

Generative AI Platforms: Selection, Architecture, and Tradeoffs



Course promise: a defensible platform selection method, not a vendor tour



Scope: 2026 landscape, reference architecture, weighted evaluation, governance, exit planning



Three platform tiers: frontier APIs, managed application platforms, self-hosted stacks



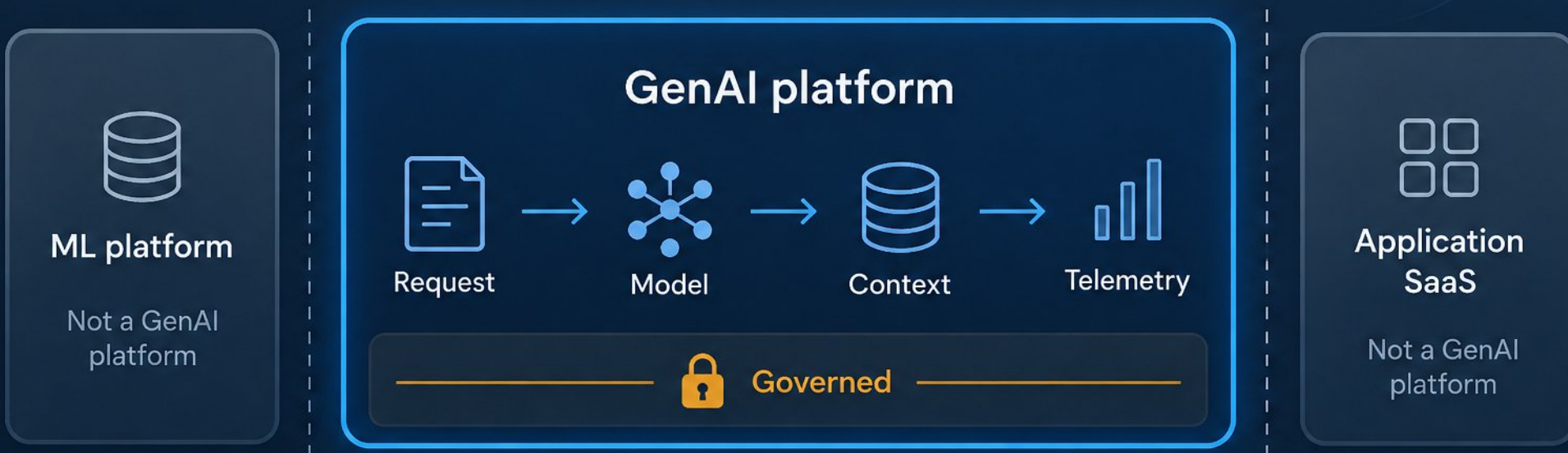
Decision context: build vs. buy vs. assemble, and product roadmap coupling



Deliverable: weighted decision matrix and defensible reference architecture sketch



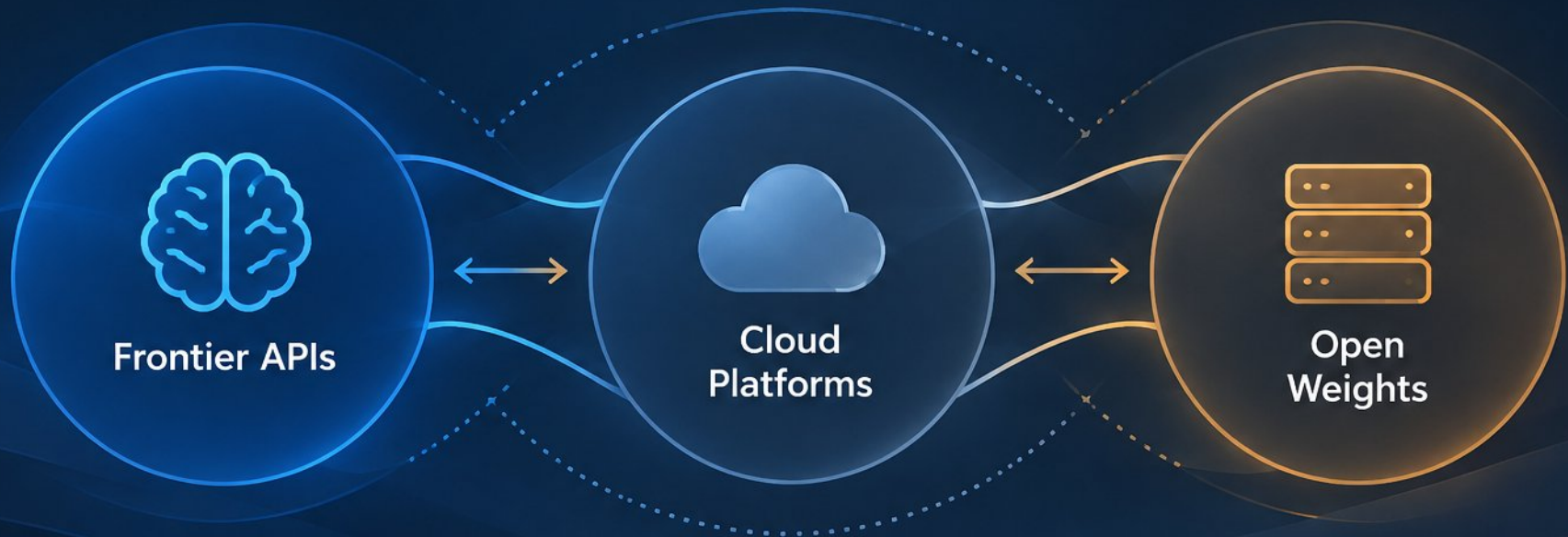
What Counts as a Platform and Why the 2026 Market Looks This Way



