Ebook

How AI Agents Actually Work

A Practical Guide for Treasury & Finance

Inside TAI, Kyriba's Trusted Agentic AI:
LLMs, Tools, MCP, Governance, & Evaluation

Kyriba

Contents

Introduction: Demystifying AI Agents for Treasury and Finance Leaders

By Felix Grevy, SVP Platform, Data & AI, Kyriba

At Kyriba, we deliver innovative, secure solutions to ensure customers thrive. With a proven record of practical, effective AI-powered capabilities, we've consistently applied cutting-edge technology to address the evolving needs of our customers. Recognizing the transformative potential of Large Language Models (LLMs), we introduced our agentic AI, TAI. A common question we hear is: how does it actually work?

To help demystify AI agents for treasury and finance, we've developed this guide to explain how they function and why understanding their capabilities—and limitations—matters for your organization.

We walk through what an agent can do and how the right approach maintains the trust, compliance, and control that treasury and finance operations demand.

What makes AI agents different: Unlike chatbots that simply respond with pre-trained knowledge, agents can reason through complex problems, call tools to gather real-time data, and recommend or perform actions.

Why treasury operations are perfect for AI agents: Treasury operations hold the richest, most fragmented financial data in the enterprise, and treasury decisions are time-sensitive, data-intensive, and governed by clear policies. Whether managing global liquidity, analyzing FX exposure, or optimizing cash positions, agents excel at aggregating multi-system data instantly and proposing actions with proper approvals and audit trails, translating to basis-point yield improvements and hours returned to treasury teams.

The Kyriba approach: Our agentic AI, TAI, is embedded within our platform, respecting role-based permissions, leaving complete audit trails, and ensuring every action follows the security and compliance frameworks you've already established. It's designed as a policy-aware teammate that explains its reasoning and proves what it did.

What TAI delivers from day one: TAI allows treasurers to improve forecast accuracy with earlier visibility into pattern shifts; optimize cash positioning to capture yield on balances that would otherwise sit idle; and accelerate exception resolution—cutting investigation time from hours to minutes.

TAI's Trust Foundation

- Privately hosted on Kyriba's secure infrastructure
- No customer data used for training
- Role-based access control enforced
- Complete audit trails for every action

Permissions, Privacy, and Control (Non-negotiables)

Trust, security, and control are at the core of Kyriba's AI agent. In finance, where decisions are high-stakes and data is sensitive, these principles are non-negotiable:

Data privacy: No customer data is used in training public models. Ever.

Role-based access: The agent can only see and do what you are authorized to see and do in Kyriba.

API scopes & least privilege: Each tool is strictly scoped (e.g., read payments vs. create payments). Sensitive actions always require explicit approval.

Data residency & logging: All calls are fully auditable. The agent passes only structured, relevant snippets to the LLM (not full datasets). Reasoning steps are transparently traced and visible to the user in the "Thinking Steps," including which tools were called. Machine-generated audit trails mapped to policies and logs of segregation-of-duties significantly shorten audit preparation time and close cycles.

Human-in-the-loop: The agent recommends; you decide. Approvals can be enforced at any action level (e.g., payments, transfers, FX), ensuring full control.

Reversibility: All agent actions are reviewable and reversible within policy, ensuring an additional safety layer for treasury operations.

How AI Agents Work for Treasury

An agent interprets the user's request and then engages in a chain of thought, a reasoning process where it breaks down the problem into steps. Based on this reasoning, the agent identifies which tools or data sources might help, calls them, observes the results, and adapts its plan if needed. This loop continues until the agent can provide a complete, reliable answer to the user.

Large Language Model (LLM)

The key component of an agent is the **Large Language Model (LLM),** the "brain" that performs reasoning and planning.

LLMs were popularized by OpenAI with the release of ChatGPT in November 2022, but the underlying breakthrough came earlier, with Google researchers' 2017 paper "Attention Is All You Need." This work introduced the **transformer architecture**, which made it possible to train models much more effectively.

