# Agents on a Leash:
## Deterministic Agentic AI for Financial Services

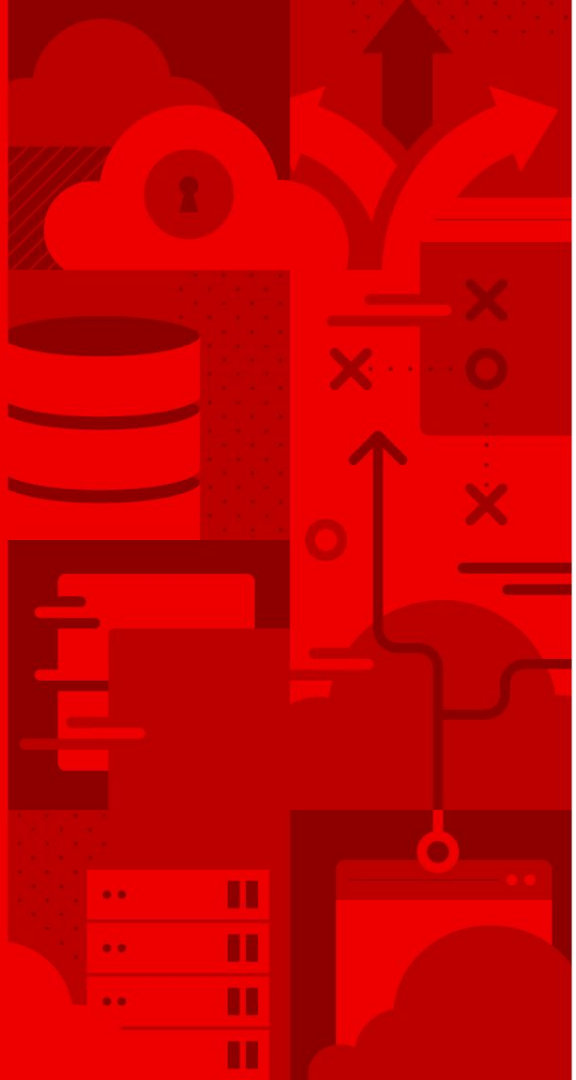


**Harish Malavade**
VP, Software Engineering
Fidelity Investments
Harish.Malavade@fmr.com



**Aric Rosenbaum**
Chief Technologist
Financial Services, Red Hat
aric@redhat.com

Red Hat



Fidelity and Red Hat are two separate companies. Unless otherwise noted, the opinions provided are those of the speaker, not necessarily those of Fidelity Investments or its affiliates.

This is not an endorsement or recommendation of any vendor, product, or service.

Red Hat

# Harish Malavade

*Fidelity Investments, VP Software Engineering*



e: Harish.Malavade@fmr.com

## Background

- Works as Engineering leader for centralized automation platform, where he helps business units across fidelity to orchestrate and increase STP by automating end to end processes and also a Lead maintainer for Fluxnova
- Prior to this role he has worked in Trade processing (FX & Equity), Fund Accounting, Share holder Reporting
- Harish has developed and also architected multiple full stack critical applications and integrated them with banks, brokers, advisors etc

## Area of Expertise

Financial services, Java, Angular, Database, Cloud(AWS), Trade Processing, Record Keeping, Engineering Leadership, Architecture

## Selected Experience

Middle office Trade Processing

Mainframe to Cloud migration

Leading Fluxnova from Inception to adoption



Red Hat

# Aric Rosenbaum

*Red Hat, Chief Technologist, Global FSI*



e: aric@redhat.com

m: +1-973-610-4671

## Background

- Aric serves as the Chief Technologist in Red Hat's Global FSI team, where he helps clients meet their strategic priorities through the use of open source technology.
- Prior to joining Red Hat, he led large, digital transformation projects at Goldman Sachs' Investment Management Division and was co-founder / CTO of several FinTechs in equity and FX trading.
- In his 20+ years of experience, Aric has designed and built multiple front, middle and back office systems and deployed his first app to the cloud in 2008.

