cognizant®

Intelligence at scale:

How Agentic AI Is reinventing Investment Research

Abstract

AI is no longer a pilot initiative - it is fundamentally transforming how organizations operate and perform. In industries like finance, healthcare, logistics and manufacturing, AI-driven processes such as document analysis, predictive maintenance, supply chain optimization and risk assessment have moved beyond experimental use. These capabilities now serve as core drivers of competitive advantage.

That same inflection point has now reached research function as well. Agentic AI marks a step change in how insights are produced and consumed. Investment research, long defined by manual rigor and linear processes, now stands at the threshold of reinvention. This paper introduces our newly created Agentic AI system for research, built on LLMs and RAG architectures. Early deployments show accuracy rates of 90-92%, along with measurable efficiency gains reshaping analyst workflows.

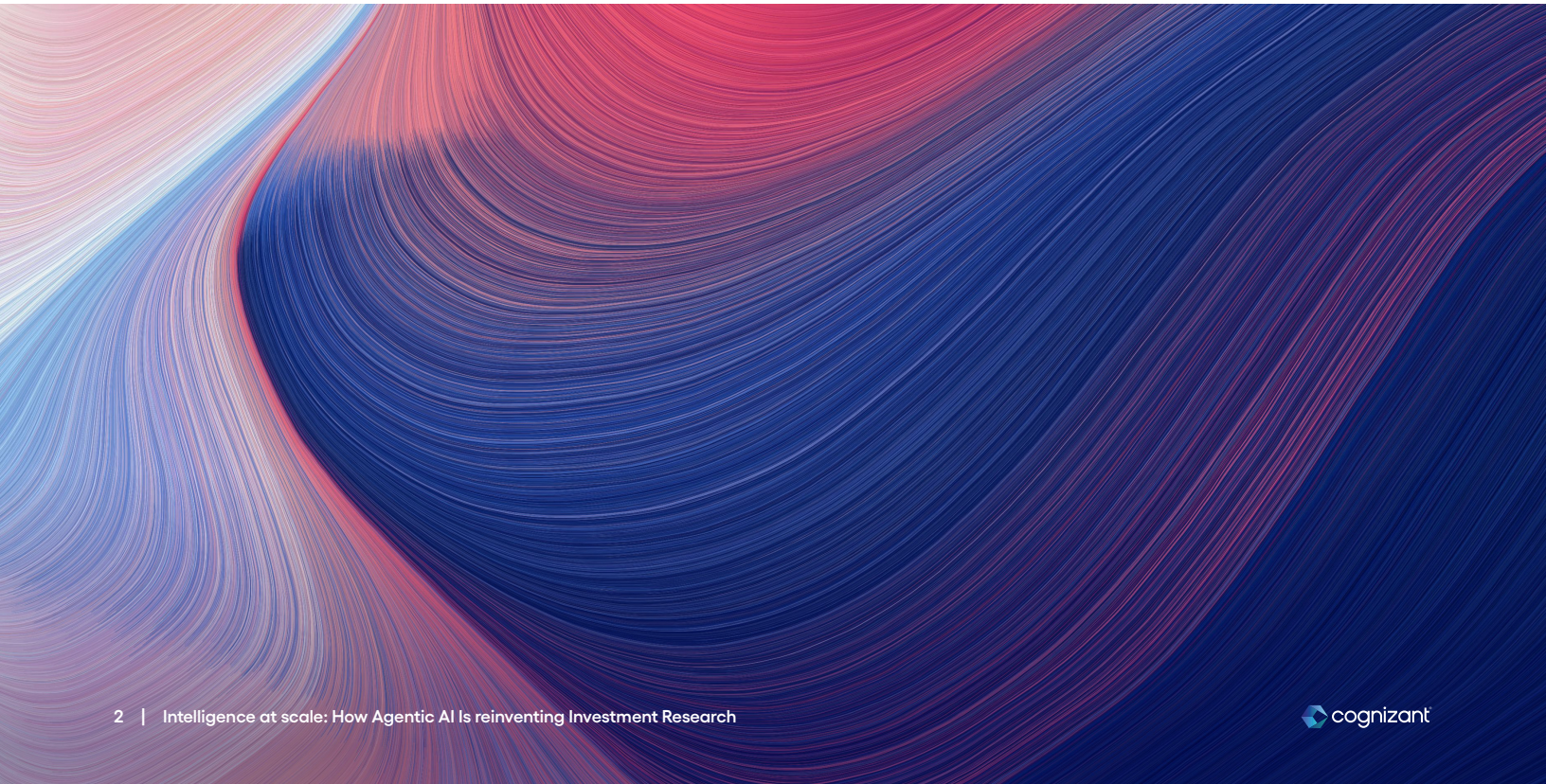
Looking ahead, future versions will expand into multilingual, multimodal and explainable systems with hierarchical reasoning - bringing broader data coverage, deeper trust and transparent recommendations.

