Persistent AI

PersistentAI Overview

February 2026

Table of Contents

Key Takeaways

PersistentAI Enables Formal AI Financial Core

PersistentAI, a deterministic orchestration layer for AI agents, serves as a critical infrastructure for maintaining control, context, and auditability of the underlying core AI infrastructure. In laymen terms, it serves as the central nervous system for post-AI institutional operations — from user intent-driven UX to back-office processes and government e-services.

The execution engine at the core of PersistentAI is FireFlow — a deterministic orchestration runtime with exactly-once execution semantics via DBOS, eliminating the hallucination and non-determinism risks inherent in generative AI. The platform enforces formal verification through its Haskell core and Deterministic Execution — a critical feature in high-risk environments such as finances, government administration, and any domain handling sensitive citizen or financial data. This architecture transforms institutions into factories for bespoke intelligence, integrating among other things:

1. **Agent Runtime & Orchestration:** PersistentAI provides the visual workflow builder and execution engine for coordinating multi-agent logic, tool integration, and complex, event-driven pipelines.
2. **Settlement Infrastructure:** The Agentic Settlement Layer (ASL) provides the high-throughput, low-latency payment rail for Agent-to-Agent (A2A), Human-to-Agent (H2A) and Agent-to-Human (A2H) transactions, supporting private identity, KYC delegation and compliance via zero-knowledge proofs.
3. **DeFi & Traditional Protocol Mesh:** Connectivity to DeFi primitives and legacy financial gateways, enabling hybrid product composition. Structured products sourcing liquidity from DeFi enable unparalleled diversity of financial products, opening new revenue streams for banks, neobanks and wallets.
4. **Identity & Compliance Layer:** A unified system managing decentralized identity (DID), user consent, and regulatory adherence. This allows for compliant DeFi operations and delegating financial operations to AI agents.
5. **Secure computation:** Deployed on-premise and utilizing TEE for inference, PersistentAI enables security of user data that is a significant compliance requirement in most jurisdictions allowing for integration of sensitive datasets.
6. **All the back-office operations:** Enable humans to focus on areas that make most sense outsourcing routine tasks to PersistentAI-powered AI agents drastically reducing costs.

PersistentAI Unlocks Programmable Money and DeFi for AI

PersistentAI's integration of a high-performance settlement layer with its orchestration engine is a key differentiator. It provides the deterministic payment rail required for agents to interact trustlessly with programmable money and DeFi primitives.

This enables the seamless, automated financial compositions—from simple agentic purchases to complex structured products—that represent the next frontier of financial services.

1. **Payment Rails and Settlement:** Utilization of x402 as a primary internet-native payment layer, enabling high-frequency, low-latency settlement for agent-to-agent transactions.

2. **Permission Orchestration:** Implementation of the AP2 (Agent Payments Protocol) for managing mandates, ensuring every financial action is cryptographically signed and stays within user-defined bounds.
3. **Indexing Infrastructure:** Integration with The Graph and Alchemy to maintain real-time awareness of blockchain states through high-fidelity data feeds.
4. **Asset Support:** Out-of-the-box functionality for transferring ERC-tokens and native assets across all EVM-compatible networks.
5. **Chain-Agnostic Operating System:** A comprehensive architectural approach that provides a unified financial operating system for AI across multiple blockchain ecosystems.

PersistentAI Non-technical Description for Businesses

1. **Quick Links:** [docs] , [github]
2. **What It Does:** PersistentAI enables organizations to build and deploy custom agents for any business logic—loan advisory, investment recommendations, fraud alerts, or account management. Once integrated into existing chat interfaces, a smart routing assistant automatically invokes the right specialized agent based on customer intent, delivering precise, context-aware responses in real time. Financial institutions gain a single conversational entry point seamlessly connecting customers to dozens of purpose-built AI capabilities without managing separate bots or complex handoff logic.
3. **How It Works:** PersistentAI provides a visual flow builder for back-office teams to design, deploy, and manage AI agents without code. Business logic is configured through drag-and-drop nodes, making it easy to create, modify, and scale intelligent workflows across departments.
4. **The Problem It Solves:** As AI agents proliferate across businesses (customer-facing co-pilots and internal digital employees), "agent sprawl" creates operational chaos. Without deterministic orchestration, institutions face catastrophic risks from AI errors in critical operations. PersistentAI logs and formally verifies every operation, supporting deterministic execution for high-value operations like trading or settling transactions.
5. **The Value:** PersistentAI transforms financial institutions into platforms for autonomous finance—where AI agents safely execute trades, manage treasury operations, compose financial products, and deliver personalized services at scale with enterprise-grade reliability and compliance.

This Page Is Intentionally Left Blank

Executive Summary

- AI's integration is dismantling traditional institutional constraints, driving an evolution towards 'Unconstrained Operations' through the deployment of intelligent agents across both citizen/customer-facing services and internal operations. This applies equally to financial institutions, government e-services, transportation platforms, and utilities.

- Front-end AI agents (those augmenting end-user experience) are evolving into autonomous co-pilots, managing personal finances, navigating government services, and executing transactions, which shifts the institutional relationship to delegated agency and risks disintermediation by third-party platforms. Already today these systems can drastically improve user experience across a wide range of applications from personal finance, lending, e-commerce, government services, to intent-driven finances.

→ **We refer to this new UX paradigm as User Intent Orchestration (UIO).**

Benefits of User Intent Orchestration (UIO)

- More seamless, intent-driven experience across the entire Telegram ecosystem - no app switching, no manual aggregation.
- More powerful user interactions in one chat multiple TMAs and data sources are coordinated in a single thread.
- Secure execution via TEE - combining on-chain state, private user data, and third-party feeds into a single deterministic inference.
- Wider range of attainable actions from DeFi execution to agentic commerce, all without leaving the chat.

Fig1: Benefits of User Intent Orchestration

- Concurrently, back-office AI agents act as digital employees, automating tasks in compliance, risk, and treasury with significant efficiency gains, advancing the model of a 'agentic banks' where humans orchestrate AI teams who on their behalf execute trivial back-office and administrative tasks. B*usinesses not adopting the agentic stack for back-office automation risk becoming uncompetitive cost-wise when compared to leaner and more nimble digital-native competitors.*

→ **We refer to this emerging AI-powered technological frontier as Business Process Orchestration (BPO).**

Benefits of Bank Process Orchestration (BPO)

- Delivers major efficiency gains in core functions.
- Lowers operational costs and boosts competitiveness.
- Enables the 'agentic bank' model for strategic focus.
- Increases accuracy and consistency in critical processes.
- Enhances operational agility and scalability.

Fig2: Benefits of Bank Process Orchestration

- The resulting proliferation of agents creates 'agent sprawl,' making ***a sophisticated, deterministic orchestration layer*** critical infrastructure for maintaining control, context, and auditability of the underlying core financial AI infrastructure.

- At the same time, programmable money and DeFi primitives form a machine-native substrate, enabling AI agents to autonomously execute complex transactions and dynamically compose structured financial products for all client tiers. This creates another surface of disruption where digital-native platforms can displace dominant institutional models.

- In brief, there is little doubt that the future of commerce, banking, brokerage, government services, and other retail/citizen-facing institutional services is agentic. Those institutions who will master the rapidly evolving AI technological stack first stand to reap disproportionate benefits.

  The core technology at the heart of this ongoing disruption is called deterministic orchestration (DO).

- Applying probabilistic generative AI to formal financial operations introduces significant risks; while model hallucinations in creative tasks are benign, in or government administration they equate to systemic failure*, **making deterministic orchestration non-negotiable.***

