Gemini Enterprise Workflow Builder.[31]	The work is shared, touches a system of record, or needs approval and audit.
1 Low-code builders and integration platforms	Departmental workflows with known steps, standard connectors and moderate risk.	Microsoft Copilot Studio and Power Automate;[42] Salesforce Agentforce builder; Workato, Zapier, Make and n8n, which can be self-hosted.[44]	You need custom logic, cross-system writes, scale or a full engineering lifecycle.
2 Pro-code agent runtimes	Cross-system agents, custom tools and production service levels.	Bedrock AgentCore; Foundry Agent Service; Gemini Enterprise Agent Platform; OpenAI Agents SDK, Claude Agent SDK and LangGraph.	The process runs for days, coordinates people, robots and systems, or needs durable cases.
3 Process orchestration	Long-running cases with human tasks, robots, rules and agents.	UiPath Maestro, IBM watsonx Orchestrate and ServiceNow workflow.	Rarely: this tier usually coordinates the others.

Use deterministic steps wherever the path is known. Insert a model only where judgment is needed: reading an unstructured document, drafting, classifying or choosing among defined options. A workflow with one model step is easier to test, cheaper to run and simpler to audit than an agent that decides every step.

TIERS 0 AND 1 STILL NEED GOVERNANCE Register every shared automation with an owner, approved connectors, data-loss controls and a retirement date. Low-code makes building easy. It doesn't make ownership optional.

Map nine layers before scoring a runtime.

"Agent platform" is too broad a phrase to support a serious buying decision. Buyers often compare a model API, a software development kit, a managed runtime, an application feature and an orchestration product as though they were substitutes. They aren't.

LAYER	DECISION
Experience	Where people or systems invoke the agent: web, mobile, Teams, Slack, CRM, ERP, portal, API, voice or event.
Agent definition	Instructions, role, tools, skills, routing, memory policy and release version.
Runtime	The service that manages turns, state, tool execution, retries, concurrency and scaling.
Models	The reasoning, language, vision, speech and specialist models available through a catalog or gateway (Decisions 2 and 3).
Tools and actions	The APIs, functions, MCP servers, RPA robots, code sandboxes, browsers and application skills the agent can use.
Context and memory	Session state, user context, retrieval, knowledge, semantic models and longer-lived memory.
Identity and policy	Human delegation, workload identity, least privilege, approvals, guardrails, data policy and secrets.
Evaluation and operations	Test sets, traces, quality and safety metrics, alerts, incident response, versioning and rollback.
Registry and governance	Inventory, ownership, risk tier, lineage, evidence, exceptions and lifecycle status across runtimes.

A vendor can be excellent at one layer and deliberately thin at another. That's no flaw if the architecture assigns the missing responsibility. It becomes a serious risk when someone discovers the gap only after a successful demonstration.

The center-of-gravity test

- Where does the authoritative data already live?
- Which system grants the user and the agent permission to act?
- Where must the transaction, case, decision or artifact be recorded?
- Which team will own the runtime at 2 a.m. when a tool, model or workflow fails?
- Which channel must users adopt, and which platform already governs that experience?

If four of those answers point to the same platform, make it a finalist. If the answers span several platforms, the general-purpose runtime and the control plane matter more than any application's embedded agent builder.

Five categories that happen to share one word.

Category boundaries are blurring. Still ask which layer is native, which is integrated and which stays the customer's responsibility.

CATEGORY	EXAMPLES	STRONGEST FIT	COMMON BOUNDARY
Cloud agent runtimes	Microsoft Foundry, AWS Bedrock AgentCore, Google Gemini Enterprise Agent Platform	Cross-system agents, custom code, enterprise cloud controls, multiple channels	Can become infrastructure-heavy; native business semantics still require integration
Application-native platforms	Salesforce Agentforce, SAP Joule Studio, ServiceNow, Oracle AI Agent Studio	Work that mostly begins, reasons and acts inside one enterprise application	Cross-domain work may expose platform boundaries, licensing and duplicated governance
Data-native platforms	Snowflake Cortex Agents, Databricks Agent Bricks	Analytical agents and workflows centered on governed data and ML engineering	Usually not the best owner of every transactional workflow or user experience
Independent frameworks and managed agent services	OpenAI Agents SDK, Claude Agent SDK and Managed Agents, LangGraph and LangSmith	Custom engineering, model-native capabilities, portability, specialist agent patterns	The enterprise must still own identity, deployment, process durability, integration and controls not supplied by the service
Enterprise orchestration and control	IBM watsonx Orchestrate, UiPath Maestro and adjacent automation platforms	Coordinating agents, robots, deterministic workflows, cases and people	May complement rather than replace a runtime; confirm feature maturity and integration depth

Cloud vendors add business connectors, application vendors add custom agent runtimes, data vendors add actions and orchestration vendors add builders. Assistants now add agents too: ChatGPT Work, Claude Enterprise and Gemini Enterprise each run multi-step work across connected applications.[24,29,31] Decide which of these is a sanctioned execution path for production work, and which remains personal productivity.

