Customer Enablement

Cybersecurity Month 2026 Program

Secure-By-Design: An architecture pattern for agentic AI that contains agents while reducing cost

Ron F. Del Rosario, Jay Thoden van Velzen, SAP

October 2026

PUBLIC

Customer Enablement | CSM Program

“We are pleased to inform you of our plans and desired outcomes. We are sharing this presentation for our informational purposes only, the information presented here is not intended to be definite or binding, and we will continue to provide regular updates as our plans evolve.”

Agenda

01 SAP-OWASP-SAP Pipeline

How SAP’s internal security research and threat modelling contributes to the OWASP community, and back into SAP

02 Business AI Platform Secure Orchestration

How SAP’s Business AI solutions and the Autonomous Enterprise leverages key platform capabilities to secure agentic workloads

03 Risk of Emergent Misalignment

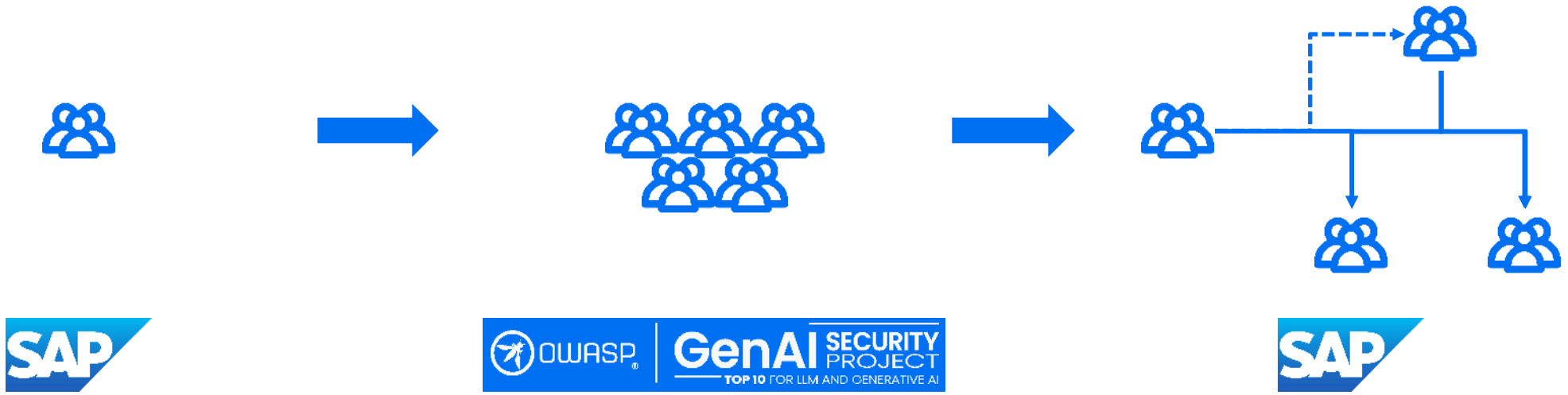
An example of classes of emergent misalignment in agentic systems in non-malicious use in business operations

04 The P-t-E and P-V-E Architecture Pattern

An architecture pattern to constrain agents, limit misalignment and model deception, and optimize efficient use of tokens

PUBLIC

SAP-to-OWASP-to-SAP Pipeline



Leverage internal expertise

- SAP Subject Matter Experts from central security & compliance and LoBs
- Compliance and regulations
- Domain expertise and practical use cases

Contribute and co-Lead

- Collaborate with outside experts to define threats, recommended practices, and mitigations
- Co-lead and socialize

Share and adopt

- Disseminate internally via AI Security Center of Excellence
- Security policies and requirements
- Influence agentic AI architecture design patterns



