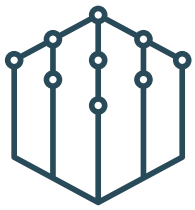
TECHNICAL BRIEF

Engineering Production-Grade AI Systems



Solving the 10 systemic challenges in AI factories

Solution overview

The primary failure mode of enterprise AI is not model performance—it is system design. As organizations move from prototype to production, they encounter a new class of problems that emerge from the interactions between components: data pipelines, retrieval systems, memory layers, orchestration engines, and governance controls. These systems are inherently distributed, probabilistic, and stateful, requiring tight coordination across layers that were never designed to operate together.

EDB Postgres® AI (EDB PG AI) Agent Factory—EDB's complete Postgres-native solution for building, testing, and deploying sovereign AI agents—addresses this by collapsing system boundaries. It co-locates data, memory, retrieval, orchestration, and governance into a single transactional environment. This architectural unification is what enables deterministic control over inherently probabilistic systems.

State, retrieval, data, and orchestration as a unified system

The first principle of a production-grade AI system is that state must be durable, consistent, and transactionally managed. In traditional architectures, agent memory is distributed across caches, vector databases, and application layers, leading to inconsistent context. By persisting both structured state and embeddings inside Postgres, the system ensures that every agent interaction operates on a consistent, ACID-compliant view of context. Memory is no longer ephemeral—it is versioned, queryable, and synchronized with system state.

Retrieval systems must evolve alongside data. Embedding drift is inevitable when data changes, yet most systems treat embeddings as static artifacts. EDB PG AI addresses this by embedding lifecycle management directly into the database. By versioning embeddings, enabling incremental reindexing, and supporting hybrid queries, the system maintains semantic alignment between data and retrieval over time. This ensures that agent reasoning remains accurate even as underlying data evolves, delivering 100x faster index builds and 2x query throughput versus standard implementations.

Data freshness is equally critical. AI systems that rely on external pipelines inevitably introduce latency and inconsistency. By eliminating data movement and enabling agents to operate directly on live Postgres data, the system aligns AI outputs with real-time business state. Change data capture and event-driven updates ensure that derived representations remain synchronized without batch delays.

Orchestration, often treated as an external concern, becomes a core system capability. By integrating with Langflow and its 700 pre-built integrations, EDB PG AI exposes this as stateful, observable workflows within the same system boundary. Execution state is persisted, retries are deterministic, and partial failures can be recovered without restarting entire workflows. This transforms orchestration from a fragile integration layer into a reliable execution engine.

Security, cost control, and deterministic behavior as foundational requirements

Security in AI systems cannot be bolted on—it must be enforced at runtime. Prompt injection and adversarial inputs are addressed by treating all inputs as untrusted and applying policy enforcement at the data and execution layers. The native security mechanisms of Postgres extend directly to AI workloads, ensuring that access control and data governance are consistently applied.

Model routing introduces another layer of complexity. Without centralized control, organizations face unpredictable costs and inconsistent performance. Model Serving in EDB PG AI supports more than 500 providers, including NVIDIA NIM and Hugging Face Hub. This enables dynamic model selection based on workload characteristics, balancing cost and latency within the system boundary.

Testing and determinism remain among the most challenging aspects of AI systems. By versioning every component—models, prompts, data, and execution paths—and capturing full execution traces, the system enables reproducibility. Workflows can be replayed, validated, and debugged with precision, transforming probabilistic systems into testable systems.

Coordination, observability, and governance as core capabilities

As systems scale, multiple agents must coordinate effectively. Without centralized control, this leads to redundant work, conflicts, and execution loops. By managing coordination through shared state and orchestration logic within Postgres, agents operate within defined boundaries, ensuring efficient task execution and preventing deadlocks.

Observability is not an afterthought—it is embedded into every interaction. Each agent workflow is treated as a traceable transaction, with telemetry captured across all layers. This enables end-to-end visibility into system behavior, allowing engineers to diagnose issues and optimize performance with precision.

Governance completes the system. AI introduces new data risks, particularly around embeddings and derived data. By extending the Postgres governance model to all AI artifacts, the system ensures that security, compliance, and auditability are enforced consistently. There is no separate security boundary—AI operates within the same trusted environment as enterprise data.