We invite research organizations to join us in shaping this journey, testing the platform and co-defining the next chapter of future research innovation.

Introduction: The Analyst's dilemma

Research analysts today face unprecedented pressures. The volume of information they must digest has exploded, deadlines have compressed to near real time, and expectations for accuracy, transparency, and compliance leave little margin for error. Bloomberg captures this reality as a "fire hose of information" - a torrent of data that no human can realistically absorb. The firm now processes more than 400 billion data points daily, surging to nearly half a trillion during periods of volatility. As Bloomberg's CTO Shawn Edwards observes, the real challenge is no longer accessing data but distilling what truly matters - at the precise moment when clients demand actionable insight (Edwards, 2025)[1].

The problem is systemic. Analysts are expected to read hundreds of pages, reconcile conflicting inputs, and deliver error-free conclusions under unforgiving time constraints. Manual approaches are reaching a breaking point: they create bottlenecks, exhaust scarce talent and risk undermining the very precision that defines research credibility.



Industry challenge: Why the old model cannot scale

Investment research has historically relied on painstaking manual workflows. Analysts trawl through filings, earnings transcripts, market reports and client datasets - often in inconsistent formats - before stitching together insights manually. Quality depends on both endurance and expertise, but the process is inherently linear, fragile and slow. Tribe AI's case study highlights how outside-in due diligence at a global consulting firm was hindered by manual research workflows, with analysts spending days or weeks piecing together fragmented insights from public documents (Tribe, 2025).

Three pressures are converging to make this model unsustainable:

- **Data overload** – The sheer quantity and heterogeneity of available information grows exponentially each year, far outpacing human processing capacity.
- **Time compression** – Decision cycles have shortened dramatically. What once could take weeks is now expected within hours - without compromise in quality.
- **Rising standards** – Clients demand outputs that are not only precise but also transparent, traceable and compliant, leaving little room for error or opacity.