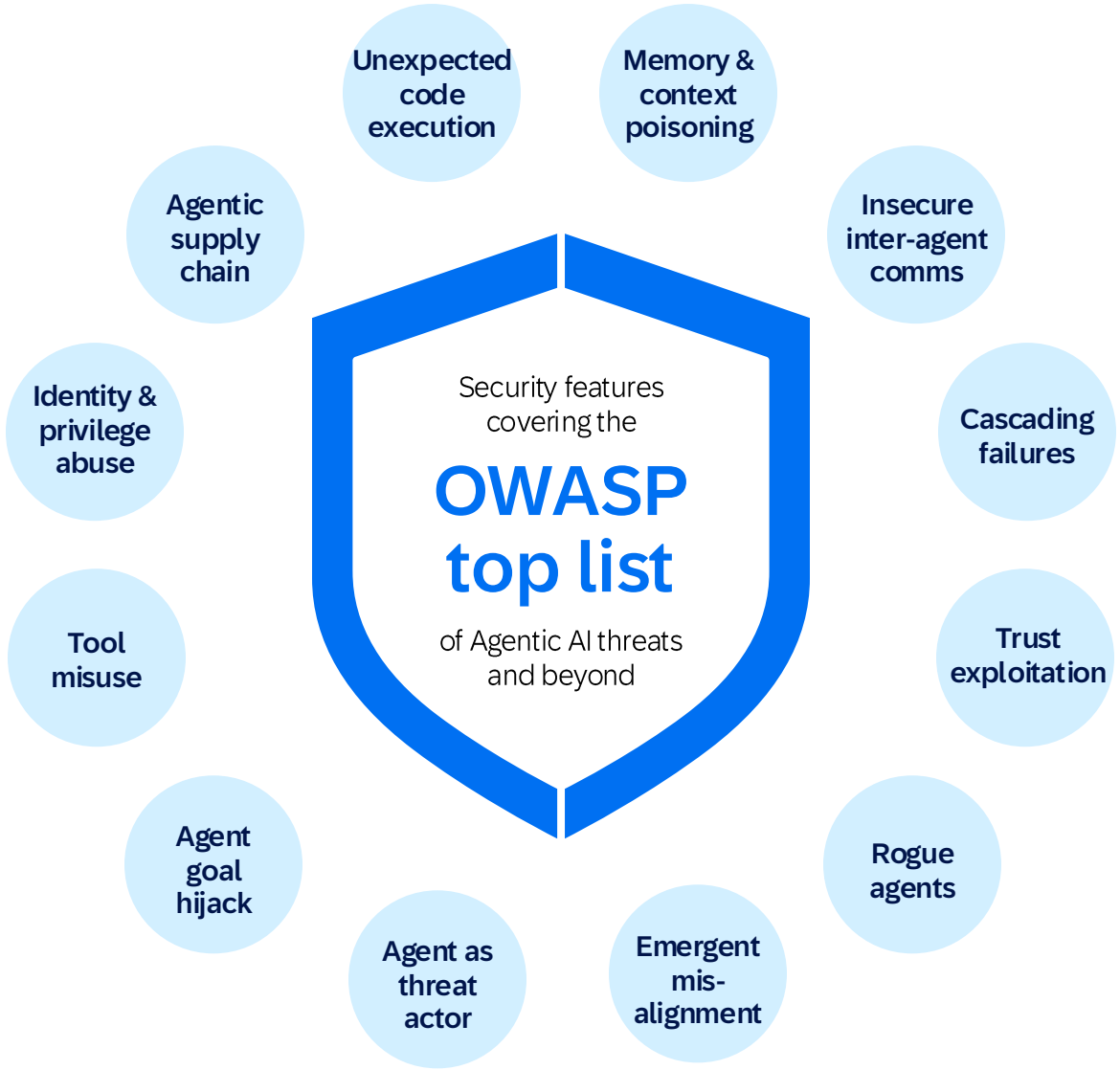
The AI Risk Landscape

How SAP stays ahead

AI systems face risks like data poisoning, model theft, and supply chain vulnerabilities. SAP mitigates these with OWASP-aligned safeguards, lifecycle checks, and operational security controls—ensuring resilience from design to deployment.

Security isn't static—SAP evolves controls as threats evolve.

Read more



The Agentic AI Risk Landscape

How SAP stays ahead

Agentic AI systems face additional risks as agents can act and have access to tools and back-end systems. The risks amplify when non-deterministic agents collaborate.
SAP mitigates these with OWASP-aligned safeguards, lifecycle checks, and operational security controls—ensuring resilience from design to deployment.