Pick the runtime that inherits your controls.

The three hyperscalers are the natural finalists for agents that cross systems, need custom code or must inherit existing cloud controls. Choose the platform that fits identity, network, data, operational tooling, engineering capability and support. The most appealing demonstration is the wrong criterion.

PLATFORM	CHOOSE IT WHEN	WHY IT CAN WIN	WATCH CLOSELY
Microsoft Foundry	Organizations standardized on Azure, Entra, Microsoft 365, Teams and Azure data services; regulated deployments that benefit from Azure networking, identity and policy integration.	Prompt agents for configuration-led work, hosted agents for custom frameworks and code, and distribution into Microsoft channels. Agent Service added GA routines, tool search and Agent-to-Agent support in September 2026.[2,40]	Don't assume Microsoft 365 distribution makes every cross-system action simple. Verify delegated access, connector depth, regional availability and the boundary between Foundry, Copilot Studio and application teams.
AWS Bedrock AgentCore	AWS-centric engineering organizations that want modular runtime, identity, memory, gateway, browser, code execution, observability, evaluation, registry and policy services with framework and model choice.	AgentCore is a set of composable services and supports popular frameworks and models inside or outside Amazon Bedrock.[3]	Distinguish AgentCore from Bedrock Agents Classic, which AWS places in maintenance mode and closed to new customers after 30 July 2026.[4] Validate the release stage of each module used.
Google Gemini Enterprise Agent Platform	Organizations centered on Google Cloud, BigQuery, Workspace, Gemini models and Google search and data capabilities.	Agent Builder and the managed runtime sit inside the broader platform, with development, deployment, governance, sessions, memory and enterprise integration.[5]	Naming changed in 2026. Confirm current service names, regional support, quotas, identity patterns and migration implications. Older Vertex AI terms may mislead you.[6]

RUNTIME AND MODEL CATALOG ARE SEPARATE PURCHASES

An AgentCore agent can call Claude or GPT models through Bedrock. A Foundry agent can call GPT or Claude models. A Google agent can call Gemini or Claude. Choose the runtime for controls and operations, then approve models through Decision 2.

Use native agents where the work truly lives.

Application-native platforms can reach production quickly, because they already understand the application's objects, permissions, workflows, channels and administration. Their edge isn't better models. It's a shorter distance from reasoning to governed action.

PLATFORM	CHOOSE IT WHEN	WHY IT CAN WIN	WATCH CLOSELY
Salesforce Agentforce	Customer, sales, service, marketing, commerce or Slack work in which Salesforce objects, flows, permissions and Data 360 are the operational center.	Grounds agents in Salesforce data and embeds them in Salesforce channels under the platform's security and automation model.[7,8]	It isn't automatically the best general runtime for finance, supply chain, engineering or data workflows outside Salesforce. Test cross-cloud access, identity propagation, action governance and consumption pricing.
SAP Joule Studio	Finance, procurement, supply chain, manufacturing and HR processes whose transactions and semantics live in SAP.	SAP documents S/4HANA access through its agent gateway, existing authorization, tool and MCP support and routing through Joule.[9,10]	Some capabilities are early adopter or limited availability. Confirm tenant, region, edition, entitlements, promotion controls and the role of SAP BTP.
ServiceNow AI Platform	IT service, employee workflows, customer service, security operations and other case domains already governed in ServiceNow.	AI Agent Studio and AI Agent Orchestrator connect agents to Now Platform workflows, skills, scripts and knowledge, with AI Control Tower as a governance layer.[11]	Avoid making ServiceNow the execution owner for unrelated product or data workloads merely because it can call them.
Oracle AI Agent Studio	Core finance, HCM, SCM and other processes in Oracle Fusion Cloud Applications.	Native access to Fusion knowledge, tools, APIs, workflow, testing, monitoring, promotion and application security context.[12]	Evaluate cross-system tool access, extensibility, release cadence and whether the process is truly Fusion-centered.

NATIVE DOES NOT MEAN UNIVERSAL

Use a native platform when the workflow truly lives there. A department owning licences is no reason. A native agent that keeps exporting data, calling an external orchestrator and acting in several other systems may be running in the wrong place.

Complementary capabilities, not one category.

Data-native platforms shine when governed data and analytical context dominate. Independent frameworks give engineering teams control. Orchestration platforms govern long-running work across many kinds of actors.