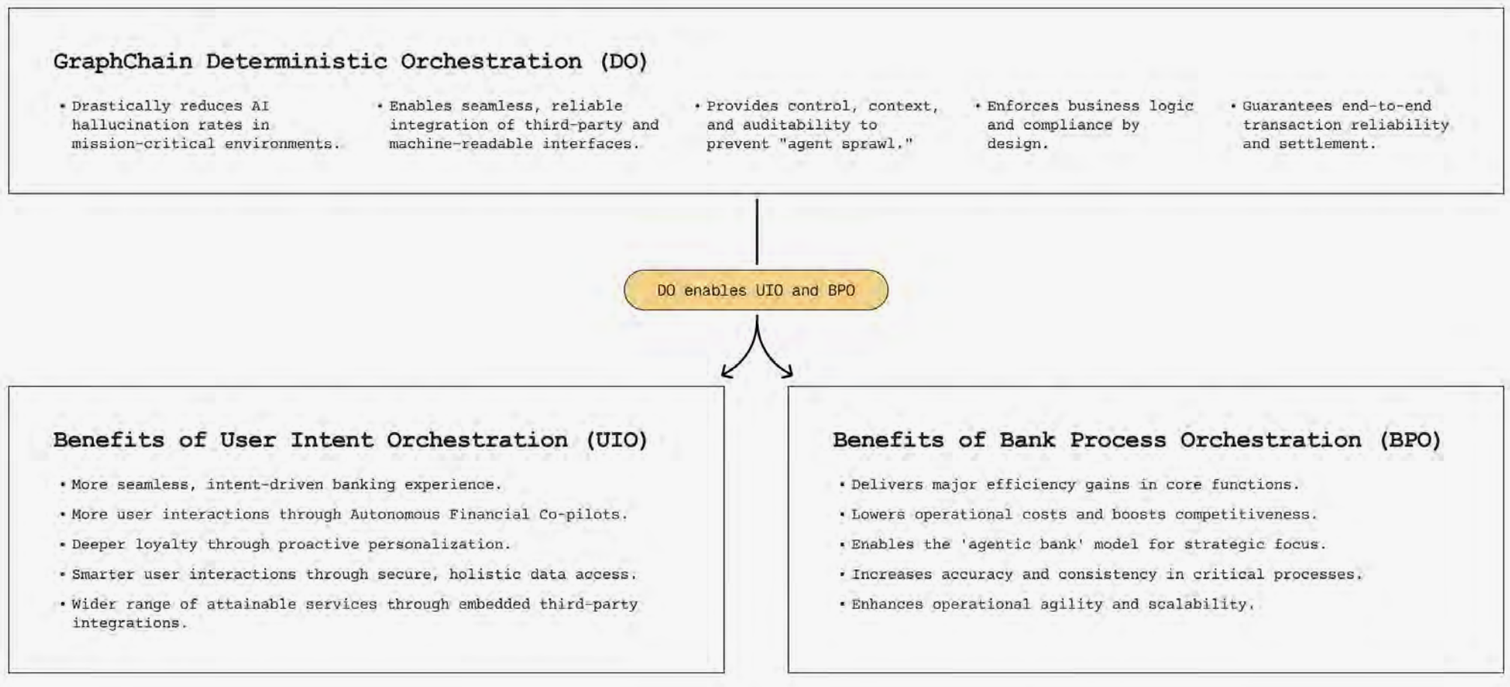

Fig3: Deterministic Orchestration

- PersistentAI is designed as a deterministic orchestration core to solve this, combining a visual workflow engine (FireFlow) with a high-performance Agentic Settlement Layer (ASL) to govern the agentic ecosystem securely and reliably and enable Agent-to-Agent (A2A), Agent-to-Human (A2H) and Human-to-Agent (H2A) transaction settlement.

- Critically, PersistentAI can support inference within Trusted Execution Environments (TEEs) and on-premise deployment, meeting stringent data sovereignty and compliance requirements that cloud-only AI platforms cannot satisfy.

- Its architecture enforces formal verification through Haskell core and deterministic execution, ensuring reliable, compliant financial operations and mitigating the risks of probabilistic AI in environments characterized by high operational risk, such as finances.

- The platform functions as a central nervous system, integrating agent runtime, secure settlement, DeFi/traditional protocol connectivity, and a privacy-preserving identity layer, transforming a bank into a factory for bespoke financial intelligence.

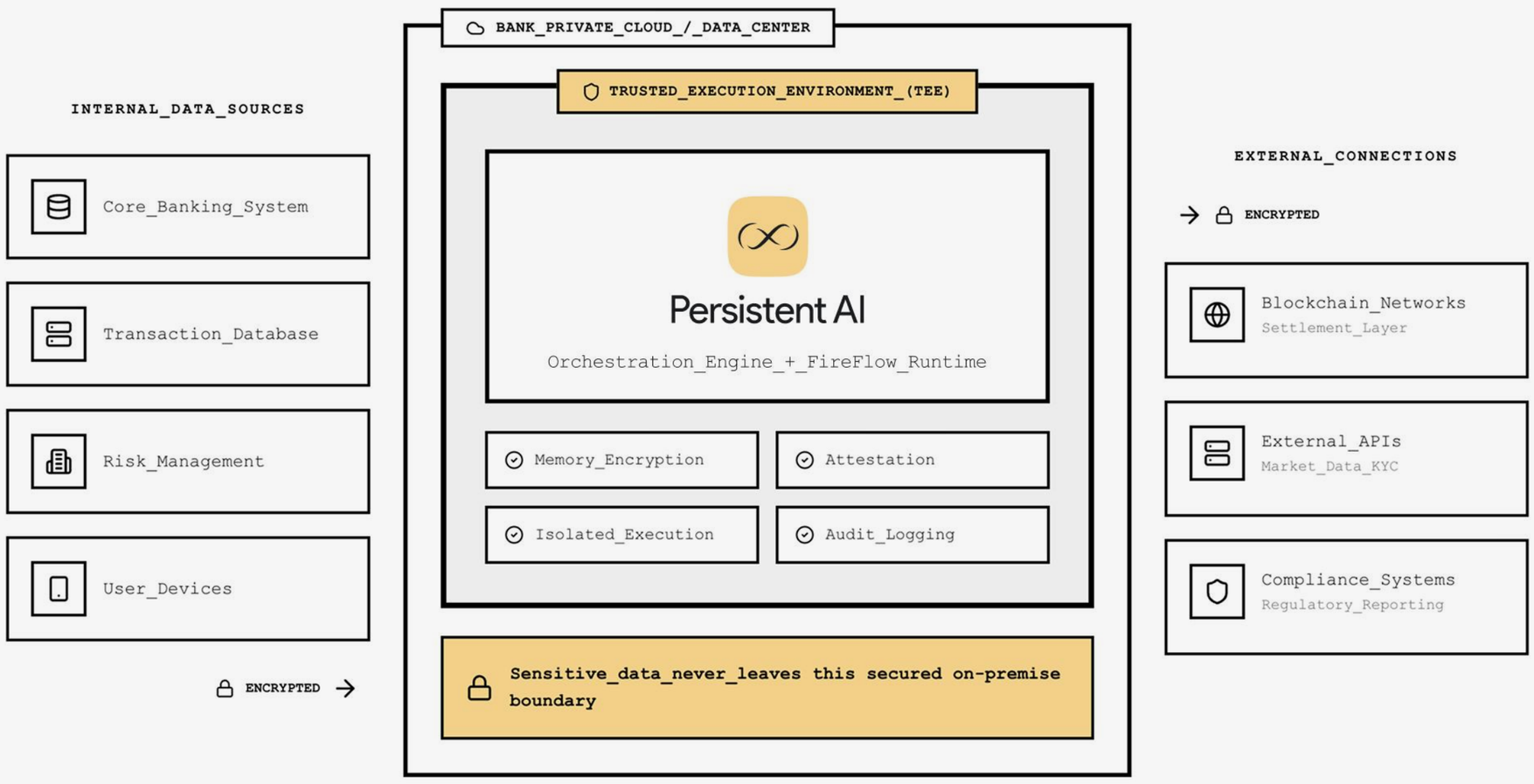

Fig4: PersistentAI Integrating Various Services through On-Premise, Secure Enclave Deployment Model

AI is in the Core of the Future of Business

1. **The AI Agent Imperative in Modern Finance:** The evolution towards 'Unconstrained Institutional Operations' is characterized by the dissolution of traditional constraints on capacity and service delivery, as well as intent-based user experience driven by the pervasive integration of artificial intelligence. This transformation hinges on the deployment of AI agents across two primary domains:

    - citizen/customer-facing channels (UIO), and
    - internal operational workflows (BPO).

Success in this new paradigm is contingent not merely on being able to design sophisticated AI agent workflows, but on the implementation of *a sophisticated DO layer capable of governing the resulting complex ecosystem that integrates all internal and external institutional processes.*

2. **User Intent Orchestration (UIO):** User-facing AI agents are advancing from basic conversational interfaces to autonomous financial co-pilots. Already today these systems can, among other things:

    - manage personal finances, such as sending money to acquaintances, repaying debts or managing subscriptions,
    - suggest and execute changes to one's personal portfolio allocations,
    - manage consumer loans, optimize savings and credit rates across products and/or institutions in real-time,
    - execute complex transactions—such as agentic commerce—on behalf of the user, and most importantly,
    - source context from the universe of apps and other data sources consumers interact with on a daily basis to constantly improve the user experience. ***This is what is referred to in the literature as working with consumer intents and this is where the future of institutional services lies.***