- A GenAI platform is the governed path from request to model, context, and telemetry
- Not an ML platform, not application SaaS — governance is what defines it
- 2026 worldwide AI models and platforms spend: \$64B, up 63% from 2025
- GenAI model spend up 117%; specialized models fastest-growing at 210%
- Buyer question shifted: where does our data and our agents live?

The Three Vendor Camps: Frontier APIs, Cloud Platforms, and Open Weights

- Frontier labs (OpenAI, Anthropic, Google, Mistral): highest capability via API, least to operate
- Prompts leave your boundary; provider sets price and deprecation cadence
- Cloud platforms (Azure AI Foundry, Bedrock, Vertex): governed environment inside your tenancy, IAM attached
- Open weights (Llama, Mistral, Nemotron): self-host for sovereignty and predictable GPU cost
- Camps blur: labs distribute through clouds; shortlists compare across all three



Hyperscaler Platforms Compared: Where the Operational Edges Differ

MODELS
PEOPLE
OPERATIONS
REAL IMPACT

SAME
CORE
DIFFERENT
EDGES



COMMON
FOUNDATION
DIFFERENT
STRENGTHS

- All three host top commercial and open-weight models, with native IAM and observability
- **Vertex AI:** broadest model catalog via Model Garden; aggressive Gemini/Gemma pricing; ADK shipped 2026
- **Bedrock:** easiest AWS procurement and compliance path; Claude parity, but narrower catalog outside Anthropic/Titan/Llama
- **Azure AI Foundry:** OpenAI exclusivity inside compliance boundary, Microsoft 365 integration; PTU pricing hurts bursty loads
- **Practice pattern:** most serious orgs run two platforms — a primary for ~80% of workloads, a secondary for specific needs