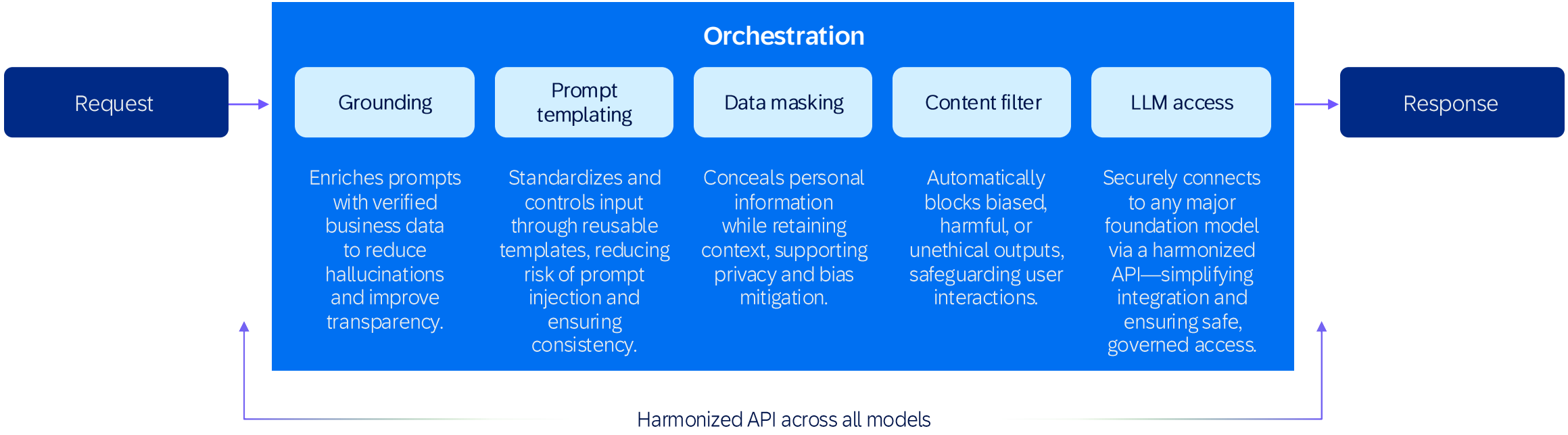
Security isn't static—SAP evolves controls as threats evolve.