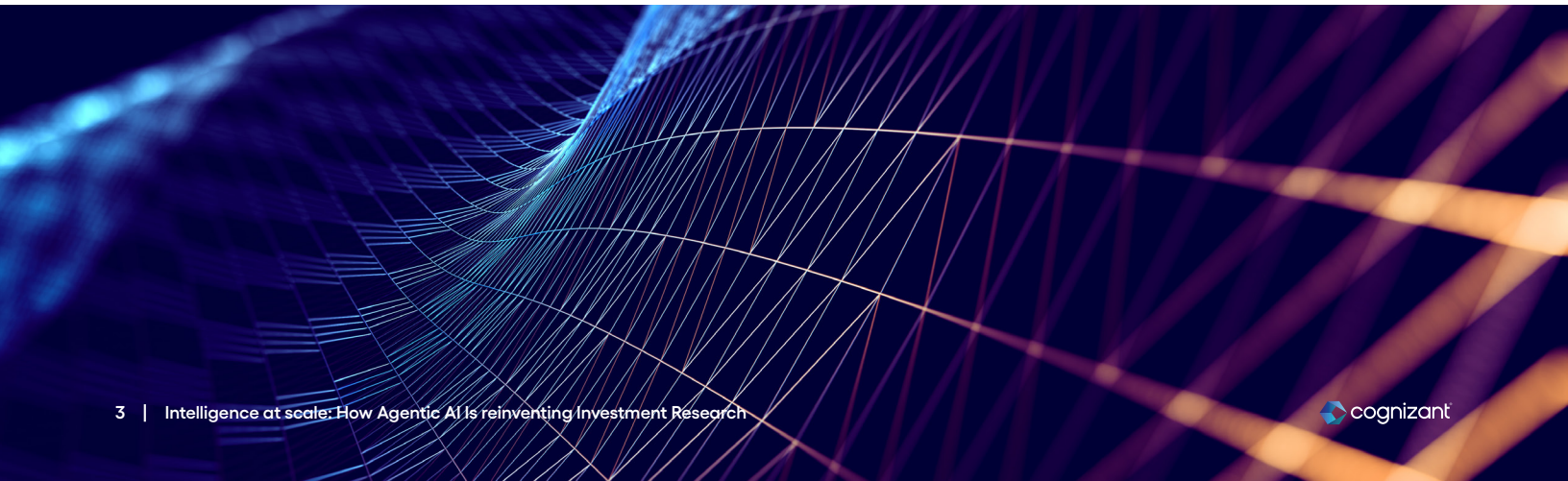
The result is a widening gap: traditional methods cannot keep pace with the scale, speed and scrutiny now required. Without intervention, research risks becoming a bottleneck rather than an enabler of market leadership.

The Investment Research Assistant: Redefining the research workflow

To address these rising pressures, we developed the Investment Research Assistant - an agentic AI platform designed not to replace analysts but to amplify their expertise. Built on a foundation of large language models (LLMs), retrieval-augmented generation (RAG) and LangChain-driven orchestration, the solution reimagines the research workflow as a coordinated system of specialized AI agents under human oversight.

Rather than following the linear, manual sequence that defines traditional research, the platform operates as an intelligent, distributed architecture. Multiple agents simultaneously extract, reconcile and validate information from diverse sources - financial filings, transcripts, analyst notes and structured datasets. Outputs are then synthesized into research-ready formats: structured discrepancy reports, executive summaries, visual dashboards and even draft slides.

At every stage, the analyst remains central. Human-in-the-loop checkpoints ensure that professional judgment governs the process: reviewing flagged discrepancies, validating complex interpretations and refining the final narrative. This balance of machine speed and human rigor delivers an unprecedented combination of efficiency, accuracy and accountability.



Understanding RAG

Retrieval-Augmented Generation (RAG) is an AI framework that enhances language models by integrating external knowledge retrieval. Instead of relying solely on pre-trained data, RAG dynamically pulls relevant information from sources like documents or databases during inference. This improves accuracy, reduces hallucinations and enables up-to-date responses. RAG combines two components: a retriever that finds relevant content and a generator that uses it to craft coherent, informed answers. It's widely used in enterprise search, customer support and research applications, offering scalable, context-aware solutions. RAG bridges the gap between static knowledge and real-time, grounded AI responses.

Understanding Agentic AI

Agentic AI refers to artificial intelligence systems that exhibit autonomous decision-making and goal-directed behavior. Unlike traditional AI, which follows predefined instructions, agentic AI can set its own objectives, adapt to changing environments and take initiative to achieve outcomes. These systems often incorporate reasoning, planning, and learning capabilities, enabling them to act independently and interact dynamically with humans and other systems. Agentic AI is central to developing intelligent, proactive machines that can collaborate, negotiate and operate with minimal human oversight - raising critical questions about safety, ethics and control in complex AI ecosystems.