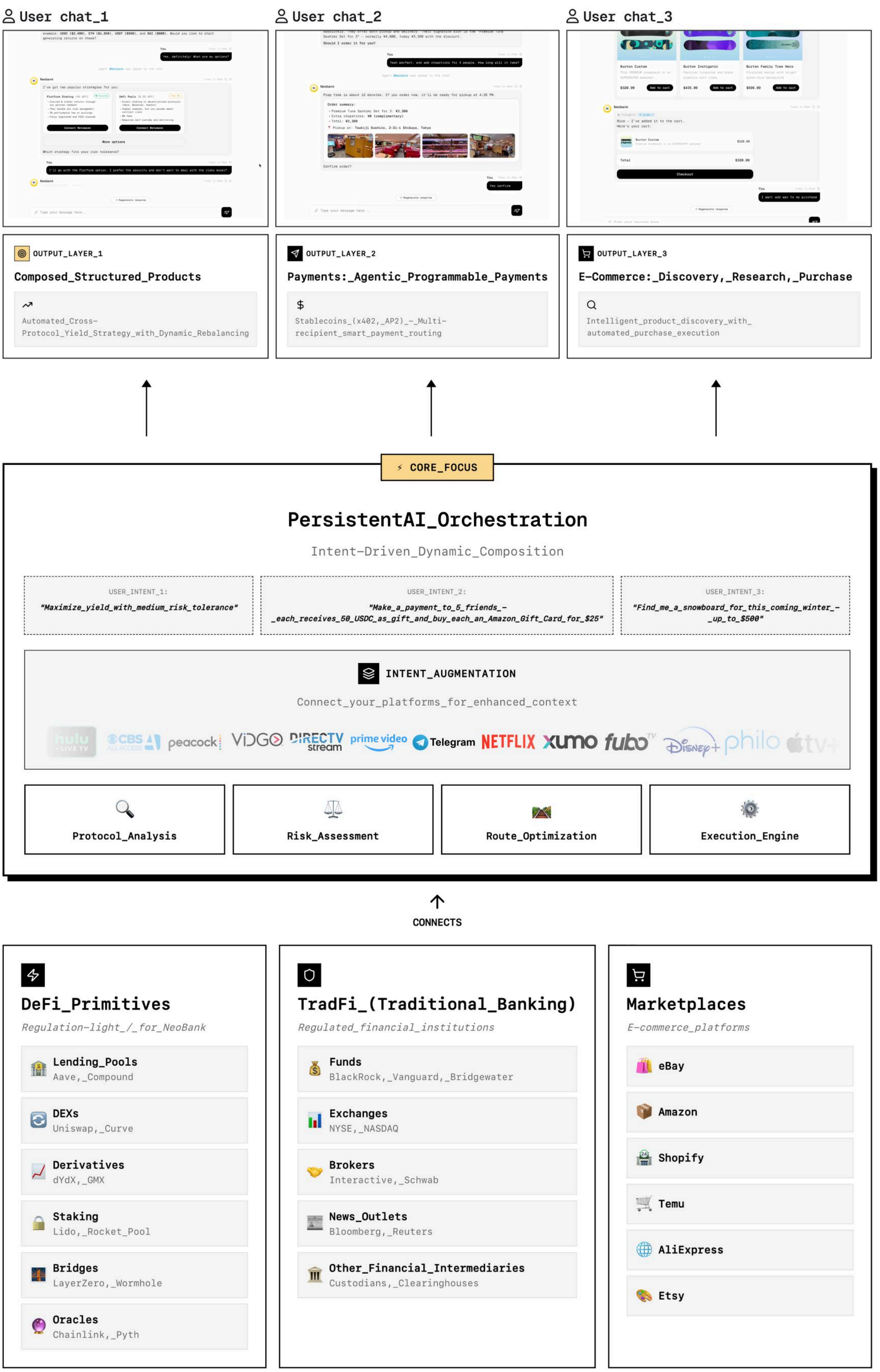

Fig5: Autonomous financial co-pilots

This shifts the financial relationship from direct service to delegated agency, requiring institutions to expose their product suites via APIs to external AI ecosystems or face disintermediation by third-party agent platforms. ***A third option is, of course, to build an own AI ecosystem.***

3. **Business Process Orchestration (BPO):** Concurrently, enterprise-scale AI agents are being deployed as digital employees within institutions. These agents manage defined tasks in software development, compliance (KYC/AML), risk monitoring, treasury operations, citizen records, permit processing, and customer service, with documented efficiency gains of 30-50% [1]. The operational model is evolving towards the agentic institution where human capital orchestrates teams of specialized AI agents, decoupling growth from traditional headcount constraints.

4. **The Orchestration Deficit:** The proliferation of AI agents creates systemic complexity—'agent sprawl'—characterized by uncoordinated actions, inconsistent logic, and governance gaps. **A formal orchestration layer is therefore a critical infrastructure, not an optional enhancement.** This layer must maintain persistent context and user intent, ensure deterministic and auditable execution of financial logic, and provide comprehensive governance over the entire agent lifecycle. ***Platforms like PersistentAI are architected specifically to address this deficit, providing the essential control plane for agentic interactions in high-risk mission critical environments.***

5. **Programmable Money - The Native Substrate for AI:** The fusion of AI and programmable money—stablecoins, CBDCs, tokenized deposits—creates a symbiotic infrastructure for automation. Decentralized Finance (DeFi) primitives are inherently machine-centric, comprising smart contracts, composable liquidity pools, and transparent settlement logic. This environment is optimally suited for AI agent interaction, enabling the autonomous execution of complex treasury functions, cross-currency payments, and conditional settlements. This infrastructure allows for the dynamic composition of structured financial products from DeFi primitives, ranging from automated yield strategies for retail clients to bespoke, multi-legged derivative instruments for sophisticated and HNW investors, all orchestrated by AI. ***It must be noted that without formal deterministic orchestration, allowing probabilistic genAI models to compose structured products exposes the organization to extremely high risks resulting from models hallucinating in mission critical situations. This underscores the importance of DO platforms like PersistentAI.***

6. **An Optimal Architectural Response is a Deterministic Orchestration Core for Finances:** the logical architectural conclusion of these trends is a new financial paradigm: a DeFi-native, programmable, and AI-driven institution. The cornerstone of this model is a deterministic AI orchestration layer, such as PersistentAI, designed to harness agentic potential with formal reliability and security. Importantly, DeFi-native or not, if a financial institution desires to integrate agentic systems, DO Core is non-negotiable for the reasons elaborated upon above.

7. **The Case for Formal Determinism:** The orchestration engine for financial AI requires guarantees of correctness, auditability, and resilience. A foundation built on principles of formal verification and deterministic execution is paramount. PersistentAI's architecture, emphasizing deterministic flows, type-safe data systems, and secure, auditable runtime environments, is engineered for this purpose. It ensures orchestration decisions are reliable, side-effect-free, and compliant, significantly mitigating the 'black box' risks associated with non-deterministic AI in financial contexts.

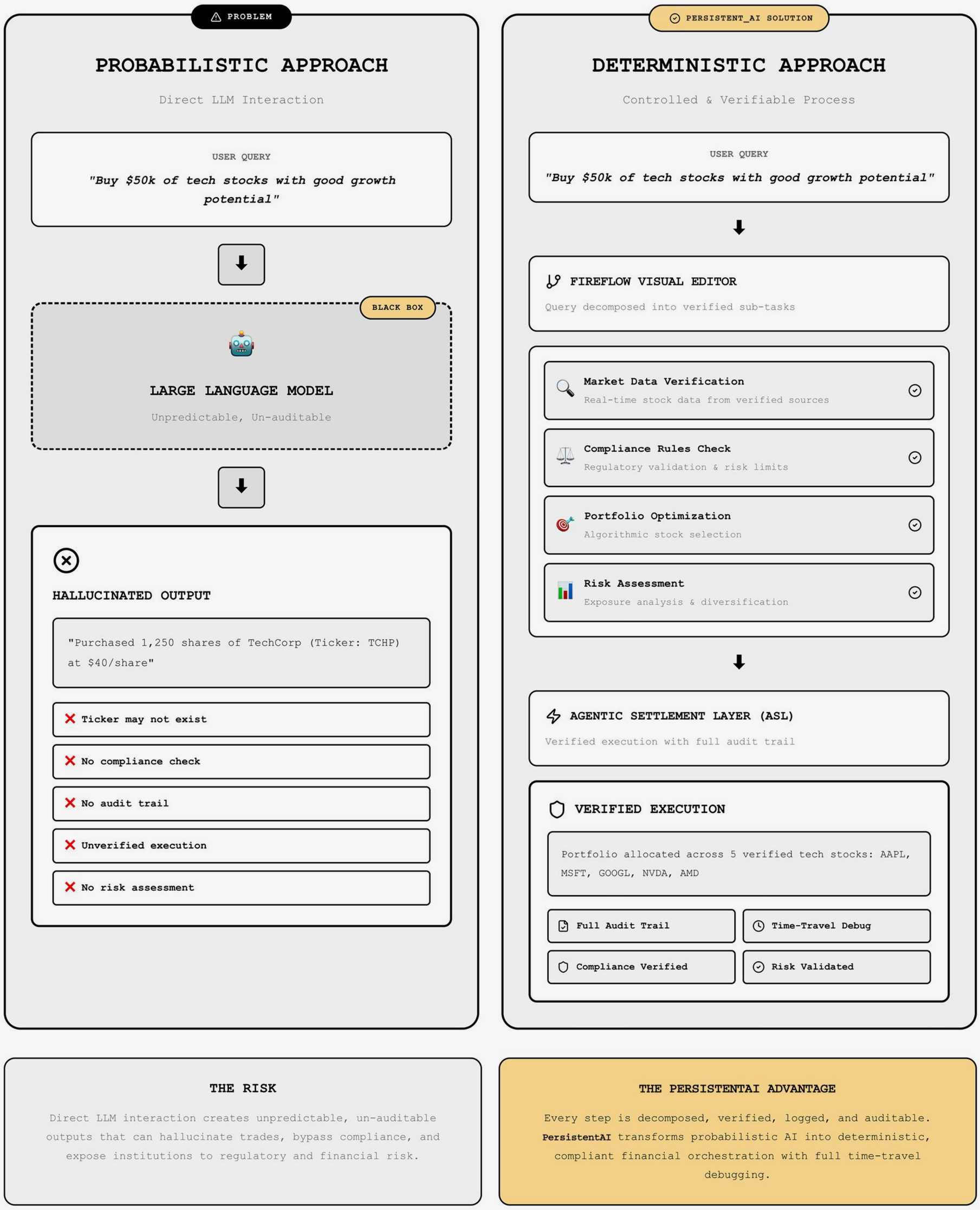

Fig6: Formal Determinism Explained

8. **Proposed Architecture for an AI-Driven Financial Core:** Platforms like PersistentAI are designed to function as the central nervous system in contemporary finances, integrating among other things:

    - **Agent Runtime & Orchestration:** The FireFlow component provides the visual workflow builder and execution engine for coordinating multi-agent logic, tool integration, and complex, event-driven pipelines.