PLATFORM	BEST FIT	BOUNDARY TO TEST
Snowflake Cortex Agents	Governed Snowflake data, semantic models, Cortex Analyst, Cortex Search and Snowflake privileges at the center. Strong for analytical assistants and data-grounded decision support.[13]	The boundary between analysis and consequential transactions. A separate application or orchestration layer may be needed for multi-system actions.
Databricks Agent Bricks	Lakehouse-centered teams wanting custom frameworks, managed app runtime, MLflow tracing, Unity Catalog controls and close integration with ML delivery.[14]	More engineering ownership than a low-code platform. Confirm production identity, resource declaration, scaling, endpoint and support patterns.
OpenAI platform and Agents SDK	Direct access to OpenAI models, hosted tools, MCP, guardrails, handoffs, sessions, tracing, realtime voice and a lightweight SDK.[15,16]	It's an agent framework and model service. It doesn't replace your landing zone, workload identity, network boundary, durable process or cross-platform governance.
Anthropic Claude Agent SDK and Managed Agents	Claude-centered custom agents, especially tool-intensive and workspace-oriented work. Anthropic separates self-operated SDK agents from managed agents that package prompts, tools, MCP servers, skills, sessions and environments.[17,18]	At our 27 October review, the documentation still called the Agents, Sessions and Environments APIs beta. Confirm terms, data handling, identity federation, regional support and operating controls for the exact service path.
LangGraph and LangSmith	Explicit stateful orchestration, framework flexibility, durable execution and a deployable agent server with tracing and evaluation.[19]	Cloud, identity, data and operational responsibilities remain yours. Review cloud, hybrid and self-hosted data-plane choices.
IBM watsonx Orchestrate	Hybrid enterprises seeking catalog, orchestration, credentials, observability, governance and control-plane capabilities across agents and applications.[20]	Availability varies by deployment and release. Validate which governance and multi-agent functions are GA and how deeply they govern third-party runtimes.
UiPath Maestro	End-to-end processes combining agents, RPA robots, APIs, deterministic workflow, cases and human approvals, especially where a UiPath estate exists.[21]	Maestro is GA, while some newer Flow, conversational, case-agent and evaluation capabilities remain preview.[22]

Standardize the core, approve the edges.

The target state spans all four decisions: a suite-aligned assistant, an approved model register, a governed access path, one primary runtime and a small number of native extensions under common controls.

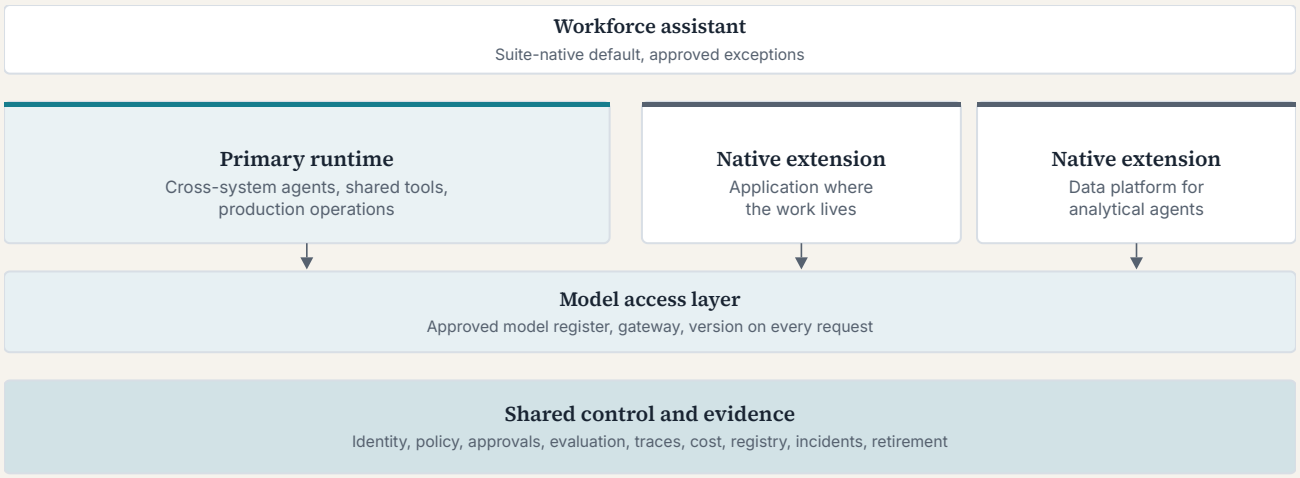


Figure 2. Recommended portfolio. One primary runtime and approved native extensions share one control and evidence pattern.

COMPONENT	WHAT IT IS FOR
Primary runtime	The default place for cross-system agents, reusable tools, shared engineering patterns and production operations, aligned to the strategic cloud, identity provider, network, delivery platform and support model.
Approved native extensions	Application- and data-native agents allowed where they reduce integration, preserve native permissions or improve the user experience. Each registers owner, risk tier, evaluation evidence, cost and telemetry.
Model access layer	A gateway or platform catalog that serves the approved models, records model and version per request, and applies rate, cost and data policies.
Shared control and evidence	Unique agent identity, named owner, approved tools, policy decisions, evaluation results, end-to-end traces, cost, incidents, versions and retirement status. You don’t need one product on day one. You do need common requirements.