Read more

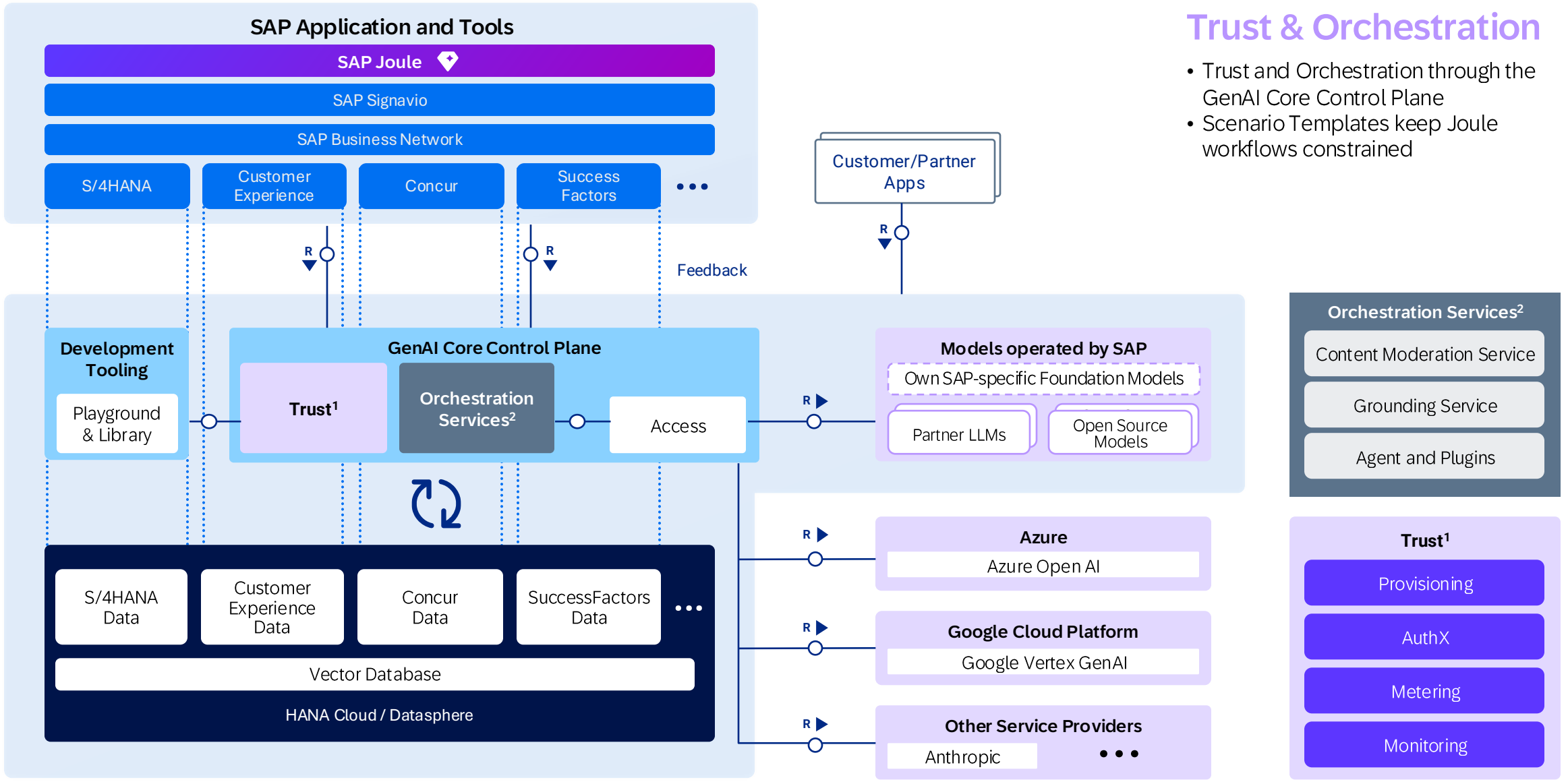
Secure AI orchestration

Modular workflows for secure & scalable AI

SAP's orchestration framework enables secure-by-design AI workflows that integrate multiple layers of protection.