- **Settlement Infrastructure:** The Agentic Settlement Layer (ASL) provides the high-throughput, low-latency payment rail for Agent-to-Agent (A2A), Human-to-Agent (H2A) and Agent-to-Human (A2H) transactions, supporting private identity, KYC delegation and compliance via zero-knowledge proofs.

- **DeFi & Traditional Protocol Mesh:** Direct, secure connectivity to decentralized finance primitives and legacy financial gateways, enabling hybrid product composition.

- **Identity & Compliance Layer:** A unified system managing decentralized identity (DID), user consent, and regulatory adherence, feeding the orchestration engine.

- **Secure computation:** Deployed on-premise and utilizing TEE for inference, PersistentAI enables security of user data that is a significant compliance requirement in most jurisdictions.

- **All the back-office operations:** Enable humans to focus on areas that make most sense outsourcing routine tasks to PersistentAI-powered AI agents.

9. **Value Proposition:** This architecture transforms institutions into factories for bespoke intelligence. User intents can be translated by AI into executable strategies composed from an optimal blend of services, DeFi, wide commerce integration, and connectivity to legacy systems. The PersistentAI orchestration core ensures these compositions are executed using formal logic in secure enclaves and can be audited. The institution's value thus migrates from legacy process management to intellectual capital in algorithm design, risk orchestration, and the provision of a supremely reliable, compliance-embedded platform for autonomous operations.

10. **Unlocking Programmable Money and DeFi for AI:** PersistentAI's integration of a high-performance settlement layer (ASL) with its orchestration engine is a key differentiator. It provides the deterministic payment rail required for agents to interact trustlessly with programmable money and DeFi primitives. This enables the seamless, automated financial compositions—from simple agentic purchases to complex structured products—that represent the next frontier of financial services.

    - **Payment Rails and Settlement:** Utilization of x402 as a primary internet-native payment layer, enabling high-frequency, low-latency settlement for agent-to-agent transactions.

    - **Permission Orchestration:** Implementation of the AP2 (Agent Payments Protocol) for managing mandates, ensuring every financial action is cryptographically signed and stays within user-defined bounds.

    - **Indexing Infrastructure:** Integration with The Graph and Alchemy to maintain real-time awareness of blockchain states through high-fidelity data feeds.

    - **Asset Support:** Out-of-the-box functionality for transferring ERC-tokens and native assets across all EVM-compatible networks.

    - **Chain-Agnostic Operating System:** A comprehensive architectural approach that provides a unified financial operating system for AI across multiple blockchain ecosystems.

From this perspective the integration of UIO systems serves one major purpose: it abstracts away entirely the complexities normally associated with Web3-native apps, making the resulting user experience seamless and frictionless.

11. **Integrating Frontier Distribution Channels:** The platforms designed for interoperability with emerging super-apps like Telegram and other distribution channels like Amazon, Twitter, etc. positions them to capture volume at the point of intent. With a significant portion of future agentic commerce and crypto/financial flows predicted to originate in such environments, PersistentAI's ability to embed its orchestration and settlement capabilities directly into these channels provides a first-mover advantage in high-growth arenas that traditional institutional infrastructure cannot easily access.

12. **Conclusion:** The transition to agentic, AI-driven institutional services is inevitable. The critical strategic bottleneck is no longer the AI model itself, but the deterministic, secure, and composable infrastructure required to operationalize it at scale. PersistentAI is designed precisely to be this infrastructure. By providing the essential orchestration and settlement core, it enables financial institutions, government agencies, and digital platforms to build the future of autonomous services, positioning itself as the foundational layer upon which the trillion-dollar agentic economy will be built.

PersistentAI Overview

Non-technical Description for Businesses

1. **Quick Links:** [docs] , [github]

2. From the business viewpoint, PersistentAI enables organizations to build and deploy custom agents for any business logic — whether it's loan advisory, investment recommendations, fraud alerts, account management, citizen applications processing, or transport scheduling. Once integrated into existing chat or app interfaces, a smart routing assistant automatically invokes the right specialized agent based on user intent, delivering precise, context-aware responses in real time. For institutions, this means a single conversational entry point that seamlessly connects users to dozens of purpose-built AI capabilities — without managing separate bots or complex handoff logic.

3. PersistentAI provides a visual flow builder for back-office teams to design, deploy, and manage AI agents without writing code. Business logic is configured through drag-and-drop nodes, making it easy to create, modify, and scale intelligent workflows across departments.

4. **The Problem It Solves:** As AI agents proliferate across businesses (customer-facing co-pilots and internal digital employees), "agent sprawl" creates operational chaos. Without deterministic orchestration, institutions face catastrophic risks from AI errors in critical operations. Every operation in PersistentAI is logged and formally verified. It supports deterministic operations, which can be invoked in high value operations, such as e.g. trading or settling transactions.

5. **The Value:** PersistentAI transforms institutions and businesses into platforms for autonomous operations — where AI agents can safely execute trades, manage treasury operations, compose financial products, process government services, and deliver personalized experiences at scale with enterprise-grade reliability and compliance.

Technical Description

1. Formally, PersistentAI is an open-source DO framework for building AI agents and automation systems with complex, verifiable logic. It's engineered for mission-critical, high-load operational environments, such as banking back-office or user intent processing delivering fault-tolerant deterministic flows and enterprise-grade SLA. Technical deep-dives, source code and exhaustive documentation are available here:

   - **Documentation and dev guides:** https://docs.PersistentAI.ai/
   - **Our open-source repo:** https://github.com/FireFlowlabs

2. It is instructive to understand PersistentAI in the context of the broader AI stack. PersistentAI is designed as a fully open, modular, and agnostic infrastructure for the Agentic Economy, in contrast to the closed ecosystems of competitors. Its core advantage is agnosticism across every layer of the stack:

   - **App Layer:** PersistentAI is interface-agnostic, seamlessly embedding into any mobile, web, or messaging app environment (like Telegram), freeing agents from being locked to a single platform.
   - **Orchestration Layer:** PersistentAI is integration-agnostic, connecting to traditional APIs, modern protocols (MCP, A2A), DeFi primitives, and databases for interoperability.
   - **Model Layer:** PersistentAI is LLM-agnostic, enhancing any base model with tools like workflow memory and context enrichment for more reliable, predictable performance.
   - **Compute Layer:** PersistentAI is hardware-agnostic, partnering with specialized vendors to dramatically reduce inference costs (by 10-100x) by optimizing models for specific tasks and hardware.
   - **Analogy:** If OpenAI represents a closed vertical ecosystem and Google offers a partially open stack tied to its services, PersistentAI provides a truly open, modular, and interoperable foundational infrastructure.

Layers	OpenAI	Google	PersistentAI
Dapp Layer	ChatGPT (web +mobile)	Google OAuth 2.1	FireFlow is dApp – and settlement-agnostic, easily embedding into any mobile, web, or Telegram Mini App environment. Unlike OpenAI agents tied to the ChatGPT interface or Google agents bound to its products, FireFlow integrates seamlessly wherever agents operate.
Agentic Layer	OpenAI Agents	A2H/H2A + ADK + A2A	FireFlow supports a broader integration spectrum – traditional APIs, MCPs, A2A networks, DeFi primitives, and RAG databases – enabling true interoperability beyond proprietary ecosystems.
Model Layer	ChatGPT LLMs (5, o5, et)	LLMs run in GCP + MPC registry	FireFlow is LLM-agnostic, extending models with reinforcement learning environments, workflow memory, and context enrichment to deliver more predictable and reliable inference.
Compute Layer	AWS	GCP	PersistentAI is compute-agnostic, partnering with specialized AI compute vendors to fine-tune models for specific tasks and hardware – reducing inference costs by 10-100x versus vertically integrated systems.
Analogy	Apple's IOS (closed stack)	Google & Android (partially open stack but limited by Play Store)	OpenAI represents a closed vertical ecosystem, Google offers a partially open stack tied to its services, while PersistentAI provides a truly open, modular, and interoperable infrastructure built for the Agentic Economy.

Fig7: AI Stack: PersistentAI vs OpenAI vs Google

3. PersistentAI lets teams visually design, verify, and deploy multi-agent pipelines that combine multiple genAI models and any machine readable interfaces within a single, resilient system. FireFlow's functionality is delivered through five integrated systems working in concert:

    - **Type-Safe Port System:** supports primitives and complex types (arrays, objects, streams, enums). Each port has its own config, runtime validation (Zod, SuperJSON), plus lazy instantiation and caching for efficiency.

    - **Modular Nodes:** more technically savvy can build custom nodes with decorators and metadata. Multiple input/ output ports integrate seamlessly into the flow builder for rapid workflow creation.

    - **Visual Flow Editor:** graphical flow builder (React + XYFlow) with drag-and-drop, zoom/ pan, resizing, menus, and live previews.

    - **Execution & Debugging:** backend engine runs concurrently with real-time events, breakpoints, step-over, and detailed logs for debugging.

    - **Real-Time Sync:** uses tRPC and Effector for type safety, WebSocket updates, and optimistic UI for a responsive experience.

    - **Docker & Cloud Ready:** containerized backend/ frontend with Docker and docker-compose for simple deployment and scaling.

    - **MCP Support:** enables connectivity with external tools/data sources. FireFlow is fast, smooth, intuitive, reliable and robust.