The idea is revolutionary: instead of hard-coding rules, the model is trained on vast amounts of text, such as content from millions of web pages and books, so that it learns statistical patterns of language and can generate coherent, well-formed sentences. Modern LLMs achieve this scale through **billions of parameters**, the adjustable weights inside the neural network. Smaller models may have around **7–8 billion parameters**, while the largest frontier models can exceed **100 billion parameters**. More parameters generally allow the model to capture more complex patterns, though they also require vastly greater computing power.

Frontier LLMs are powerful; Kyriba's variant runs in a private, governed environment so no customer data trains any public models.

In treasury, where operations are rule-bound and high-stakes, agents don't just "chat"—they execute policy-aware workflows that close the last-mile gap from answers to actions.

To improve usefulness and safety, an additional process called **Reinforcement Learning with Human Feedback (RLHF)** is applied. In this step, humans review outputs and guide the model on how to answer questions in ways that are more accurate, relevant, and aligned with user expectations.The initial training of an LLM, compressing vast amounts of knowledge into billions of parameters, requires enormous computational power, typically using large clusters of GPUs or specialized AI accelerators. Additionally, the Reinforcement Learning with Human Feedback (RLHF) process adds weeks or months of fine-tuning, as humans guide the model to produce more useful and aligned answers.

Overall, it can take **several months to produce a single model**, which explains why new generations (e.g., from GPT-3.5 to GPT-4) are released on a **multi-month cycle** rather than continuously.

Tokens

LLMs process language using **tokens**, small units of text that can be as short as a single character or as long as a word, but on average about four characters in English. Instead of reading sentences directly, the model breaks everything down into these tokens.

The model's core task is to predict the next token in a sequence, based on all the tokens it has already seen. It does this using the transformer architecture: layers of attention mechanisms and neural networks that apply advanced mathematics and statistical patterns learned during training.

This process, called **inference**, is what happens each time you interact with the model. Unlike the months-long training process, inference is much faster and requires far less computation, which makes real-time use possible.

Prompt

Every interaction with an LLM begins with a **prompt**, the input text that frames the task. A prompt can be as simple as a direct question ("Explain corporate treasury"), which the model can answer from its trained knowledge.

However, there is a limit to how much text an LLM can consider at once. This limit is called the **context window**, and it is measured in tokens. Depending on the model, context windows today range from around 32,000 tokens (roughly 20–25 pages of text) to more than 200,000 tokens (an entire book). This ebook, for example, has roughly 2800 tokens. Prompts must not exceed the context window size.

With TAI, the agent sends the LLM only the structured, relevant snippets needed to answer each query—never full datasets.

Smart context window management lets Kyriba's agentic AI, TAI, compare multi-bank positions, forecasts, and payment queues in a single conversation—tasks that are hugely time-consuming when done ad hoc.

In the initial version of ChatGPT, the main innovation was a chat interface that allowed users to have natural conversations with the LLM. The model responded using patterns and knowledge it had learned during training, but that knowledge was static and limited to the information available up to its training cut-off date.

This limitation meant that ChatGPT could answer general questions, explain concepts, or generate text in many styles, but it could not access the latest news, facts, or real-time data, since it wasn't connected to the internet or external systems.

Tools

The next breakthrough came with the introduction of **tools**. Instead of relying only on its static training, the LLM could now decide to call an external tool, for example, a search engine. Based on the user's request, the model could trigger a search query, process the results, combine them with the ongoing conversation, and then produce a final answer.

For instance, if you ask *"What's the weather tomorrow?"*, the LLM must call a tool to fetch real-time weather data, since that information isn't part of its training. Once the result is returned, the LLM integrates it into the conversation and provides the final answer.

This capability removed one of the biggest limitations of early LLMs: it allowed them to access up-to-date information and respond accurately to questions about events or facts that happened after the training cut-off date.

Tools are the way for an LLM to observe a system and perform actions.

They extend the model beyond language prediction: instead of only generating text, the LLM can decide to call a tool to look up data, run a calculation, query a database, or trigger an external process.

In treasury, tools must follow least-privilege design: each tool is scoped precisely—"read balances" does not grant "create payments." Sensitive tools are segregated and always approval-gated.