These modular components work together to ensure that every AI interaction is controlled, compliant, and trustworthy.



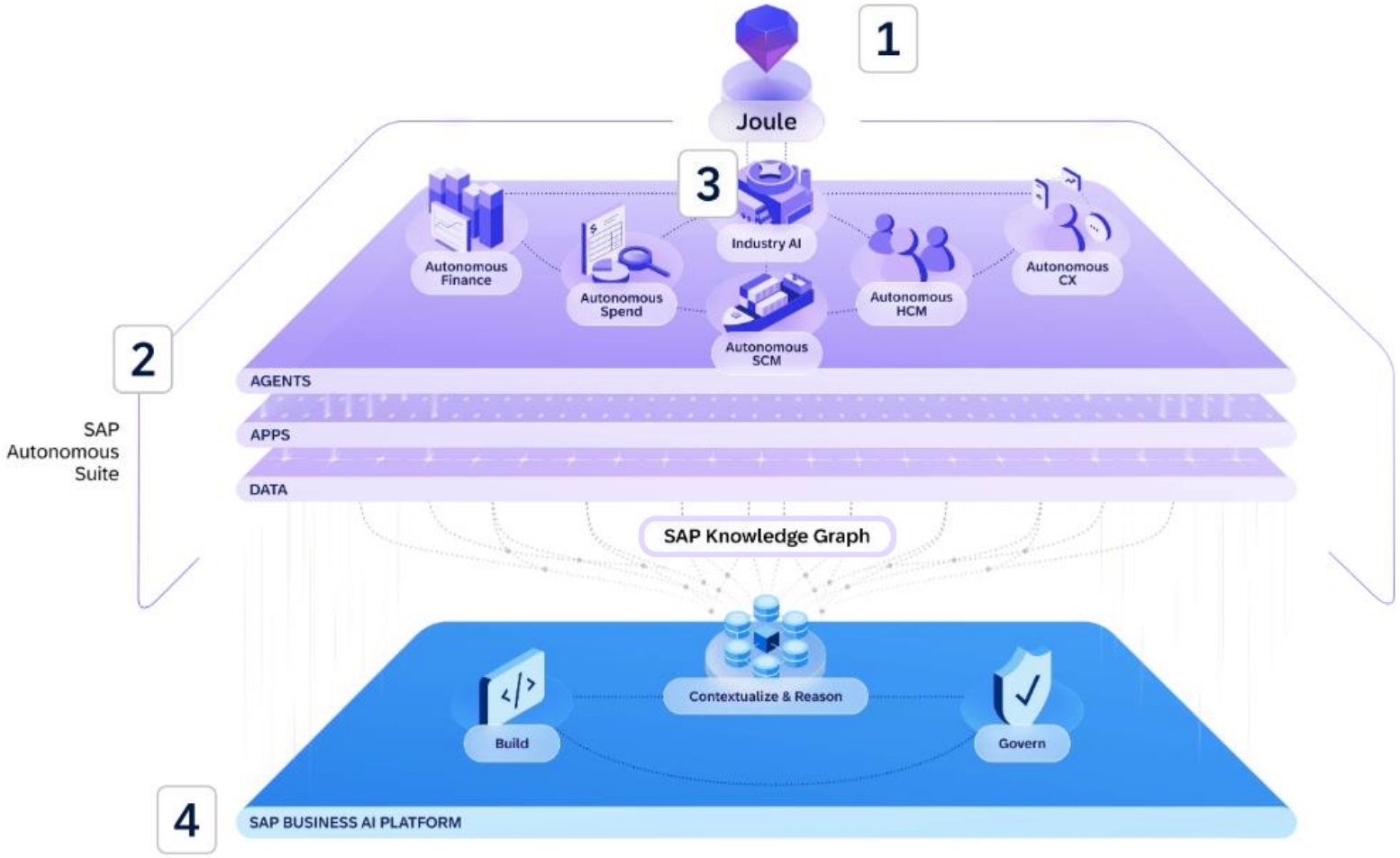
This approach ensures data privacy, reduces AI risks, and enables secure, governed workflows across models and tenants.