Eliminate before you rank.

Weighted scoring creates false precision when a platform fails a mandatory control. Treat these as elimination gates for the named workload. Record pass, conditional pass or fail, with evidence and an accountable owner.

GATE	EVIDENCE REQUIRED
Identity	Can the platform represent the user, the agent and the service separately, and enforce delegated authority with short-lived credentials and no shared user secrets?
Data and residency	Can every model, memory store, tool and trace meet location, retention, encryption, private connectivity, logging and data-use requirements?
Model access path	For each approved model on this platform, who processes the data, in which region, under which terms, at which release stage?
Tool authorization	Can each tool call be constrained by role, purpose, resource, environment and transaction value, with approval required for dangerous tools?
Evidence	Can you reconstruct context, model and version, policy decision, tool calls, outputs, approvals, data access and final action under one request identifier?
Reliability	Are timeout, retry, idempotency, compensation, checkpoint, recovery, concurrency and regional failure patterns suitable for the process?
Lifecycle	Can teams test, version, promote, compare, roll back, retire and inventory agents and models within existing software and model risk processes?
Commercial and legal	Do terms cover data use, intellectual property, subprocessors, audit, support, incident notice, service levels, portability and termination?

A CONDITIONAL PASS MUST EXPIRE A roadmap promise isn’t a production control. If a finalist gets a conditional pass, document the compensating control, owner, cost, delivery date and the date you must revisit the decision.

Score only the finalists, with evidence.

Rate each factor from zero to five, multiply by the weight and attach evidence. These weights suit a cross-system enterprise agent. Change them before vendors demonstrate. Never after the results are known.

FACTOR	WEIGHT	WHAT TO TEST
Workflow and channel fit	20%	How directly the platform supports the named process, users, channels, latency and action model
Identity and authorization	15%	User delegation, workload identity, least privilege, approvals, secrets and policy enforcement
Data and context	15%	Authoritative sources, retrieval quality, semantics, memory, lineage, residency and data controls
Tools and integration	15%	API, event, application, MCP, RPA, browser, code and transaction support
Operations and reliability	10%	Deployment, scaling, retries, checkpoints, rollback, support, incident response and regional resilience
Governance and evidence	10%	Registry, ownership, evaluation, traces, risk controls, audit evidence and policy reporting
Model choice and access	5%	Approved models available on GA endpoints in your region, version pinning and fallback
Portability and standards	5%	Export, open protocols, framework choice, data portability and exit effort
Economics	5%	Cost per successful business outcome, including platform, model, tools, data, people and controls

Version 2 replaces the former engineering-fit factor with model choice and access; assess engineering skills and delivery tooling inside operations and reliability.

CONFIDENCE MATTERS A high score built on marketing material is weaker than a lower score demonstrated in your own environment. Add a confidence rating. Subtract for unresolved dependencies, preview features, manual operations or unpriced consumption.

Starting stacks for common situations.

These are dated starting hypotheses for October 2026. They aren't universal prescriptions. Each names an assistant, models and access path, and an agent runtime. The proof should confirm the boundary and show where a second runtime is justified.

SCENARIO	LIKELY STARTING STACK	WHAT DECIDES THE OUTCOME
Regulated insurer or bank on Azure and Microsoft 365	Assistant: Microsoft 365 Copilot, with Anthropic models enabled only after data-path review. Models: GPT-6 and Claude through Foundry, Azure-hosted where data must stay in Azure. Runtime: Foundry Agent Service; native agents where work is concentrated.	Entra identity, network controls, evidence, human approval, core-system integration and model risk lifecycle
AWS-centric, engineering-led enterprise	Assistant: suite-native, plus Claude Enterprise or ChatGPT Enterprise for engineering and analysts. Models: Claude, GPT and an open-weight fast model through Bedrock. Runtime: Bedrock AgentCore.	IAM delegation, AgentCore module release stages, Bedrock quotas and regions, and evidence export
Google Workspace and BigQuery organization	Assistant: Gemini Enterprise. Models: Gemini plus Claude through Google Cloud. Runtime: Gemini Enterprise Agent Platform.	BigQuery semantics, edition entitlements, preview features and identity federation for non-Google systems
Sales and service centered on Salesforce	Assistant: suite-native. Models: platform-provided within Agentforce; approved cloud models for specialist services. Runtime: Agentforce first; a cloud runtime for cross-domain services.	Salesforce permission behavior, Data 360 grounding, channel fit, cross-system actions and consumption economics
SAP finance, procurement or supply chain	Assistant: suite-native. Models: through Joule and SAP BTP for SAP work. Runtime: Joule Studio and BTP first; the general cloud runtime for non-SAP domains.	SAP authorization, business object semantics, tenant availability, transports, approvals and non-SAP integration

Starting stacks for common situations.