Architecture Layers: From Request Ingress to Governed Response

A layered architecture for safe, reliable, and adaptable AI systems



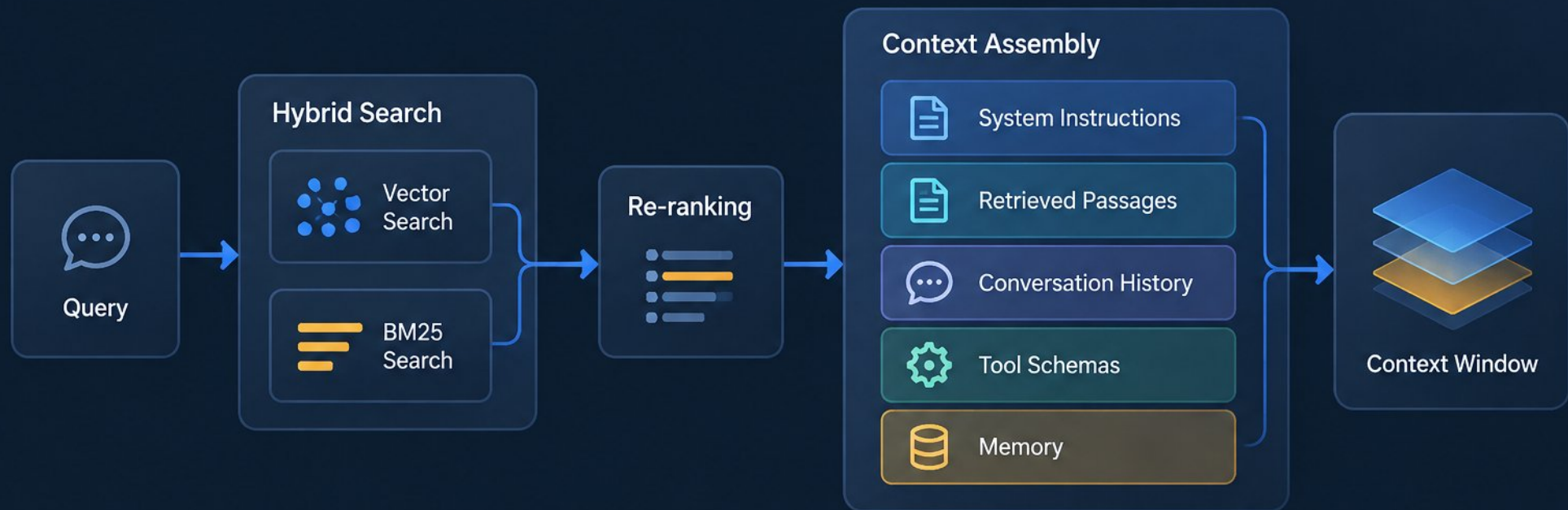
- Seven layers: ingress, gateway, context assembly, model execution, validation, observability, feedback store
- The raw model call is only 10–15% of the work
- Cross-cutting planes — observability, governance, model lifecycle — span every layer
- Layer separation lets retrieval, guardrails, or telemetry change independently
- Anti-pattern: building the agent as a sidecar instead of inside the enterprise

The Gateway as the Highest-Leverage Layer



- Thin internal service every model call passes through—one interface, many providers
- Four concerns attach at one chokepoint: routing, guardrails, cost caps, observability
- Model aliasing ("chat-standard" → concrete model) turns provider swap into config change
- Fallback chains, semantic caching (30-60% hits), per-tenant budgets live here
- Production reality: fallback capacity under-provisioned; streaming locks you to one provider

Retrieval and Context Engineering as the Real Quality Lever



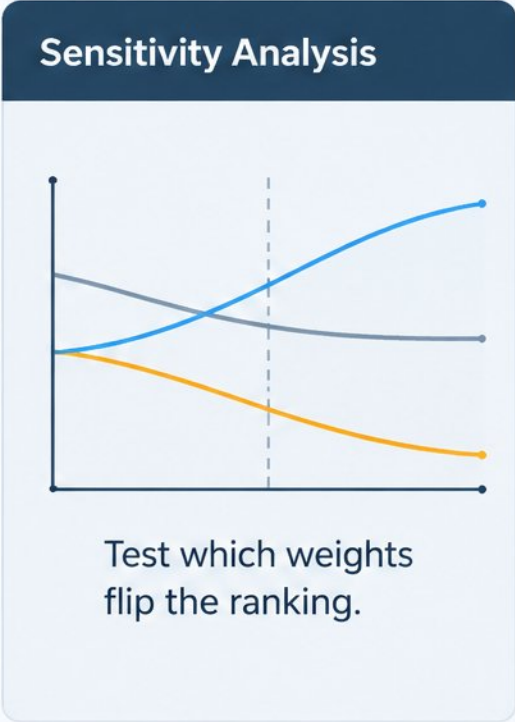
- Context assembly is the biggest quality lever — and the most neglected
- Context engineering now beats prompt engineering: what to include, order, compression
- Hybrid search (vector + BM25) with re-ranking for exact-match queries
- Query-time retrieval for changing corpora; cached context for stable knowledge
- Weak retrieval produces weak output — retrieval and policy boundaries are undervalued



Evaluation Criteria and a Weighted Decision Matrix That Survives Review

BETTER
DECISIONS
STRONGER
OUTCOMES

Evaluation Criteria	Weight (by use case)	Candidate A	Candidate B	Candidate C
Capability	●			
Cost	●			
Latency	●			
Scalability	●			
Security	●			
Ecosystem	●			
Operability	●			
Strategic fit	●			
Hard gates (data residency, regulatory class)		◆	◆	◆





Criteria families

Criteria families: capability, cost, latency, scalability, security, ecosystem, operability, strategic fit.