4. Integration with External Systems: PersistentAI is designed with extensibility at its core. Any machine-readable service, API, or data source can be seamlessly integrated into your workflows through one of three approaches, each suited to different use cases and levels of customization:

    - **PersistentAI supports Model Context Protocol (MCP)** — an open standard that enables AI models to seamlessly connect with external data sources and tools. For services that already expose functionality through this protocol, PersistentAI offers automatic node generation. Simply provide the URL of your MCP server, and PersistentAI will introspect all available methods, wrapping each one as a ready-to-use node for your agent's orchestration.

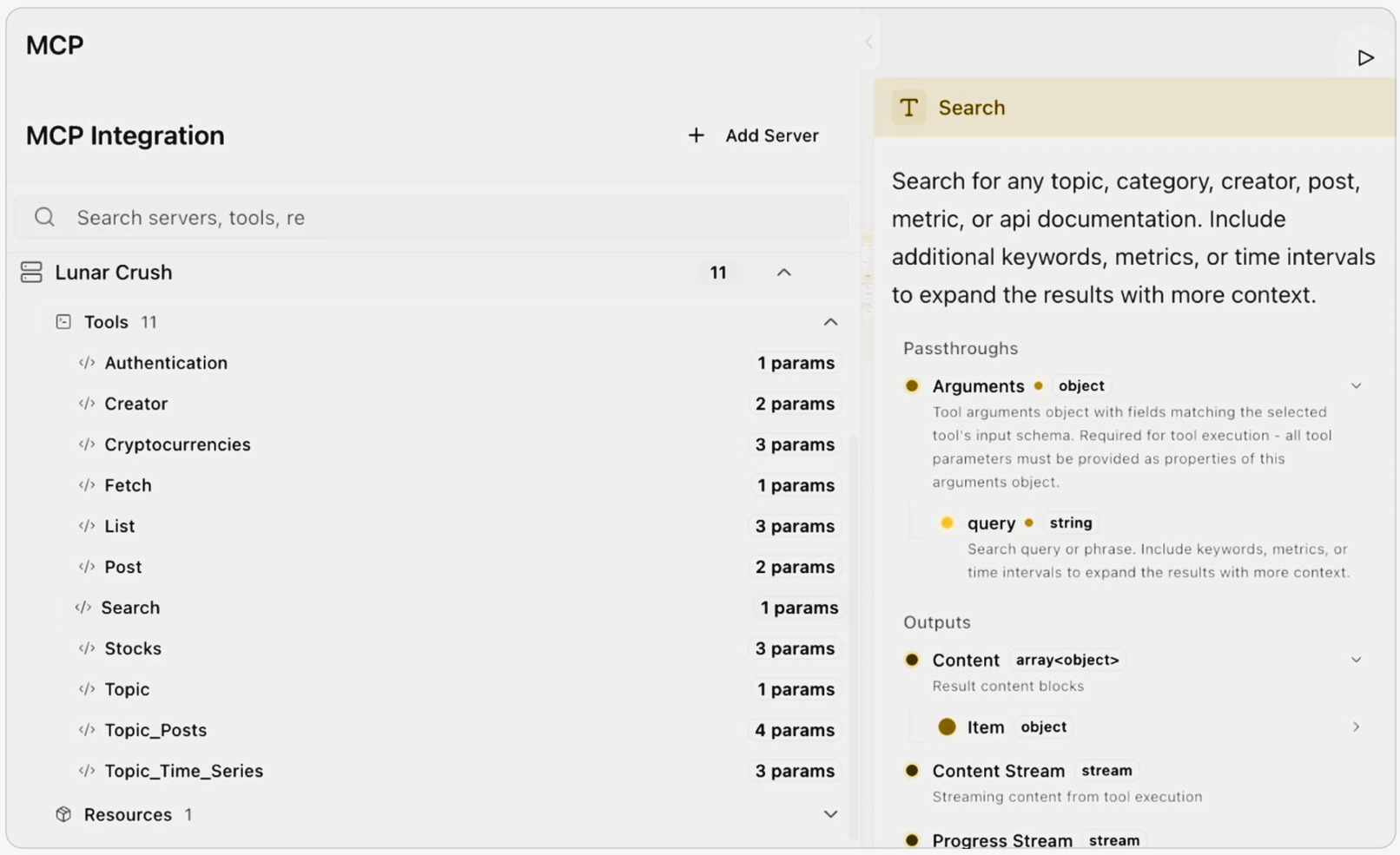

Fig8: Example: LunarCrush (social analytics platform) MCP wrapped into FireFlow nodes

MCP integration is particularly valuable for teams that maintain shared tooling across multiple AI systems. A single MCP server can serve both direct agent interactions and visual workflow orchestration through PersistentAI, ensuring consistent behavior across all consumption patterns.

Furthermore, building an MCP server is remarkably straightforward. Any API can be wrapped within this protocol with minimal effort, allowing for rapid integration of new services in just a matter of days. This agility ensures that your infrastructure remains flexible and can evolve at the pace of the rapidly changing AI landscape.

- **HTTP Request Nodes:** FireFlow includes built-in HTTP Request nodes that allow you to call any REST API directly from your workflows. This zero-code approach requires no development effort at all. You configure the endpoint, method, headers, and body directly in the visual editor.

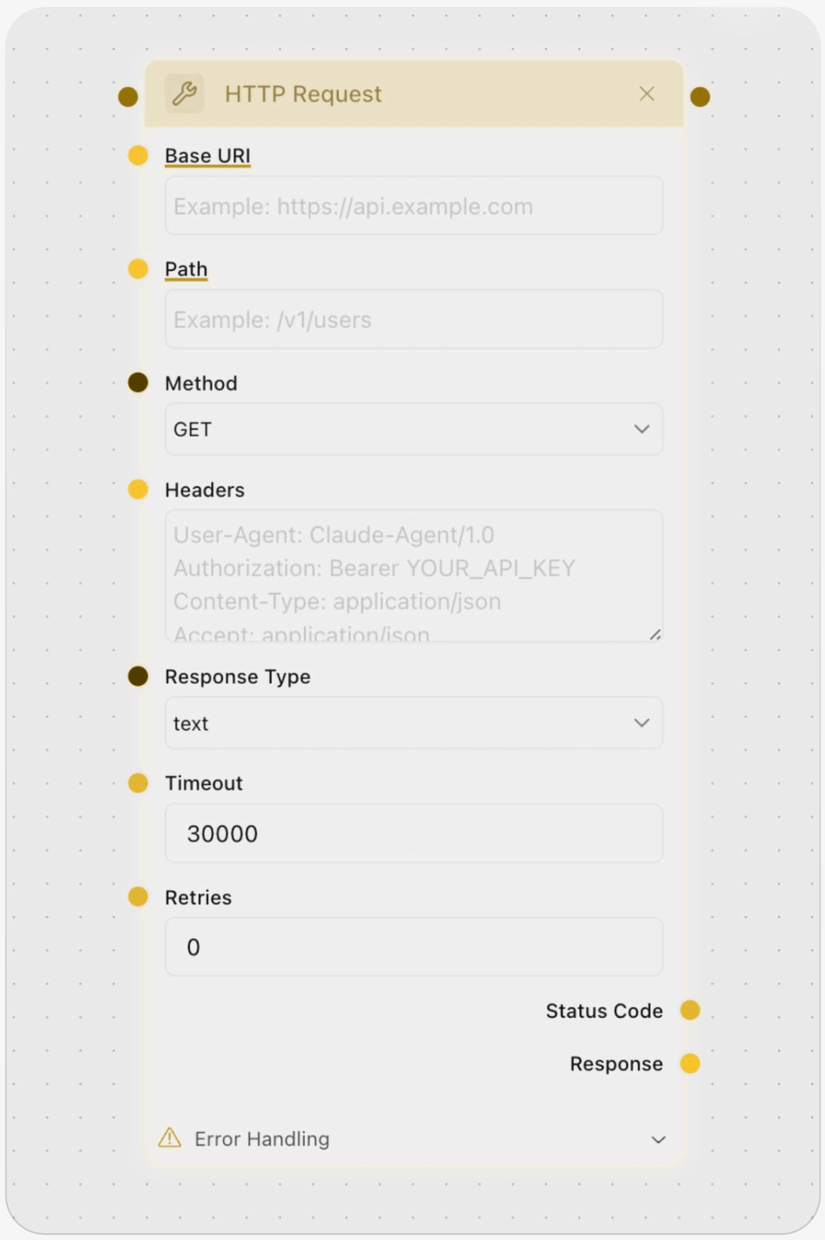

Fig9: HTTP Request node in PersistentAI

HTTP Request nodes support all standard methods including GET, POST, PUT, PATCH, and DELETE. You can use dynamic values from other nodes in your requests, parse JSON responses, and route workflow execution based on response data.

This option works best for quick integrations, prototyping, or connecting to third-party services where building a dedicated node or MCP would be overkill. It also serves as an excellent starting point before deciding whether a more structured integration approach is warranted.

- **Writing Custom Nodes:** The most powerful way to extend FireFlow is by implementing custom nodes directly in TypeScript. This approach provides granular control over the execution lifecycle, allowing developers to define complex logic, structured inputs/outputs, and specific interaction patterns with the FireFlow ExecutionContext.