The second set covers service, analytics, lean and sovereign situations. Revisit every recommendation at each quarterly model review.

SCENARIO	LIKELY STARTING STACK	WHAT DECIDES THE OUTCOME
IT, employee or customer service workflow	Assistant: suite-native. Models: platform-provided in ServiceNow; approved cloud models for specialist agents. Runtime: ServiceNow for the workflow and case layer; an external runtime only as needed.	Knowledge quality, tool permissions, workflow ownership, AI Control Tower scope and incident evidence
Analytics and decision support on Snowflake or Databricks	Assistant: suite-native. Models: the platform catalog, chosen per task by evaluation. Runtime: Cortex Agents or Agent Bricks; separate orchestration for consequential transactions.	Semantic quality, role and execution context, citations, evaluation and handoff to action systems
Lean PE-backed portfolio company	Assistant: one, aligned to the suite. Models: the strategic cloud or one vendor API. Build tiers 0 and 1 first; one lightweight SDK only for a differentiated workflow.	Time to value, managed operations, vendor count, repeatability across portfolio companies, unit economics and exit path
Sovereign, defense or air-gapped workload	Assistant: only a sovereign or on-premises edition approved for the classification. Models: self-hosted open weights or Gemini on Google Distributed Cloud, air-gapped. Runtime: self-managed framework inside the boundary.	Accreditation, weight licensing, capacity, patching, evaluation and safety controls without provider support
Multi-cloud enterprise with several agent platforms	Assistant: one primary, one approved exception. Models: an approved register spanning two providers through a gateway. Runtime: a primary cloud runtime plus approved native platforms.	Inventory coverage, federated ownership, evidence normalization, identity consistency and platform sprawl

Prove the decision, not the demo.

The proof should resemble production closely enough to test the buying decision, without trying to cover full scope. Prove the architecture, control path, operating model and economics with one thin vertical slice. Use it to choose models as well as the runtime.

TIMING	OBJECTIVE	WORK
Week 0	Name the decision	Choose one production-shaped workflow, business owner, risk owner, finalists, gates, score weights, and baseline cost and performance.
Week 1	Establish the control path	Configure identity, data access, tools, logging, approvals, test data and environments before optimizing prompts. Build the golden set.
Week 2	Build the thin vertical slice	Complete one end-to-end path with the real channel, representative data, at least one real action and a human handoff.
Week 3	Evaluate and break it	Run at least two models through the same harness. Test normal, edge, adversarial, permission, stale-data, tool-failure, latency and recovery cases.
Week 4	Price the outcome	Measure cost per completed case or decision, including integration, review, rework, platform, support and control effort, for each model tested.
Week 5	Make the portfolio decision	Choose, reject or extend with evidence. Record platform boundaries, approved models, exceptions, production gaps and the next two agents that would validate reuse.

Minimum exit evidence

- Named business, technical, risk, data and operations owners
- End-to-end trace from user request through policy, model, tools, data, approval and action
- Evaluation results for quality, safety, permissions, recovery, latency and cost, per model
- Production gap register with cost, owner, dependency and delivery date
- Architecture decision record stating where this platform and these models are the default, and where they are not

Price the outcome, not the token.

Compare the cost of a completed business outcome at the required quality and control level. The denominator should be a closed case, a resolved incident, a completed review, an approved transaction or another measurable unit of work.

COST COMPONENT	EXAMPLES
Platform	Runtime, seats, sessions, environments, connectors, memory, registry, tracing, evaluation and support
Model	Input, cached input, output, reasoning, embeddings, speech, image and fallback calls
Tools and data	API consumption, search, vector store, data warehouse, browser, code execution, RPA, egress and third-party services
People	Engineering, product, risk, data stewardship, content, review, operations and change management
Quality and failure	Human review, rework, retries, escalations, incidents, latency, abandoned sessions and incorrect actions
Control	Identity, monitoring, testing, audit, policy, approvals, legal review, documentation and evidence retention

Model cost levers, in order of typical impact

- Route volume work to fast models and reserve frontier models for the tasks that need them.
- Cache stable context. Bedrock lists a 90 percent discount on cached input for OpenAI's GPT-5.6 models.[35]
- Use batch processing for work that does not need an immediate answer.
- Cap assistant spend by role or group rather than restricting access entirely.[29,24]

The cheapest platform on a token comparison can be the most expensive at the process level, if it needs more integration, human review, duplicate controls or operational support. A native platform with a higher unit price can be economical when it removes bespoke integration and speeds adoption. See AI Token Economics for the forecasting method.

Ask every finalist the same questions in writing.

Require answers that identify the product, edition, region, release stage and the contractual document on which each answer depends.