Method

Method: normalize, weight by use case, score, then sensitivity-test which weights flip the ranking.



Hard gates vs. soft preferences

Hard gates (data residency, regulatory class) differ from soft preferences (SDK ergonomics, docs).



TCO buckets

TCO buckets: inference, storage, egress, evaluation, human review, platform engineering headcount.



Bias traps

Bias traps: demo-driven choice, incumbent bias, ignored switching costs, list prices over loaded cost.

Cost, Latency, and Quality:

Managing the Trilemma Instead of Solving It

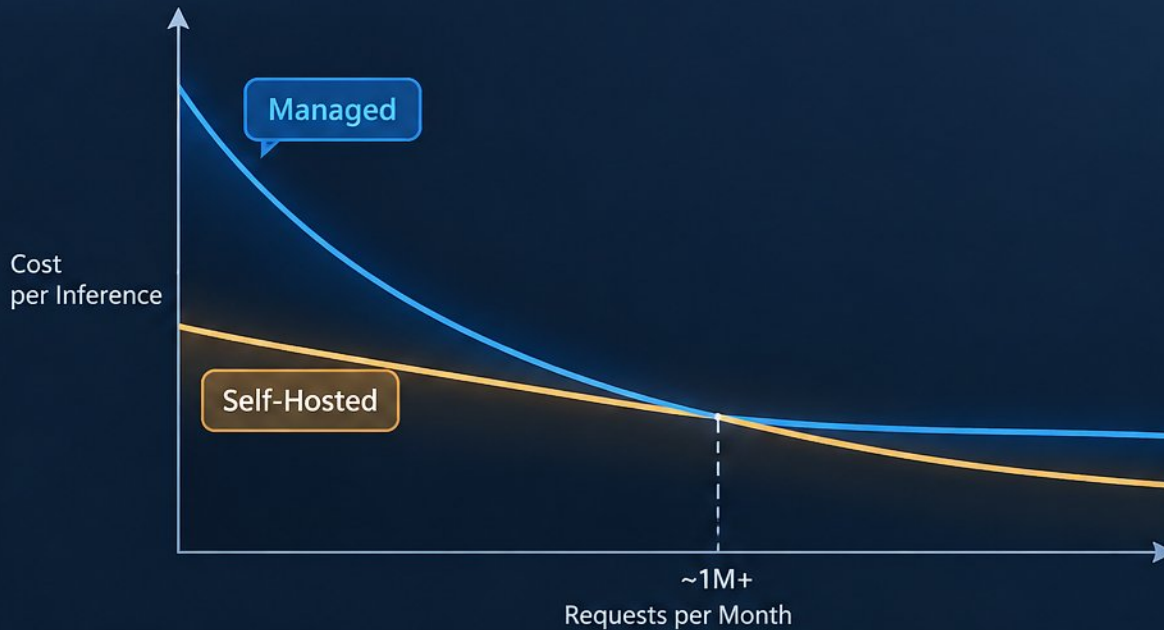
TRADE OFFS
DRIVE
BETTER SYSTEMS

- The three axes are coupled: improving one shifts the other two
- Define SLOs per budget: quality, p95 latency, cost per success, capacity
- Loaded cost per accepted result beats raw token rate
- Highest-leverage levers: route to cheapest sufficient tier, prompt caching, output caps, batch deferrable work
- Latency levers: streaming, parallel tool calls, smaller context, precomputed answers