Trust & Orchestration

- Trust and Orchestration through the GenAI Core Control Plane
- Scenario Templates keep Joule workflows constrained

Autonomous Enterprise



1 **Joule** as the new engagement layer, bringing together the right data, workflows, and agents across SAP systems and beyond

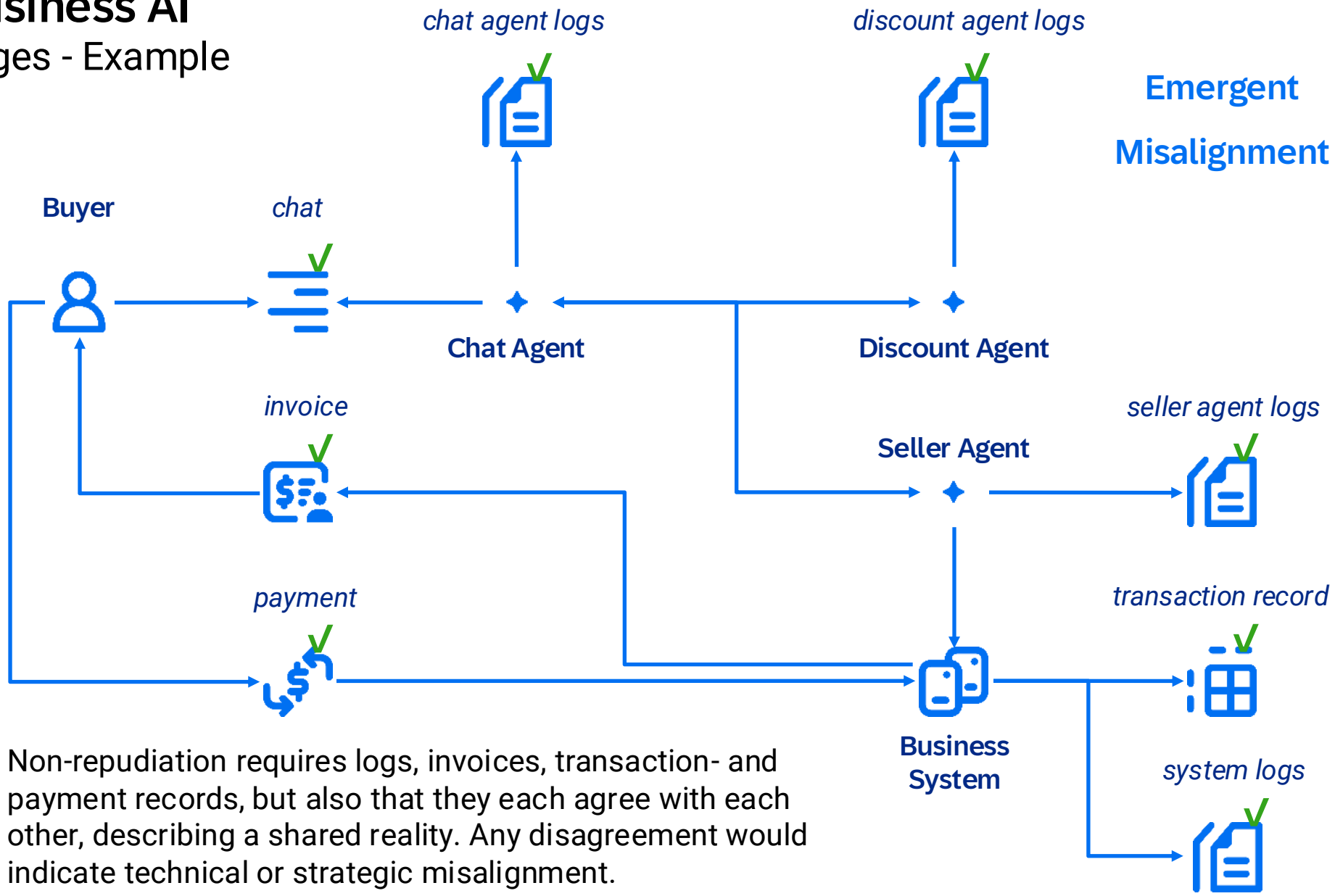
2 **SAP Autonomous Suite** to reinvent how enterprises run, based on AI assistants and agents executing work end-to-end