Platform questions

#	QUESTION
1	Which components are GA, preview, beta, early access, or restricted by region and edition?
2	What exactly is metered: users, conversations, sessions, tokens, tool calls, compute, memory, connectors, traces, evaluations or actions?
3	Can we use our own models, vector stores, identity provider, keys, network, logs and observability tools?
4	How are user delegation and workload identity represented across model calls, tools and downstream systems?
5	What data is retained in prompts, memory, traces, evaluations, support logs and abuse-monitoring systems, and for how long?
6	Can we export agent definitions, prompts, test sets, traces, memory, evaluation results and audit evidence in usable formats?
7	What happens when a model, tool, region or provider is unavailable? Which recovery features are native and which must we build?
8	How does the platform handle versioning, approval, promotion, rollback, decommissioning and ownership changes?
9	Which standards are implemented natively, and which are adapters with feature or security gaps?
10	What support, incident response, service levels, roadmap protection and commercial remedies apply to the complete production path?

Ask every finalist the same questions in writing.

Assistant and model providers need their own questions. The answers often differ by endpoint even for the same model.

Assistant and model questions

#	QUESTION
11	On this endpoint, which party processes our prompts and outputs, in which regions, and under whose terms?
12	Which models and features are GA on this endpoint compared with the vendor's own API, and what is the usual lag?
13	What notice do you give before a model version is retired, and what migration support is provided?
14	Can administrators restrict models, connectors and agent actions by role, require approvals, and set spend limits by group?
15	Is our data used for training, abuse monitoring or human review, and can retention be reduced to zero for this workload?
16	Does model consumption through this route count toward our existing cloud commitments, and at what price relative to the vendor's list?

How to use the answers

Score the answers into the gates and scorecard. Read in isolation, they mislead. An answer that depends on a preview feature, a roadmap item or an unsigned amendment is a conditional pass with an expiry date. Ask for the governing document by name, and keep it with the architecture decision record.

SEPARATE THE DEAL FROM THE DESIGN	Procurement can negotiate seat bundles and commitments in parallel. Don't let a bundle decide the architecture. A discounted assistant that can't meet a gate saves nothing.

Most platform mistakes are predictable.

Each pattern below has cost real organizations months. Use the list as a pre-mortem before the selection committee meets.

FAILURE PATTERN	WHY IT MATTERS
Buying the demonstration	The demo proves model behavior but not identity, real permissions, data quality, recovery, operations or unit economics.
Equating model choice with platform choice	You can usually reach the best model for one task through several platforms. The decision that lasts is the runtime and operating boundary.
Choosing the assistant by popularity	The assistant employees asked for may not see the documents, permissions and channels where their work happens.
Mandating a single model vendor	One vendor’s lead rarely lasts a year. A mandate removes your fallback and your bargaining power.
Chasing the leaderboard	Public rankings measure someone else’s task. Switching models without your own evaluation trades known errors for unknown ones.
One platform per department	Local speed creates duplicated identity, integration, tools, testing, monitoring, cost and governance.
One platform for every workload	Forced standardization ignores application-native semantics and creates unnecessary integration around systems that already govern the work.
Treating MCP or A2A as governance	Open protocols improve interoperability. They do not by themselves provide authorization, data policy, approval, testing, reliability or accountability.
Ignoring preview boundaries	A compelling feature may be beta, preview, early access, region-limited or unavailable under the intended contract.
Measuring cost per token	The enterprise optimizes a small variable cost while ignoring integration, review, failure and operating costs that determine the business case.

Buy the runtime; prove the complete workflow.

Use the same reference architecture to compare cloud, application-native and data-native platforms. Ask each finalist to demonstrate, not describe, each responsibility.

RESPONSIBILITY	ASK EACH FINALIST TO DEMONSTRATE
Identity and context	Verified user and workload identities, tenant isolation, delegation, context provenance and expiry. Who maintains these connections?
Retrieval and semantics	Current source permissions, versioned business definitions, citations and permission-change propagation. Which paths are native, custom or preview?
Tools and approval	Typed tools, action-level authorization, approval bound to payload, stale-state rejection and safe handling of uncertain writes.
Lifecycle and evidence	Versioned releases, representative evaluations, protected exportable traces, denial tests and accountable change approval.
Operations and exit	Latency, failure recovery, costs per completed workflow, incident ownership and export of contracts, policies and evidence.

Demonstrate one invoice workflow with the same identities and source records across finalists: explain a hold; deny an unauthorized read; propose a correction; require the right approval; reject a changed or stale command; reconcile an uncertain outcome. Record what the supplier provides, what your team must build and who operates each dependency.

Application-native platforms may reduce integration work inside their system of record. A broader runtime may suit work across several systems. Neither choice removes data rights, semantic ownership or command controls. Price the complete operating arrangement before you score convenience. See The Federated Data Platform and Context Before Query for the design detail.