Technology stack and models

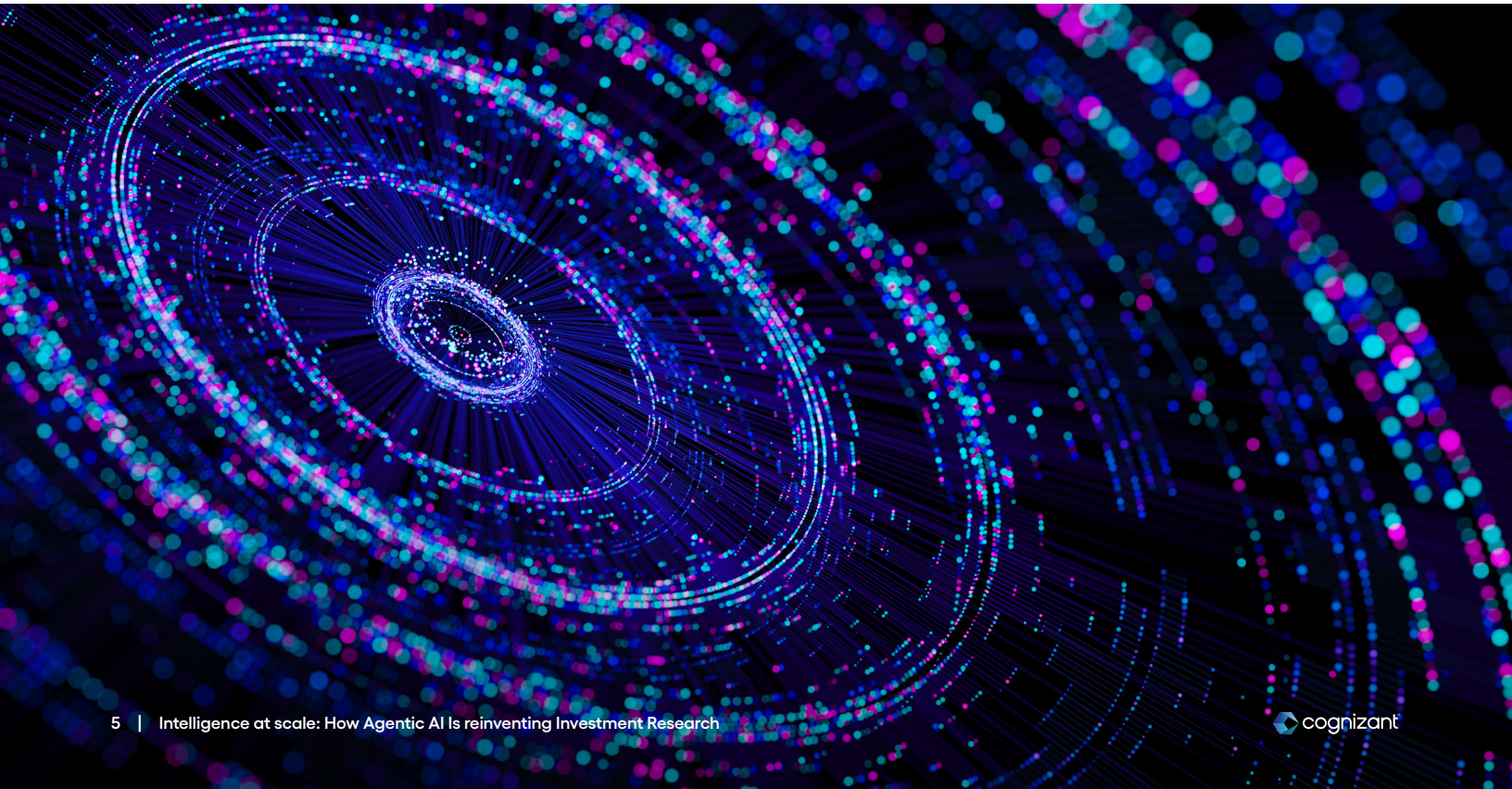
The Investment Research Assistant is built on a cloud-native, enterprise-grade architecture designed for scale, transparency and sustainability. At its core, the solution employs LangChain as the agentic orchestration framework, enabling specialized AI agents to collaborate seamlessly. The architecture is fully composable, so new features and models can be added in a plug-and-play manner as research needs evolve.