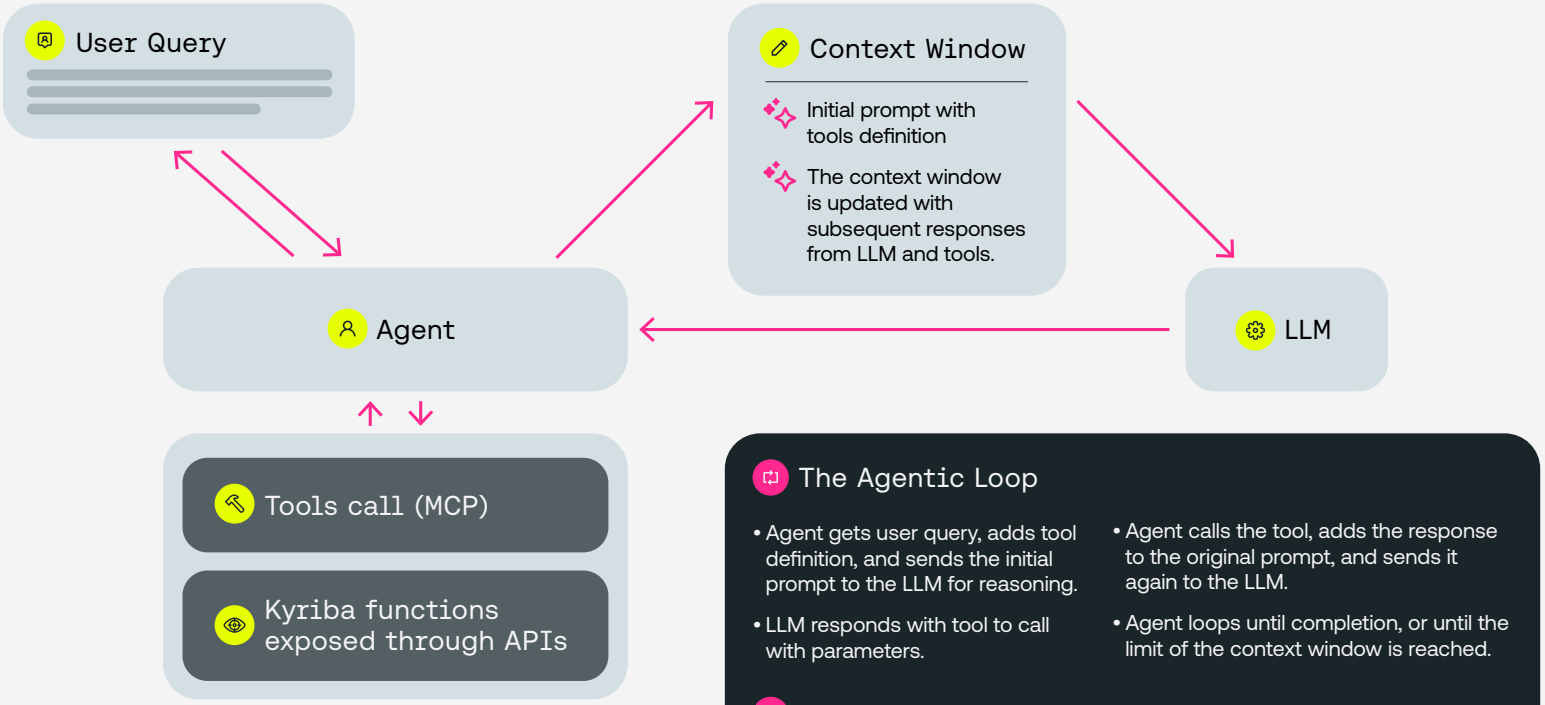
Reasoning Loop

A typical agent works through a **reasoning loop:**
Plan → Act (call tools) → Observe → Refine → Answer
The agent repeats this loop until it has gathered enough information to respond—or, in some cases, to request approval before taking an action.

In treasury operations, this loop applies domain-specific intelligence. For example:

- **Plan:** Identify trapped cash and yield opportunities.
- **Act:** Call tools to retrieve balances and forecasts.
- **Observe:** Screen urgent payments for approval.
- **Refine:** Respect minimum balances and cutoff times.
- **Answer:** Propose transfers and FX hedges with policy citations.

The effectiveness of this loop also depends on the context window of the LLM, which defines how much of the conversation and tool outputs it can keep in memory at once. A larger context window allows the agent to consider longer conversations, detailed reports, or multiple tool calls in sequence, making its reasoning more accurate and reliable.