Twelve steps from question to decision.

Use this sequence to run the selection. Each step produces an artifact the next one needs.

STEP	ACTION
1	Name one production-shaped workflow and its business outcome.
2	Split the question into the four decisions and name an owner for each.
3	Confirm the workforce assistant: suite-native default, exceptions proven by cohort test.
4	Classify the workflow’s task types and shortlist two or three models per class.
5	Choose the access path for each candidate model and record who processes the data.
6	Pick the lowest build tier that can carry the workflow’s risk.
7	Map the nine platform layers and identify the center of gravity.
8	Shortlist the strategic cloud runtime, the relevant native platform and at most one independent alternative.
9	Apply the hard gates in the intended region and edition.
10	Set scorecard weights before demonstrations and require evidence for every score.
11	Run the five-week proof with at least two models and measure cost per successful outcome.
12	Record the default, approved models, exceptions, review date and exit path.

FINAL PERSPECTIVE

The winning platform isn’t the one with the broadest demo. It’s the one that lets a named team run a valuable workflow with the right authority, data, models, controls, reliability and economics. Enterprises that make these boundaries explicit can use several assistants, models and agent platforms without losing control.

Sources and currency.

Product names, release stages and capabilities change quickly. Revalidate availability, region, edition, terms and pricing during procurement. Vendor descriptions are evidence of documented capability, not independent proof of production fitness.

[1] Microsoft. *Microsoft Foundry documentation*. https://learn.microsoft.com/en-us/azure/foundry/

[2] Microsoft. *Foundry Agent Service overview*. https://learn.microsoft.com/en-us/azure/ai-foundry/agents/overview

[3] AWS. *Amazon Bedrock AgentCore documentation*. https://docs.aws.amazon.com/bedrock-agentcore/latest/devguide/

[4] AWS. *Amazon Bedrock Agents Classic maintenance mode*. https://docs.aws.amazon.com/bedrock/latest/userguide/agents-classic-maintenance-mode.html

[5] Google Cloud. *Agent Builder and Gemini Enterprise Agent Platform documentation*. https://docs.cloud.google.com/agent-builder

[6] Google Cloud. *Gemini Enterprise Agent Platform release notes*. https://docs.cloud.google.com/gemini-enterprise-agent-platform/release-notes

[7] Salesforce. *Agentforce overview*. https://help.salesforce.com/s/articleView?id=ai.copilot_intro.htm&type=5

[8] Salesforce. *Agentforce trust and security*. https://help.salesforce.com/s/articleView?id=sf.copilot_trust.htm&type=5

[9] SAP. *Joule Studio agents*. https://help.sap.com/docs/joule-studio/joule-studio/agents?version=CLOUD

[10] SAP. *Joule Studio development flow*. https://help.sap.com/docs/joule-studio/joule-studio/general-development-flow

[11] ServiceNow. *Now Platform glossary and AI agent platform terms*. https://www.servicenow.com/docs/r/glossary/now-platform-glossary.html

[12] Oracle. *AI Agent Studio for Fusion Applications*. https://docs.oracle.com/en/cloud/saas/fusion-ai/26c/aiaas/index.html

[13] Snowflake. *Cortex Agents*. https://docs.snowflake.com/en/user-guide/snowflake-cortex/cortex-agents

[14] Databricks. *Deploy agents with Agent Bricks*. https://docs.databricks.com/aws/en/agents/deploy/

[15] OpenAI. *OpenAI Agents SDK*. https://openai.github.io/openai-agents-python/

[16] OpenAI. *OpenAI Agents SDK tools*. https://openai.github.io/openai-agents-python/tools/

[17] Anthropic. *Claude Platform API overview*. https://platform.claude.com/docs/en/api/overview

[18] Anthropic. *Claude managed agent definition*. https://platform.claude.com/docs/en/managed-agents/agent-setup

[19] LangChain. *LangSmith data plane*. https://docs.langchain.com/langsmith/data-plane

[20] IBM. *watsonx Orchestrate governance and observability*. https://www.ibm.com/products/watsonx-orchestrate/governance-and-observability

[21] UiPath. *Maestro overview*. https://docs.uipath.com/maestro/automation-cloud/latest/user-guide/overview

[22] UiPath. *Maestro September 2026 release notes*. https://docs.uipath.com/maestro/automation-cloud/latest/release-notes/september-2026

Sources and currency.