The stack combines LLaMA 3.3 (7B) for language tasks, FastAPI for high-performance service delivery, Qdrant as the vector database for retrieval and similarity search and Azure Blob Storage for secure data persistence.

A deliberate decision was made to adopt open models such as LLaMA rather than depend solely on closed, proprietary APIs. Larger models - including LLaMA 70B and GPT-4 class systems - were tested and found to deliver only 1-2% improvements in accuracy while consuming significantly more energy and compute resources. By contrast, the LLaMA 3.3 (7B) model achieved near-parity accuracy with ~60% less energy use, ~50% lower cost and faster response times. This ensures not only efficiency and sustainability, but also greater transparency, control and adaptability - critical for compliance-driven industries such as financial services.

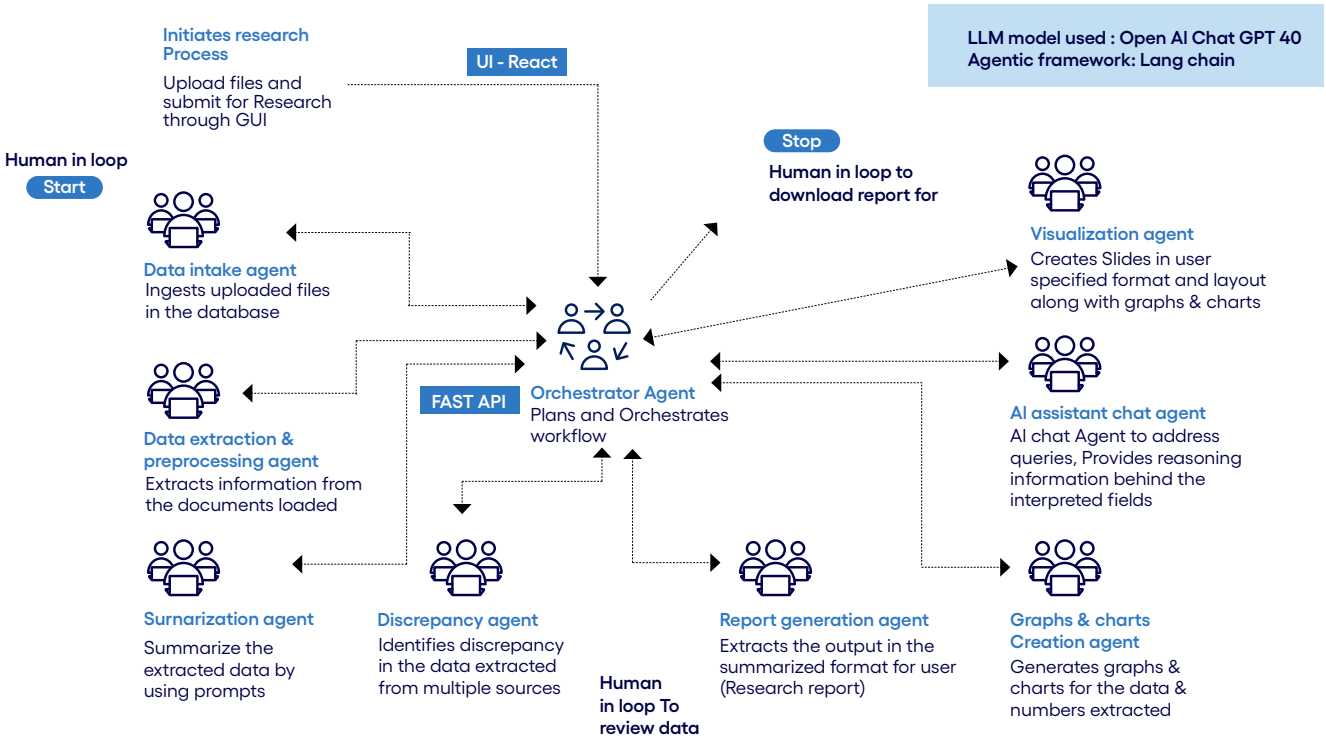
The philosophy guiding model selection is simple: choose the model that delivers the required accuracy with the lowest energy, lowest cost and smallest carbon footprint. By embedding this principle into the platform's DNA, we align technical choices with enterprise priorities around compliance, sustainability and cost discipline.