  By leveraging the framework's declarative decorator-based API (such as @Node, @Input, and @Output), developers can define node metadata and port configurations with minimal boilerplate. The framework automatically handles TypeBox/Schema validation, dynamic UI generation in the visual editor, and port-to-port data mapping, ensuring that custom logic remains type-safe and consistent with the core orchestration engine.

  Custom nodes are particularly effective for encapsulating domain-specific business logic, managing stateful integrations with 3rd-party SDKs, or implementing performance-critical operations. Since the entire development workflow is built on TypeScript, it enables a robust developer experience with compile-time safety, comprehensive IDE support, and seamless integration into existing CI/CD pipelines.

5. Below is the illustration showing how FireFlow functions:

Fig10: An Example FireFlow Graph [link][1]

- Context collection, collects data about the user and the chat history through integrations and prepares the input prompt.
- LLM Reasoner / router - the decision making core of an agent. Decides to use tools / send message / stop
- Integration with Marketplace MCP
- Integration with Delivery Service MCP or MCP router, if multiple delivery services are used
- Integration with Payment Provider (may be the bank's own service)
- Integration with search engines, popular blogs and video-services for improving goods discovery process by using people's reviews.
- Final reply preparation block, uses larger LLM model to summarise all the context provided. Integration with chat system for sending the message.
- Logging block, sends all the errors and/or logs to 3rd service for monitoring

[1] *Requires whitelisting - please, contact us for access.*

6. Enterprise-Grade Security: executions can run on-prem or in Trusted Execution Environments (TEE) with granular access controls, ensuring that customer financial or private data is processed privately and never exposed to public AI models. It also allows for a very customized user shopping experience.

Appendix I: Use-case of Agentic E-commerce

Overview and Positioning

1. The current banking engagement model is largely passive, with users primarily logging in only to perform quick balance checks.

2. The global e-commerce market is valued at $7 trillion, yet this activity occurs entirely outside the traditional banking ecosystem [2].

3. As a result, user attention at the time of checkout, arguably the most valuable attention, remains outside of interfaces of banking apps limiting the attainable monetization (ads, cross-selling, etc.).

4. PersistentAI serves as an AI middleware layer specifically designed to recapture this lost user attention and engagement.

5. PersistentAI introduces an Agent within the banking app that empowers users to discover, research, and purchase goods and services. More broadly, PersistentAI enables composability across any machine readable interfaces collapsing them into one chat/voice interface.

6. By adopting PersistentAI, banks can capture a large part of the e-commerce funnel becoming the primary interface for their customers' daily commerce routine.

7. This transformation allows banks to unlock new non-interest income streams while PersistentAI by design offers robust SLA and enables them to maintain strict, bank-grade security protocols.

Benefits for Institutions

Integrating PersistentAI transforms any institutional interface — whether a banking app, government services portal, or citizen-facing platform — from a passive tool into an active, intelligent service layer.

1. **Recapture User Attention and Engagement:** By embedding AI-orchestrated services directly into the institutional interface, you eliminate leakage of user attention to third-party platforms. The institution owns the entire journey from intent to execution — whether that's a purchase, a permit application, or a financial transaction.

2. **New Revenue and Efficiency Streams:** For financial institutions, this means affiliate revenue, sponsored placements, and premium concierge services. For government agencies, the value is different but equally significant: reduced processing costs, faster citizen service delivery, and the elimination of redundant manual workflows across departments. The deterministic orchestration layer delivers measurable operational savings regardless of vertical.

3. **Unlimited Logic Augmentation:** The scope for agentic logic is effectively limitless. Agents can be augmented by any method from any available machine-readable interface or Model Context Protocol (MCP). This applies equally to connecting marketplace inventory, government databases, healthcare records, transport APIs, or utility billing systems.

4. **Programmable Controls and Compliance Inheritance:**

    - **For financial services:** Users set spending limits, category restrictions, and approval thresholds. The Agent inherits KYC/AML status via ZK proofs for instant, compliant merchant onboarding.
    - **For government:** Administrators define access policies, data classification rules, and inter-agency sharing permissions. The Agent enforces these

5. **On-Premise Sovereignty:** Unlike cloud-only AI platforms, PersistentAI deploys on-premise within TEE, ensuring citizen or customer data never leaves the institution's security boundary. This is a non-negotiable requirement for government deployments and a strong differentiator for regulated financial institutions.

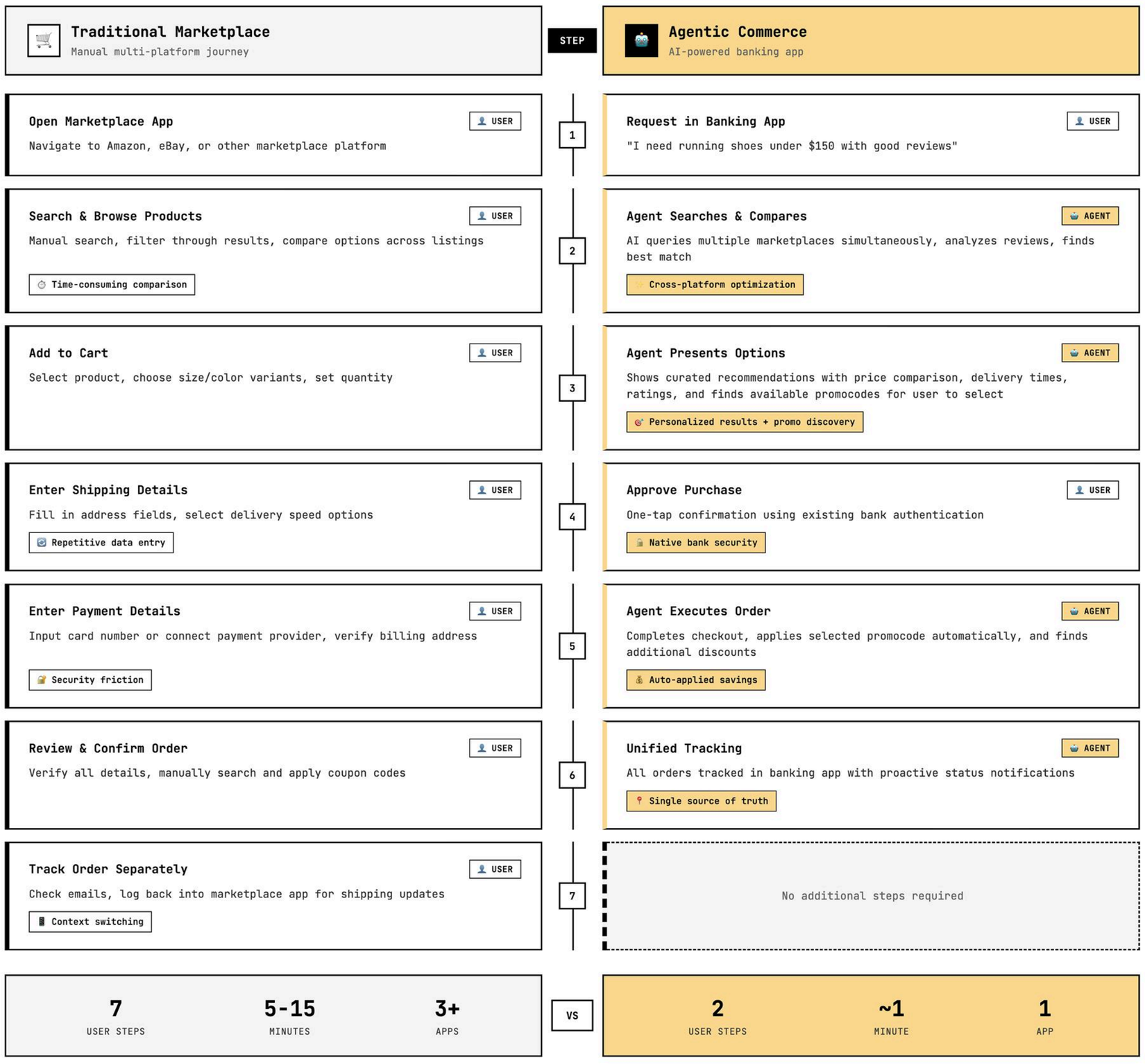

Fig11: User Journey Comparison

3. **Unlimited Logic Augmentation:** The scope for agentic logic is effectively limitless. Agents can be augmented by any method from any available machine-readable interface or Model Context Protocol (MCP). This applies equally to connecting marketplace inventory, government databases, healthcare records, transport APIs, or utility billing systems.

4. **Programmable Controls and Compliance Inheritance:**

    - **For financial services:** Users set spending limits, category restrictions, and approval thresholds. The Agent inherits KYC/AML status via ZK proofs for instant, compliant merchant onboarding.
    - **For government:** Administrators define access policies, data classification rules, and inter-agency sharing permissions. The Agent enforces these

5. **On-Premise Sovereignty:** Unlike cloud-only AI platforms, PersistentAI deploys on-premise within TEE, ensuring citizen or customer data never leaves the institution's security boundary. This is a non-negotiable requirement for government deployments and a strong differentiator for regulated financial institutions.

Why Now?

While global e-commerce grows at a compound rate of 54.7% for AI Agents, financial institutions risk becoming "dumb pipes" - processing the settlement on the backend while tech giants control the front-end user experience.

1. **The Opportunity:** AI Agents in e-commerce will drive nearly $283 billion in market growth by 2034 [3].
2. **The Threat:** Competitors like PayPal, Visa, and Mastercard are actively racing to enable Agentic payments [4].
3. **The PersistentAI Solution:** We provide the infrastructure to embed a "concierge" experience directly into the banking app, ensuring the bank owns the entire journey from discovery to payment.