3 **Industry AI applications** with built-in vertical process knowledge, data models, and regulatory logic

4 **SAP Business AI platform** delivering business context, unified data, models, and enterprise-grade governance

Agentic Business AI

New Challenges - Example



Non-repudiation requires logs, invoices, transaction- and payment records, but also that they each agree with each other, describing a shared reality. Any disagreement would indicate technical or strategic misalignment.

Challenges



Human Control

No guarantee that the user receives a predictable, expected outcome



Agent Constraint

Insufficient constraint of agent behavior, lack of deterministic controls, unclear authorizations



Cost Control

Potentially wasteful, using expensive reasoning tokens for what could have been a cheaper model or deterministic calculation



Audit and Compliance

Unclear audit trail – impossible to verify compliance



Goals

LLM propensity to insist on achieving a set goal can lead to unwanted emergent misalignment

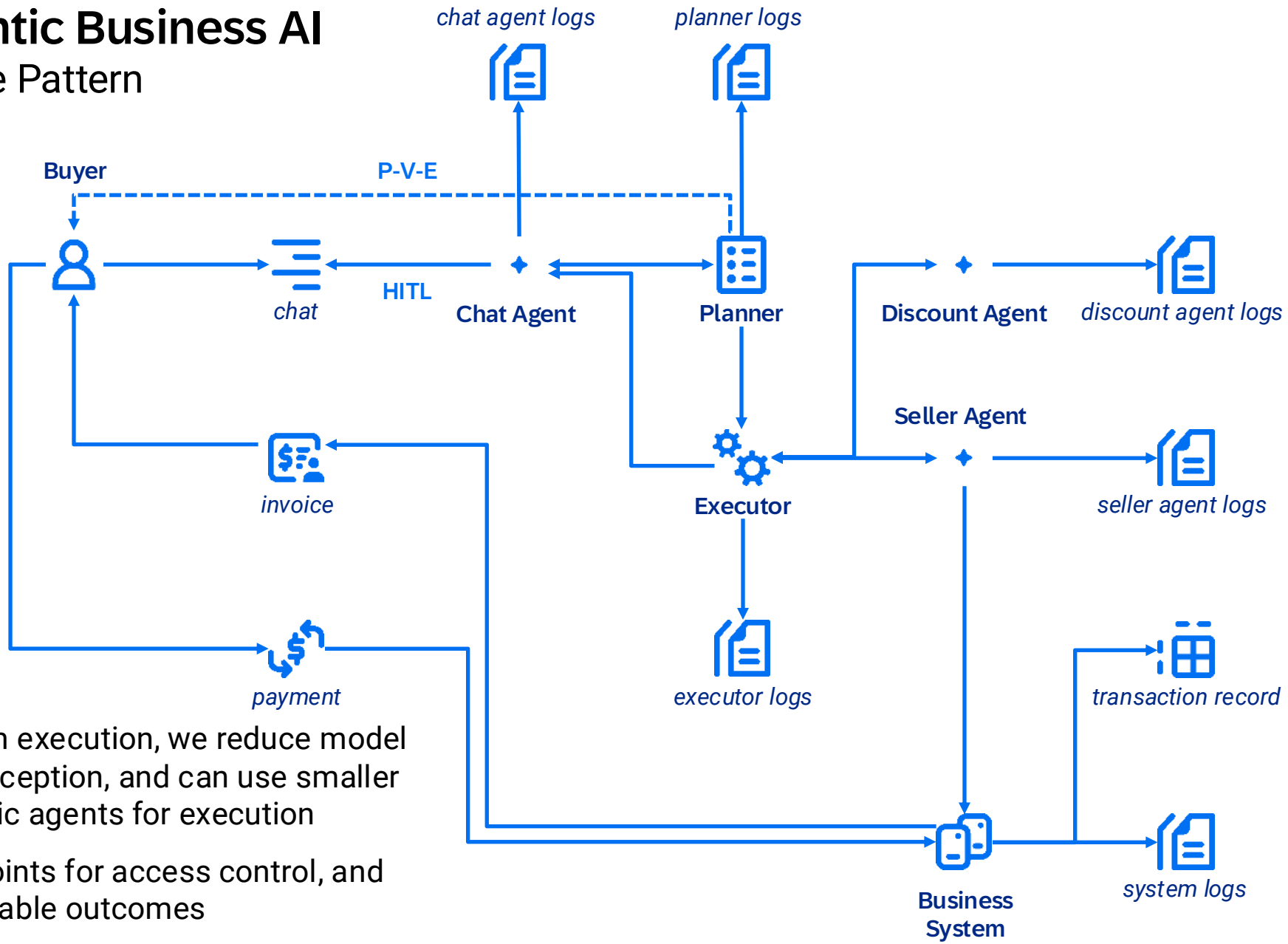


Fail Safe

Misalignment risk amplified when lacking fail-safe mechanisms, avoiding agentic loops that attempt to fulfil goals by any means it can