This architecture is not an abstract design - it is the engine behind the gains already demonstrated. By orchestrating specialized agents, enabling real-time retrieval and optimizing for efficiency, the platform enables not only faster research but a reimagined operating model: analysts spend less time wrangling documents and more time applying judgment, context and creativity - the qualities that define competitive advantage.



Architecture – Investment solution

As depicted in Figure 1, at the heart of the Investment Research Assistant lies an agentic architecture: a coordinated system of specialized AI agents, each responsible for a distinct task in the research workflow. Unlike monolithic models, this architecture mirrors how research teams operate - dividing work across roles while maintaining constant coordination and oversight.



The system comprises eight specialized agents, orchestrated by LangChain and designed for composability:

- **Data intake agent** – ingests unstructured and structured sources such as filings, transcripts and analyst notes.
- **Extraction agent** – identifies key entities, metrics and signals from raw data.
- **Summarization agent** – condenses documents into concise, accurate research-ready outputs.
- **Discrepancy agent** – detects conflicts or inconsistencies across multiple inputs.
- **Report generation agent** – produces structured research reports aligned to analyst formats.
- **Visualization agent** – translates insights into charts and dashboards.
- **Slide creation agent** – drafts presentation-ready materials to accelerate client-facing deliverables.
- **Conversational agent** – enables interactive Q&A with analysts, guiding interpretation and clarifications.

These agents operate in parallel, which dramatically compresses cycle times. Just as importantly, the architecture is built around human-in-the-loop checkpoints. Analysts validate flagged discrepancies, review draft outputs and bring judgment and context that no model can replicate. This combination ensures that outputs are fast, consistent and reliable - yet always anchored in human expertise and accountability.

The impact of this design is profound. By distributing work across intelligent agents while retaining human governance, the platform replaces the bottlenecks of linear research with a networked, collaborative workflow. Tasks that once consumed days are resolved in hours. Analysts shift their focus from document handling to higher-order thinking, enhancing both productivity and the quality of insight delivered to client

Business impact

The Investment Research Assistant is not an incremental improvement; it represents a step-change in research productivity and quality. By re-engineering workflows around agentic orchestration and human oversight, the platform has delivered measurable impact at scale.

Research teams can now achieve a 40% faster turnaround for reports, enabling insights to reach decision-makers within hours instead of days. Summarization efficiency has improved by 45%, allowing analysts to cover broader datasets without compromising accuracy. End-to-end research cycles have been reduced by 42%, compressing the time from raw data ingestion to client-ready deliverables.

The implications extend beyond efficiency. Analysts can now devote more time to interpretation, context and strategic judgment - the qualities clients value most - rather than manual document handling. This shift enhances both the quality of client deliverables and the job satisfaction of analysts, reducing burnout and freeing capacity for higher-value work.

For enterprises, the benefits compound. Faster research cycles mean accelerated decision-making, giving clients an edge in competitive markets. Improved accuracy and transparency strengthen trust with regulators and stakeholders, reducing compliance risk. And by leveraging energy-efficient, cost-conscious models, firms realize not only financial savings but also progress toward sustainability commitments.

In short, the Investment Research Assistant is more than a productivity tool - it is a strategic enabler. It equips firms to handle exponential data growth, meet unforgiving deadlines and uphold rising expectations for quality and compliance, all while enhancing the role of the analyst and reinforcing trust in the research function.