How the system solves the 10 core challenges

The following table summarizes how each of the 10 systemic challenges is resolved within this unified architecture.

AI System Challenges → Postgres-Native Technical Solutions

Challenge	System problem	Technical resolution (EDB PG AI)	Resulting system property
State management and memory	Fragmented, inconsistent context	Unified relational and vector persistence with transactional consistency	Deterministic, durable agent state
Retrieval quality and drift	Embeddings degrade over time	Versioned embeddings, incremental reindexing, hybrid query execution	Stable, accurate semantic retrieval
Data freshness	Stale or delayed pipelines	In-database pipelines with CDC and real-time access	Real-time, consistent data alignment
Tool orchestration	Fragile, failure-prone workflows	Stateful orchestration with retries, idempotency, and execution tracking	Reliable, recoverable workflows
Prompt/policy injection	Unsafe or adversarial inputs	Runtime validation, policy enforcement, data-layer security controls	Secure, controlled execution
Model routing and cost	Inefficient model usage	Dynamic routing with cost/performance optimization	Predictable cost and performance
Determinism and testing	Non-reproducible behavior	Full versioning plus execution trace replay	Testable, debuggable systems
Multi-agent coordination	Conflicts and loops	Centralized orchestration plus shared state management	Scalable, coordinated execution
Observability and debugging	Opaque system behavior	End-to-end tracing with correlated telemetry	Transparent, diagnosable systems
Governance and security	Data leakage and compliance risk	Native RBAC, RLS, audit logging applied to all AI data	Enterprise-grade compliance

*Competitive comparisons are based on publicly available information and are subject to change as vendor offerings evolve and new information is made available. All product names, trademarks, and registered trademarks are the property of their respective owners.

Independent research validates the approach. McKnight Consulting Group found organizations using EDB PG AI can reduce development timelines from 28 weeks to 9 weeks—a 67% reduction in effort—with 38% lower long-term maintenance costs. In addition, Everest Group documented 55%+ reduction in AI workflow complexity, 90%+ elimination of integration steps, and 50% better total cost of ownership versus fragmented architectures.

Architectural implication: Eliminating fragmentation

This table demonstrates architectural consolidation rather than just feature coverage. Each challenge is not solved by adding another component but by eliminating the boundaries that created the problem in the first place.

Instead of:

- Separate memory stores
- External vector databases
- Independent orchestration engines
- Disconnected observability tools

... the system unifies these capabilities within a single operational substrate. This eliminates:

- Data movement latency
- Cross-system inconsistency
- Integration complexity
- Operational overhead

EDB Postgres AI: The sovereign data and AI platform for the agentic enterprise

EDB PG AI brings together a unified data layer, governance, sovereign control and orchestration, and an agent runtime environment, giving enterprises a trusted foundation for AI on infrastructure they own and control. The platform unifies transactional, analytical, and AI workloads in a single Postgres-based architecture—eliminating ETL, data movement, and operational fragmentation. And you choose where and how to deploy: on-premises, cloud, managed, or certified appliance.

The outcome: production-ready sovereign AI in days or weeks, not months.

Conclusion: AI as a system, not a stack

The 10 challenges outlined here are not isolated; they are emergent properties of fragmented architectures. Attempting to solve them individually leads to increasing complexity. Solving them systemically requires a unified approach.

By collapsing data, memory, retrieval, orchestration, and governance into a single Postgres-native system, organizations can build AI systems that are:

- **Consistent** in state and behavior
- **Accurate** in retrieval and reasoning
- **Real time** in data access
- **Reliable** in execution
- **Secure** by design
- **Observable** end-to-end
- **Governed** within enterprise constraints

This is the difference between assembling AI components and **engineering AI systems**.

And that distinction is what determines whether AI remains experimental or becomes production grade.

EDB POSTGRES AI

EDB Postgres® AI (EDB PG AI) is the sovereign data and AI platform for the agentic enterprise. Built on Postgres, the world's leading open source database, EDB PG AI unifies transactional, analytical, and AI workloads in a single governed architecture, on-premises and across clouds. To learn more, visit www.enterprisedb.com.