The Agentic Loop

- Agent gets user query, adds tool definition, and sends the initial prompt to the LLM for reasoning.
- LLM responds with tool to call with parameters.
- Agent calls the tool, adds the response to the original prompt, and sends it again to the LLM.
- Agent loops until completion, or until the limit of the context window is reached.

Why This is Auditable

- Every tool call is logged with user, scope, timestamp, and payload hash.

Kyriba

Kyriba's Agentic AI Architecture (At a Glance)

To deliver the best agentic experience for finance professionals, where safety and trust are paramount, Kyriba has built its agentic AI solution following industry best practices and the most advanced AI concepts.

Agentic orchestration

The reasoning loop (Plan → Act → Observe → Refine → Answer) is orchestrated using LangGragh, built on the LangChain ecosystem. These frameworks are purpose-built for robust, explainable, and auditable agent workflows, refined by Kyriba for enterprise-grade scalability and constraints.

Large Language Model (LLM)

For complex queries, Kyriba leverages a frontier LLM with over 100 billion parameters, delivering both accuracy and depth of reasoning. Kyriba also applies an evaluation framework that tests LLM and MCP tool performance across a large set of treasury-related questions, using LLM-as-a-Judge techniques to benchmark and select the most suitable model.

Safety & compliance

The LLM is privately hosted within Kyriba's secured environment (Databricks and AWS). All usage is governed by contractual Data Processing Agreements, ensuring that customer data is never used for training and cannot leak outside Kyriba's environment. Strict controls on data residency, access rights, and auditability ensure the agent always operates within the highest standards of compliance and trust.

Tool calling via MCP

Tool usage is managed through the Model Context Protocol (MCP), an emerging industry standard initiated by Anthropic. Much like USB-C for AI, MCP provides a universal way for LLMs to call APIs. In Kyriba's setup, MCP executes API calls on behalf of the user, always respecting data permissions, scopes, and segregation of duties.

All interactions between Kyriba, the agent, and MCP flow through an **API Gateway**, which provides:

- **Centralized security:** Authenticates, encrypts, and controls access for every call.
- **Data integrity:** Ensures payloads cannot be altered in transit.
- **Traffic management:** Scales safely while preventing misuse or overload.
- **Monitoring & auditability:** Logs every call and makes each one traceable for compliance.

Our continuous investment in API-first strategy to shape Kyriba around secure, scalable, well-governed APIs has laid the foundation for our agent capabilities.

Why This Architecture Unlocks Treasury Value:

Speed:
Intraday liquidity moves before cut-offs.

Breadth:
Aggregates ERP, banks, and payments telemetry.

Consistency:
Executes policies the same way every time.

User → Kyriba API Gateway (auth/encrypt/log) → TAI → MCP → LLM → User

(No lateral escape paths)