[23] OpenAI. *API models*. https://developers.openai.com/api/docs/models

[24] OpenAI. *ChatGPT is now a partner for your most ambitious work, 9 July 2026*. https://openai.com/index/chatgpt-for-your-most-ambitious-work/

[25] OpenAI. *ChatGPT business pricing and plan comparison*. https://openai.com/business/pricing/

[26] Anthropic. *Claude models overview*. https://platform.claude.com/docs/en/about-claude/models/overview

[27] Anthropic. *Claude Opus 5.5 model page*. https://platform.claude.com/docs/en/models/opus-5-5/overview

[28] Anthropic. *Claude Models on Google Cloud*. https://platform.claude.com/docs/en/build-with-claude/claude-models-on-google-cloud/overview

[29] Anthropic. *Claude Cowork and chat come together for Enterprise, 23 October 2026*. https://claude.com/resources/articles/unified-claude-enterprise

[30] Google. *Gemini API release notes*. https://ai.google.dev/gemini-api/docs/changelog

[31] Google Cloud. *AI agents for Gemini Enterprise editions*. https://cloud.google.com/gemini-enterprise/agents

[32] Google Cloud. *Gemini is now available anywhere (Google Distributed Cloud), 28 August 2025*. https://cloud.google.com/blog/topics/hybrid-cloud/gemini-is-now-available-anywhere

[33] Google Cloud. *OpenAI models on Google Cloud, private preview, 21 October 2026*. https://cloud.google.com/blog/products/ai-machine-learning/google-cloud-brings-openai-models-to-customers

[34] AWS. *GPT-5.5, GPT-5.4 and Codex generally available on Amazon Bedrock, 1 June 2026*. https://aws.amazon.com/about-aws/whats-new/2026/06/amazon-bedrock-openai-models-codex-generally-available/

[35] Amazon. *OpenAI models on Amazon Bedrock (release timeline)*. https://www.aboutamazon.com/news/aws/bedrock-openai-models

[36] AWS. *Amazon Bedrock data protection*. https://docs.aws.amazon.com/bedrock/latest/userguide/data-protection.html

[37] AWS. *gpt-oss and Nemotron models on Amazon Bedrock in AWS GovCloud (US), April 2026*. https://aws.amazon.com/about-aws/whats-new/2026/04/openai-gpt-oss-nvidia-nemotron-govcloud/

[38] AWS. *Amazon Bedrock Intelligent Prompt Routing generally available*. https://aws.amazon.com/about-aws/whats-new/2025/04/amazon-bedrock-intelligent-prompt-routing-generally-available

[39] Microsoft. *Claude models in Microsoft Foundry*. https://learn.microsoft.com/en-us/azure/foundry/foundry-models/concepts/claude-models

[40] Microsoft. *What's new in Microsoft Foundry, September 2026*. https://devblogs.microsoft.com/foundry/whats-new-in-microsoft-foundry-september-2026/

[41] Microsoft. *Foundry model retirement schedule*. https://learn.microsoft.com/en-us/azure/foundry/openai/concepts/model-retirement-schedule

[42] Microsoft. *Microsoft Copilot Studio documentation*. https://learn.microsoft.com/en-us/microsoft-copilot-studio/

[43] Directions on Microsoft. *M365 Copilot adds choice and risk with Anthropic's Claude, updated 12 June 2026*. https://www.directionsonmicrosoft.com/reports/m365-copilot-adds-choice-and-risk-with-anthropics-claude

[44] n8n. *Hosting and installation documentation*. https://docs.n8n.io/hosting/installation/

Sources 1 to 22 support the platform review researched on 27 October 2026. Sources 23 onward, and the dated market statements in sections 01 to 11, were checked on 29 October 2026. Source 43 is independent analysis rather than vendor documentation. Model names, release stages and catalog contents are expected to change within weeks.

DISCLOSURE

No vendor reviewed, sponsored or paid for this paper. Products and models are named as examples of patterns and as candidates to test, not endorsements or test results. Shortlists reflect vendor documentation and RJL practice; your own evaluation decides.

ABOUT

Richard White

I was global firm chief data officer at McKinsey & Company, owning data, AI and platforms end to end, and before that chief data officer at New York Life, where I built the chief data office and reported data risk to risk committees at multiple levels. Earlier I led data at the Federal Reserve Bank of New York and enterprise data governance at Citigroup. I founded RJL Solutions in 2024. My book, *The Day After the Demo*, is free at rjlsol.com/book.

RJL Solutions connects strategy, architecture, engineering, governance and operation through Frame, Prove, Build, and Run & Transfer. Each phase is separately scoped and priced in a written proposal before you commit. The first sixty-minute working session is free.

ECONOMICS **AI Token Economics**
Forecast cost per accepted task and set spend controls.

ARCHITECTURE **The Reference Implementation**
AI, data and security as one system.

PATTERNS **Patterns and Product Fit**
Retrieval, enforcement and product fit at enterprise and lean scale.

RJL SOLUTIONS

WEB rjlsol.com

EMAIL rwhite@rjlsol.com

Product and model statements are dated and describe vendor documentation, not RJL testing. Examples are illustrative and describe no client engagement or client result. Revalidate availability, terms and pricing before you commit.