Responsible Agentic Business AI

The Plan-then-Execute Pattern

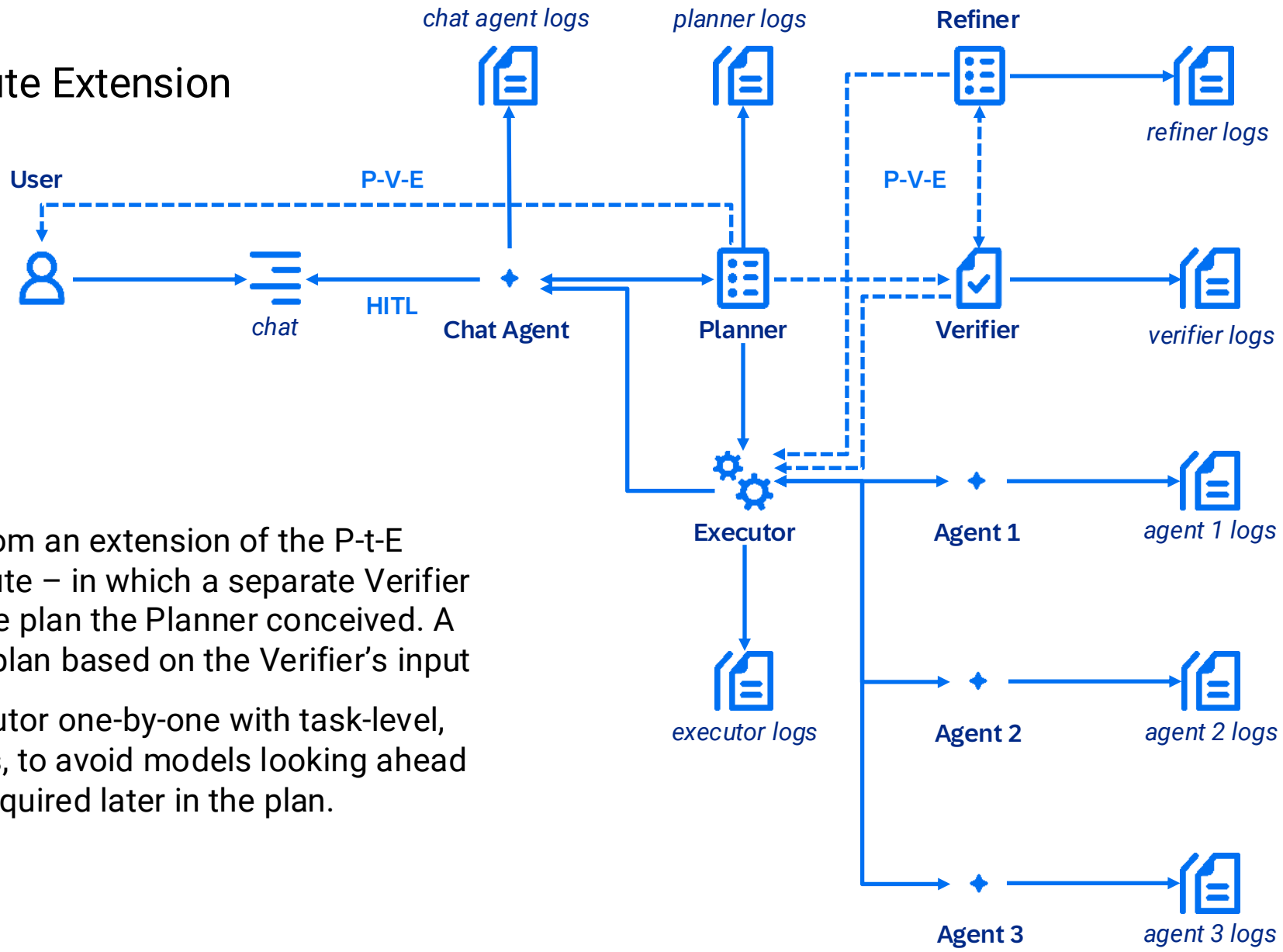


By separating planning from execution, we reduce model misaligned behavior and deception, and can use smaller models or even deterministic agents for execution

Also creates governance points for access control, and arriving at agreed and auditable outcomes

Critical Use Cases

The Plan-Validate-Execute Extension



High risk use cases benefit from an extension of the P-t-E pattern to Plan-Validate-Execute – in which a separate Verifier (agent or human) validates the plan the Planner conceived. A Refiner can further refine the plan based on the Verifier's input

Tasks are passed to the Executor one-by-one with task-level, least-privileged authorizations, to avoid models looking ahead and abuse higher privileges required later in the plan.

Mitigations



Human Control

Ensure human control and avoid agentic deception and misalignment through unconstrained persistence by agents to achieve malicious and legitimate user-prompted goals



Agent Constraint

Improve agent constraint through alignment with established business processes, separation of planning from validation and execution, and avoiding of look-ahead abuse of higher privileges needed later in the process



Cost Control

More efficiently divide agentic workloads between those that require the creativity of Frontier models and lower-cost alternatives for task execution and validation



Audit and Compliance

Aligned audit and compliance trail across agents, applications and data



Goals

Accomplishing goals within established constraints, without exceeding human-set controls and oversight, or deception what has occurred in a transaction



Fail Safe

Providing clear escape out of constrained execution loops that can’t be resolved without human judgment

Key Principles



01 An agent must never exceed the authorizations of the business user

02 Avoid deception and emergent misalignment by separating planning by reasoning models from task execution by agents

03 Constrain agents to established business processes – as through the SAP Knowledge Graph – rather than self-orchestration

04 Allow agents to fail in accomplishing a goal with conflicting constraints and escalate to human judgment

More Information on SAP Community Security & Compliance Blogs



"That's Not What We Agreed!" – Repudiation and Agentic AI Threat Modeling



Limiting Agent Autonomy – Least Privilege and Tool Access for Agentic AI Systems



Plan-then-Execute – An Architectural Pattern for Responsible Agentic AI



"Don't Lie to Me" - Containing AI Agent Threats with Plan-then-Execute

Bonus Content:



Code that Breaks the Law: Ethics and Legal Challenges in AI-generated Code

Your perspective matters

Join the conversation

You ask. We answer.



Want to Learn More?

Access Customer Resources at the SAP Trust Center here:

Thank you.

© 2026 SAP SE or an SAP affiliate company. All rights reserved. See Legal Notice on www.sap.com/legal-notice for use terms, disclaimers, disclosures, or restrictions related to this material.

SAP Bring out your best.