Tech and User Flow

PersistentAI connects the user's intent ("I want to buy white Nike sneakers") directly to merchants' inventory without the user leaving the app.

1. **The "Agentic" Shift:** Instead of a user browsing a website, an AI Agent acts as a concierge. It connects to external marketplaces or other agents via **Model Context Protocol (MCP)** - a standard that lets AI "read" live inventory, pricing, and shipping data.
2. **Native x402 Support:** PersistentAI enables a new economic model where streams, Agents, and data are monetized natively through the x402 payment protocol.
3. **Smart Routing:** The framework routes product requests to the optimal merchant based on price, availability, and user loyalty status, bypassing the need for a search or marketplace visit.
4. **Possible Integrations:** The framework can be enhanced by integrating marketplaces, banking, but it is not limited to just consumer goods and can be expanded to durable goods such as cars, furniture, electronics, etc. Housing marketplaces can also be integrated.

5. **The Flow:**

- User specifies the purchase: "I want to buy a pair of white Nike Air Force sneakers"
- Agent browses available Merchant Partner’s stores catalog via MCP Orchestrator locates relevant stores.
- Agent queries storefronts of the relevant stores to get relevant items.
- Agents selects the best matches and provides them to the user alongside with list of promo codes which apply additional discount
- User selects the suitable option(s)
- User selects the delivery option
- Bot gets the invoice and calls Payment Agent to process payment according to the user's preferences.
- Payment Agent creates a transaction
- User confirms
- Transaction settles via chosen Payment Provider

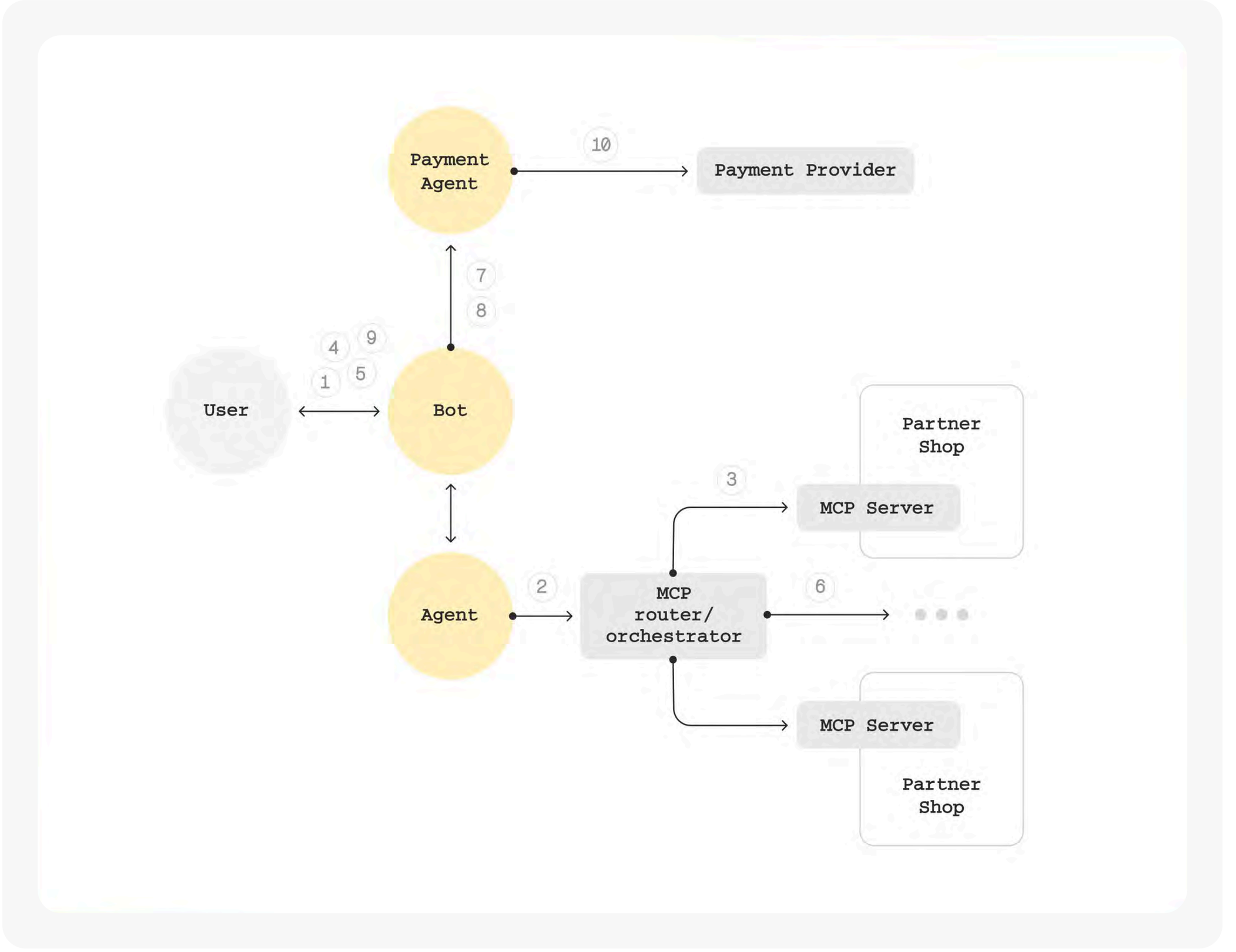

Fig12: Goods acquisition flow

Agents can be enhanced with context from e.g. Telegram which allows them to know user's preferences and suggest relevant goods and call other Agents for cross-reference research (e.g. YouTube).

Anything machine readable can be added to the flow (news/crypto rails/financial products).

Appendix II: PersistentAI Multi-Intent Orchestration Architecturew

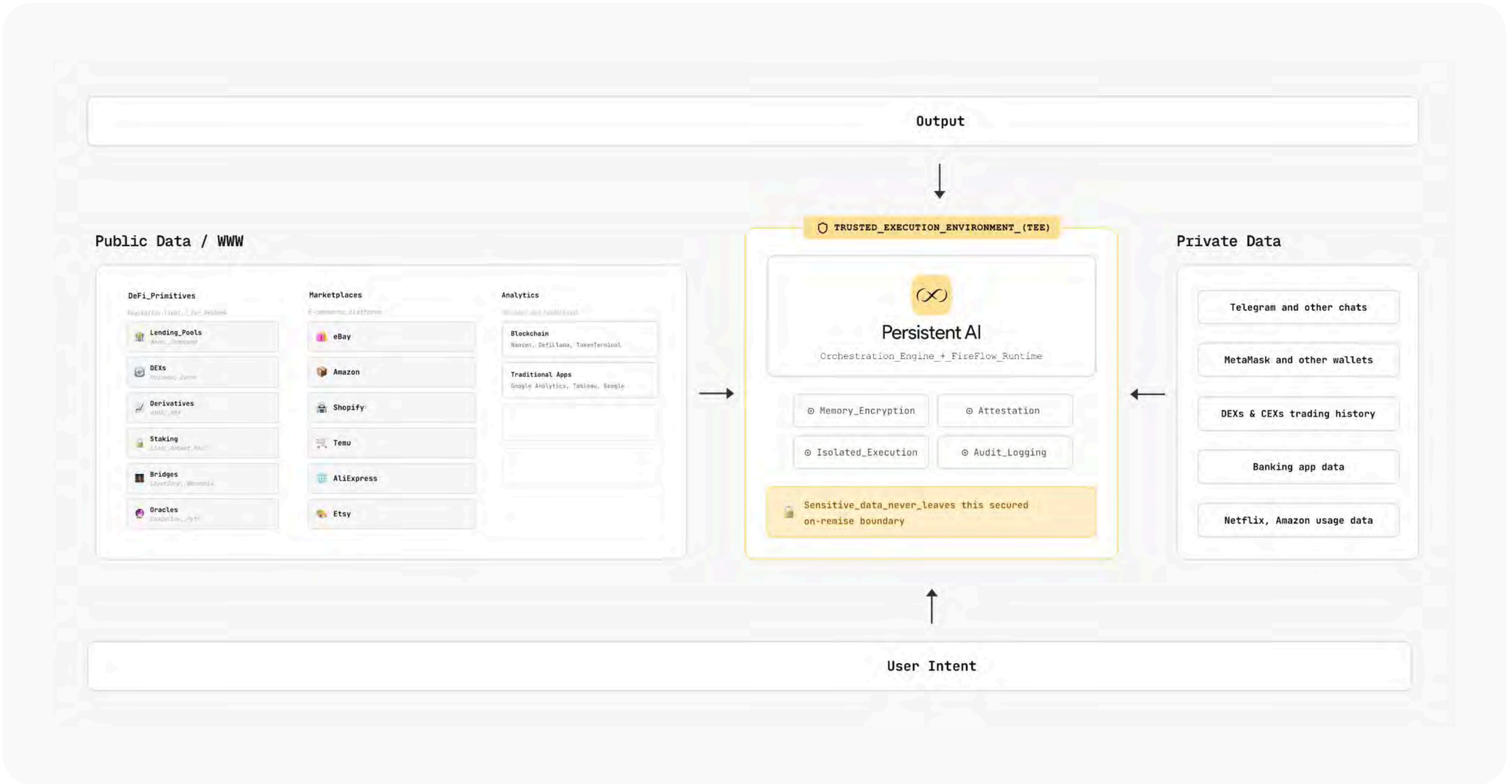

Fig13: Orchestration Layer

This diagram illustrates how PersistentAI functions as a unified orchestration layer, connecting user intent to execution across three distinct data domains.