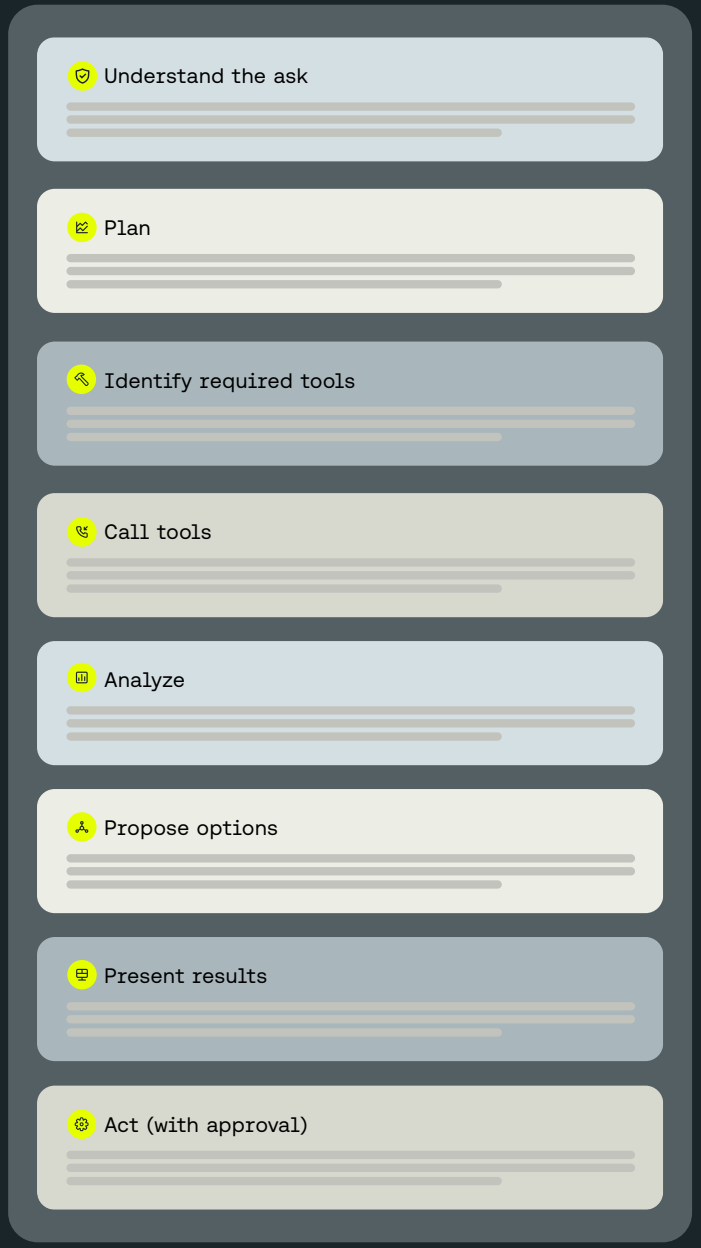
PRIVATE-HOSTED | API-FIRST | MCP-GOVERNED | LEAST-PRIVILEGE | FULL AUDIT

Kyriba

Walkthrough Example

“Where can we safely free up $10M by Friday?”

- **Understand the ask**: The agent parses the request and extracts constraints (amount, deadline, currencies, entities) and applicable policies (investment guidelines, cut-off times).
- **Plan:** Decide what to fetch: list of entities, current bank balances, forecasted inflows/outflows, committed payments, available credit.
- **Identify required tools:** Select the relevant MCP tools.
- **Call tools:**
  - For each entity: getBankBalances(entity=X, asOf=today)
  - For each entity: getForecast(entity=X, horizon=7d)
  - getReport(PendingPayment)
  - getReport(CreditLine)
- **Analyze:** Compute liquidity headroom by bank/account, consider FX effects, and ensure minimum balances and policy limits are respected.
- **Propose options:** For example, *“Move $6.2M from Entity A to Entity B today; convert €3.8M to USD tomorrow; draw $1M on Facility C.”*
- **Present results:** Show numbers, sources, and cut-off risks clearly to the user.
- **Act (with approval):** The agent can execute recommended actions directly in Kyriba, always subject to:
  - **Segregation of duties:** Creator doesn’t mean Approver; dual-control enforced; approvals recorded with reason codes.
  - **Full auditability:** Each recommendation links back to source reports and time stamps.



Limitations to Keep in Mind

Even with advanced orchestration and controls, it's important to recognize that LLMs have inherent limitations:

Context Window

An LLM can only "see" a limited amount of text at once. If a request or dataset is too large, the model may struggle to produce coherent results. While we apply smart context management and frugal AI techniques—optimizing MCP tool responses to minimize unnecessary tokens and focus on what truly matters—it is often best to start a new conversation to reset the context window, especially when switching to a different topic.

Hallucinations

LLMs can sometimes generate plausible-sounding but incorrect information. Because finance actions must be grounded in tool outputs and reports, free-form text can never trigger payment creation or policy changes. To safeguard financial integrity, Kyriba's agent always grounds its answers in tool outputs and reports, with sources visible to the user. Free-form text generation is never used alone to trigger financial actions. However, results should always be double-checked, as hallucination remains an inherent limitation of today's LLMs.

Mathematics

LLMs excel at language but are less reliable with precise calculations, since they rely on next-token prediction rather than true arithmetic. To address this limitation, Kyriba's agent uses a dedicated calculator tool, ensuring that mathematical operations are accurate and auditable.