Organizations seeking to accelerate their research transformation can connect with us to access and deploy the Investment Research Assistant.

Future work: Advancing to the next version

While the current version of Investment Research Assistant is already set to deliver transformative gains, its evolution is just beginning. The next version will expand capabilities, strengthen trust and push the platform closer to autonomy.

Functionally, the system will broaden its reach with multilingual support, enabling analysts worldwide to work seamlessly in their native languages. It will grow into a multimodal, explainable assistant with hierarchical reasoning to connect and synthesize diverse data streams. These include official filings and audited statements for factual baselines; regulatory and legal databases for compliance risks; financial market feeds for pricing, ratings and volatility signals; media and news archives to surface controversies and social/professional platforms like LinkedIn, Glassdoor and Twitter/X to capture governance sentiment. Audio and video - such as earnings calls, interviews, or YouTube content - will provide management tone and stress cues, while alternative datasets like shipping manifests, satellite imagery, or app downloads will validate operational claims. Each source will be processed independently, then aggregated through a super reasoning model to deliver comprehensive, transparent recommendations.

On the technology side, the platform will address current AI limits. Reinforcement guardrails will reduce hallucinations by rewarding factual accuracy. Hierarchical reasoning will separate planning from execution, improving depth and reliability. Explainability by design will ensure every output is auditable and traceable, while integrated sustainability metrics will monitor energy use and carbon footprint. Ethical and governance principles, rooted in Constitutional AI, will guide all recommendations.

Together, these enhancements will move the Research Assistant from a powerful productivity tool to a trusted, sustainable and globally inclusive research partner.

Our future work is not just an upgrade - it lays the foundation for a new operating model where human expertise and intelligent agents combine to deliver faster, deeper and more accountable insights.

We invite research agencies and companies to join us in shaping this journey - partnering to determine the functional viability of our vision and together, to define the future of research.

References:

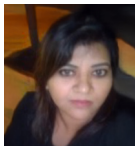
- https://www.bloomberg.com/company/press/bloomberg-cto-shawn-edwards-constellation-research-ai150/
- **BCG (2025): How insurers can supercharge their strategy with AI** How insurers can supercharge strategy with artificial intelligence | BCG
- **Forrester, Predictions 2025: Insurance**
  Predictions 2025: Can AI Deliver On Its Promises For Insurance?

Authors

Dr. Venkatesh Upadrista

Global Head of Transformation, BFSI - IOA
Drvenkatesh.upadrista@cognizant.com

Dr. Venkatesh Upadrista leads global transformation for BFSI IOA vertical at Cognizant. In this role, he is responsible for driving AI transformation across the unit, while ensuring customer success and strengthening delivery of modern business operations for the Financial Services and Insurance sector.



Sanghamitra Das

Delivery Leader, BFSI-IOA
Sanghamitra.Das@cognizant.com

Sanghamitra Das brings over 20 years of experience in leading financial spreading, investment research and compliance operations for top-tier banks in the BFSI. She is recognized for her deep domain expertise, outstanding client engagement capabilities and a consistent track record of delivering sustainable value.

Manikantan Nair

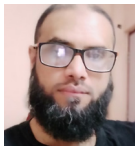
Global Head of IB & WM Delivery, BFSI - IOA
Manikantan.Nair@cognizant.com

Manikantan Nair leads IB & WM Delivery for BFSI IOA vertical at Cognizant. In this role, he is responsible for delivery of modern business operations and AI led solutions for the Financial Services sector.

Yayati Tyagi

Process Leader, BFSI-IOA
Yayati.Tyagi@cognizant.com

Yayati Tyagi leads research delivery for Cognizant's global investment research clients. He is responsible for driving AI-led transformations in his domain to help clients optimize research quality, operational efficiency and turnaround time.



Mirza Mehdi Ali

Senior Manager – Process Excellence, BFSI-IOA
MirzaMehdi.Ali@cognizant.com

Mirza Mehdi Ali is responsible for driving transformation for clients in the capital markets sector.