Managed Versus Self-Hosted: Break-Even Points and Hybrid Architectures

Cost per Inference



Underutilized GPUs
Cost More



Hybrid Architecture



Start managed for the first 6-18 months; simplicity is dramatic



Self-hosting wins at ~1M+ requests/month, sovereignty, or deep customization



Underutilized GPUs (below ~60%) cost more per inference than APIs



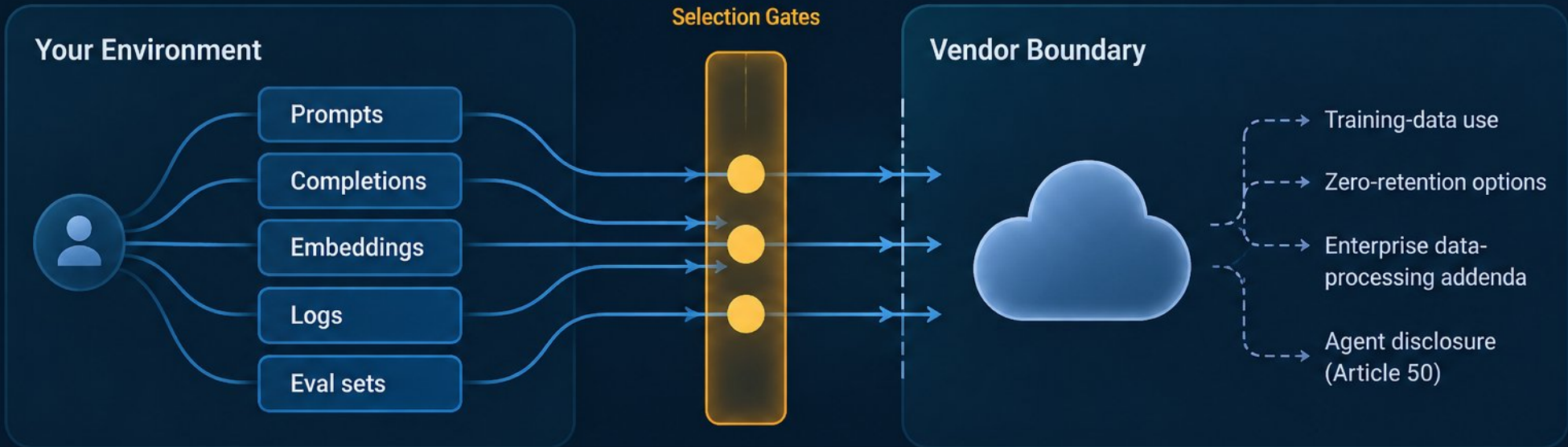
Managed dedicated endpoints often beat self-managed once 40+ eng hours/month count



Hybrid wins most: self-hosted small models plus frontier APIs behind one router

Data, Security, Privacy, and Regulatory Posture as Selection Gates

RESPONSIBLE AI
HIGHER STANDARDS
REAL IMPACT



- Map data flow first: prompts, completions, embeddings, logs, eval sets cross boundaries differently
- Verify vendor terms: training-data use, zero-retention options, enterprise data-processing addenda
- No substantiation from a vendor is itself a risk signal — document it
- EU AI Act 2026: Article 50 transparency live since 2 August 2026; GPAI enforceability 2026
- Annex III high-risk deferred to 2 December 2027; Annex I to 2 August 2028
- Agents must disclose artificial nature and the principal represented — across multi-agent routing
- Controls: tenant isolation, secret management, injection/exfiltration defenses, 6-month log retention

FROM DATA FLOWS
TO TRUSTED OUTCOMES

Observability, Evaluation, and Release Engineering for Non-Deterministic Systems



Standard APM misses semantic failure: 0% errors, 99.9% uptime, confidently wrong answers



Offline evals and online observability are two halves of one loop



Cheapest method first: assertions, then LLM-as-judge, then human review



Prompts, models, retrieval, and tools ship as one pinned manifest



Gate on risk slices, not aggregate scores that hide broken cases



Canaries watch corrections, escalations, and refusals, not just errors



Operating Model, Team Topology, and FinOps Governance for Token Spend

1

Shared platform contract: gateway, eval harness, guardrails — ship in days

2

Centralized enablement with federated execution; small teams, embedded champions

3

Stamp every request with team/project/agent tags at the gateway

4

Chargeback on cost per validated outcome, not raw token consumption

5

Staff against: retrieval engineering, evaluation design, AI security, inference ops, FinOps



Migration, Exit Planning, and the Decision Workshop

- **Integration surfaces set true cost:** IdP, data platform, API gateway, MLOps registry, SIEM, ITSM queue
- **Portability:** gateway abstraction, vendor-independent eval harness, prompt versioning, embedding export paths
- **Migration patterns:** parallel run, shadow evaluation, progressive cutover by segment, pre-agreed rollback criteria
- **Lock-in anti-patterns:** scattered provider SDK calls, un-exportable vector indexes, no eval suite to re-tune
- **Workshop:** regulated assistant, high-volume consumer feature, internal copilot — scored, architected, risk-registered



Portable integration hooks



Lock-in integration hooks



PersonWise.

PersonWise • AI digital-human course platform



Scan the code or click anywhere to watch this course live, with voice Q&A

personwise.ai/t/a78c032a/public