Harness AI Agents for Treasury and Finance

Agents represent the next step in applied AI for finance, moving beyond simple chatbots into intelligent teammates that can reason, plan, and act—while staying firmly within the guardrails of trust, compliance, and control.

By combining:

Frontier LLMs
(hundreds of billions of parameters, hosted securely),

Agentic orchestration frameworks
(LangGragh, built on the LangChain ecosystem),

Standardized tool connectivity
(MCP, respecting roles and permissions),

Auditability and human approvals,

Kyriba delivers an enterprise-grade agent designed for finance professionals who need both speed and safety.

The customer outcome: CFOs, treasurers, and finance teams can confidently leverage AI to answer complex liquidity questions, optimize cash management, and take action, always backed by the same trust and rigor that define Kyriba's platform.

As the AI industry evolves—with larger context windows, more capable models, and new orchestration methods—Kyriba will continue to adopt the latest advancements. Our AI agent for finance will remain state-of-the-art, empowering finance teams with innovation that is trustable, responsible, and future-ready.

Learn more about our agentic AI, TAI, and see how it can bring your treasury to the next level.

Why CFOs Trust TAI:

- Private hosting
- Least-privilege tools
- Segregation of duties for approvals
- Comprehensive audit logs
- No training on customer data

Sources

Foundational Research

- Apple Machine Intelligence Research (2024). The illusion of thinking in Large Language Models.
- Brown, T., et al. (2020). Language models are few-shot learners. NeurIPS. (Introduces GPT-3).
- OpenAI (2022). Introducing ChatGPT.
- OpenAI (2025). Why language models hallucinate.
- Ouyang, L., et al. (2022). Training language models to follow instructions with human feedback. (Introduces RLHF).
- Vaswani, A., et al. (2017). Attention is all you need. NeurIPS.

Agentic Frameworks & Tool Use

- Anthropic (2024). Model Context Protocol (MCP).
- LangChain Documentation – https://python.langchain.com
- LangGraph Documentation – https://www.langchain.com/langgraph

Risks & limitations

- Borji, A. (2023). A categorical archive of ChatGPT failures.
- Ji, Z., et al. (2023). Survey of hallucination in natural language generation.
- OpenAI (2023). GPT-4 technical report.
  (Discusses context window, limitations).

Industry / Enterprise AI Context

- Databricks (2024). Building state-of-the-art enterprise agents
- Gravitee (2025). Google's Agent-to-Agent (A2A) and Anthropic's Model Context Protocol (MCP).
- IDC (2025). Agentic AI in Treasury: Navigating Trust and Realizing Potential.
- Kyriba (2025). From DX to AX: Why the future of AI depends on agent experience and open standards like MCP.
- Kyriba (2025). Mews: A partnership delivering innovation, efficiency, and trust in AI-enabled treasury.
- Kyriba (2024). API Integration with Large Language Model.
- McKinsey & Company (2023). The economic potential of generative AI.

Kyriba

Cooke Aquaculture, a global aquaculture company operating in 15 countries, achieved 100% cash visibility and reduced manual treasury workload by 83% through Kyriba's Liquidity Performance Platform.

Kyriba

Select Brands Using Kyriba

爱回收 · Baxter · Cenveo · GRAFF · HCSC Health Care Service Corporation · KOCH

ŞİŞECAM · Spotify · SOCIETE GENERALE · SPECIALIZED · TRANE TECHNOLOGIES · Cooke

About Kyriba Corp.

Kyriba is the global leader in liquidity performance, empowering CFOs, Treasurers and IT leaders to connect, protect, forecast and optimize their liquidity amid economic complexity. As a secure and scalable SaaS solution trusted by more than 3,000 customers, Kyriba delivers trusted intelligence and financial automation through innovative technologies—including its agentic AI, TAI—bringing precision, efficiency, and insight to financial operations. With an expansive ecosystem of banking, technology and consulting partners, Kyriba's platform powers more than 3 billion bank transactions across 9,900+ banks annually with $15T annual payments orchestrated with enterprise-grade controls, helping companies gain enterprise-wide visibility, ensure financial stability, and outperform their business strategy.
For more information, visit www.kyriba.com.