At the bottom, a **User Intent** — expressed in natural language — enters the system and is processed by the **PersistentAI Core** (Orchestration Engine + FireFlow Runtime). The Core consists of a Router for intent parsing and node selection, DBOS for durable exactly-once execution, Audit Logging for full traceability, Type-Safe Ports for structured data flow, and Private Inference for secure on-premise model execution. Every operation is logged and auditable by design.

The Core draws from three categories of data sources:

Public Data / WWW (left): Accessed via MCP or HTTP, these include DeFi primitives (lending pools, DEXs, derivatives, staking, bridges, oracles), e-commerce marketplaces (eBay, Amazon, Shopify, Temu, AliExpress, Etsy), and analytics providers — both on-chain (Nansen, DefiLlama) and traditional (Google, Yahoo Finance) as well as social intelligence feeds (LunarCrush, Kaito).

Private Data (right, encrypted): User-specific data processed within the secure enclave — including messaging history, wallet activity, exchange trading data, banking app data, platform usage patterns, health and fitness data, and email/calendar context. This data never leaves the institution's security boundary.

User Sourced Data (far right): Explicitly provided context such as social media feeds, weather data, financial models, custom knowledge bases, research papers, uploaded documents, and proprietary datasets.

Below the Core, the **TetraChain — Agentic Settlement Layer** provides verifiable settlement with fraud proofs and sub-second finality, while the Connectivity & Interoperability layer enables x402 payments and A2A, A2H, H2A transaction routing.

The processed intent flows upward to **Output** — which can take the form of composed structured products, programmable payments, e-commerce transactions, government service actions, or any other orchestrated result.

Example: A user says: *"I want to buy running shoes under $150 with good reviews, and pay with my USDC balance."* The Router parses this as a multi-step intent. It queries marketplace MCPs (Amazon, Shopify) from Public Data for inventory and pricing, pulls the user's size preferences and past purchases from Private Data, cross-references YouTube review sentiment from User Sourced Data, presents curated options, and upon user confirmation, settles the transaction through TetraChain — all within a single deterministic, auditable flow.

Appendix III: DO, UIO, BPO In Detail

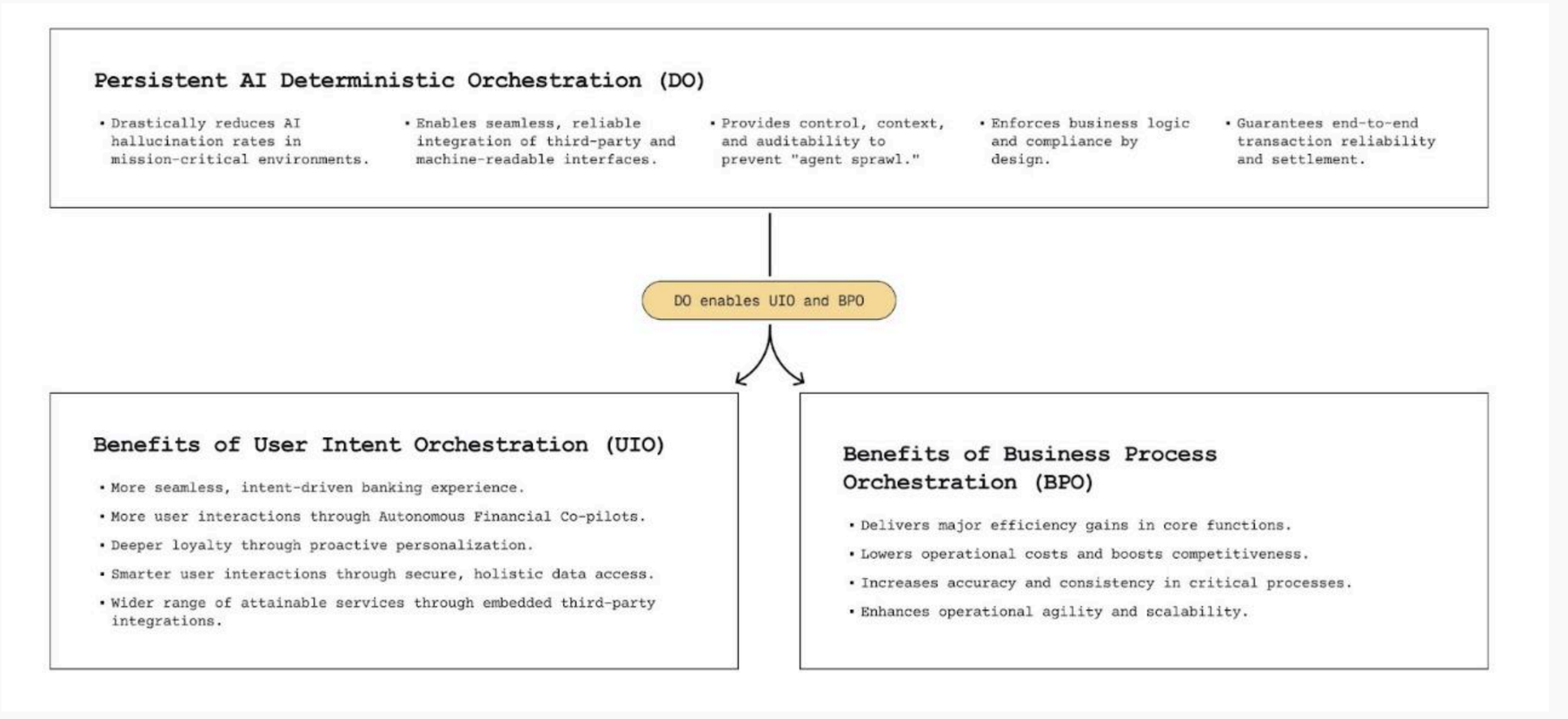

Fig14: Deterministic Orchestration

PersistentAI Deterministic Orchestration (DO) in Detail

Deterministic orchestration is the critical enabling technology for both UIO and BPO:

1. **Drastically reduces AI hallucination rates in mission-critical environments.** Unlike probabilistic AI, deterministic orchestration enforces strict, rule-based workflows and formal verification, ensuring every operation executes exactly as programmed. This eliminates the unpredictable and unacceptable risks posed by generative AI “guessing” in contexts where errors equate to financial loss or systemic failure.

2. **Enables seamless, reliable integration of third-party and machine-readable interfaces.** A deterministic orchestration layer acts as a universal adapter, reliably connecting diverse AI agents, core systems, DeFi protocols, government databases, and data feeds.

3. **Provides control, context, and auditability to prevent “agent sprawl.”** As institutions deploy hundreds of specialized AI agents, deterministic orchestration is the essential central nervous system. It maintains a single source of truth and context, logs every action in an immutable audit trail, and ensures all agents operate cohesively according to institutional policies.

4. **Enforces business logic and compliance by design.** Critical rules are hard-coded into the deterministic workflow itself. This ensures that every process automatically adheres to regulatory and internal policy requirements at every step, drastically reducing compliance risk.

5. **Guarantees end-to-end transaction reliability and settlement.** Deterministic orchestration ensures that a chain of actions will either complete fully and correctly or fail safely with a clear rollback. The Agentic Settlement Layer (ASL) powered by TetraChain provides the certainty required for trustworthy automation.

Benefits of User Intent Orchestration (UIO) in Detail

1. **More seamless, intent-driven experience.** UIO moves beyond reactive, menu-driven apps to a proactive, contextual interface. The system interprets a user's broader goal and automatically orchestrates the necessary services, dramatically reducing friction and decision fatigue.

2. **More user interactions through Autonomous Co-pilots.** Users delegate routine and complex tasks to a trusted AI agent, creating a constant, low-effort engagement layer.

3. **Deeper loyalty through proactive personalization.** UIO enables hyper-personalization that anticipates needs, fostering emotional loyalty beyond mere utility.

4. **Smarter interactions through secure, holistic data access.** UIO platforms, with user consent, can securely aggregate and analyze data from across a user's ecosystem, providing radically more insightful advice.

5. **Wider range of attainable services through embedded third-party integrations.** By acting as a secure orchestrator, UIO seamlessly connects users to best-in-class third-party services directly within the institutional interface.

Benefits of Business Process Orchestration (BPO) in Detail

BPO involves back-office AI agents that automate internal workflows, acting as digital employees to streamline operations:

1. **Delivers major efficiency gains in core functions:** Automating repetitive tasks in compliance, risk management, treasury operations, and administrative processing leads to significant improvements in processing speed and resource allocation.

2. **Reduces operational costs and boosts competitiveness:** Implementing an agentic stack for back-office automation is critical for cost control. Real-world implementations show cost reductions of over 30% while handling increased volume.

3. **Enables the 'agentic institution' model for strategic focus:** This paradigm allows human employees to orchestrate teams of AI agents, freeing them to focus on complex problem-solving and high-value tasks.

4. **Increases accuracy and consistency in critical processes:** AI agents drastically reduce human error, improving data accuracy in document processing from 92% to 99.3% while cutting handling time by 78%.

5. **Enhances operational agility and scalability:** Adaptive process orchestration allows institutions to dynamically adjust workflows, identify bottlenecks, and continuously improve processes.