## Area of Expertise

Financial services, capital markets, FinTech, microservices, continuous integration / delivery, agile, cloud computing, hybrid cloud

## Selected Experience

At Goldman Sachs, Aric co-led a global team of 60+ engineers building the next generation client lifecycle management system.

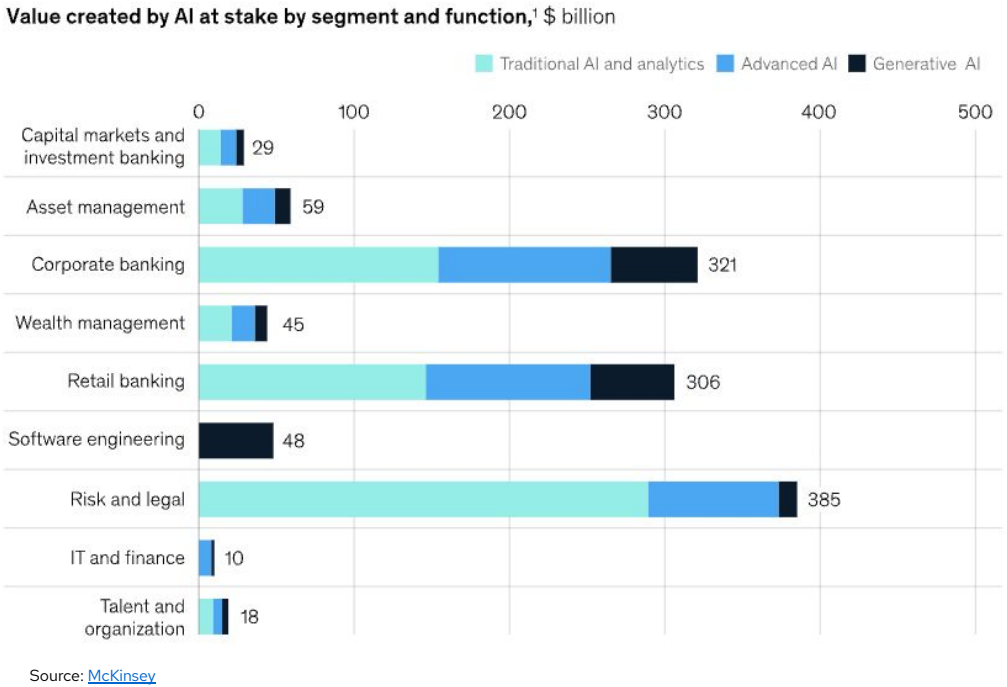
Aric served on the Governing Board of the Fintech Open Source Foundation (FINOS) from 2022-2024.



Red Hat

# Gen AI Will Deliver Significant Value to Banks

McKinsey: $200 - $340 billion (3-5% of industry revenues, 9-15% of operating profits)




Source: McKinsey

## Use Cases

- Customer service (ChatBots, Avatars)
- Document summarization / gen
- AML / KYC
- Virtual experts (policies, regs)
- Application development
- Legacy code migration
- Client onboarding



Red Hat

# But 95% of GenAI pilots are stalled

MIT: “Most GenAI systems do not retain feedback, adapt to context or improve over time”




Source: MIT

> Tools like ChatGPT and Copilot are widely adopted. Over 80 percent of organizations have explored or piloted them…just 5 percent reached production. Most fail due to brittle workflows, lack of contextual learning, and misalignment with day-to-day operations.

> What's really holding it back is that most AI tools don't learn and don’t integrate well into workflows.



Red Hat

# Challenges of Gen AI in Financial Services

Firms must identify and mitigate risks for Gen AI to be successful



## Multimodal Data

Enterprise workflows depend on both structured and unstructured data. (70-90% of enterprise data is unstructured e.g. alternative data)



## Data Privacy

Protect PII, MNPI and IP. Guard against data leakage. Don't include data in public LLMs.



## Accuracy

Vertical Domain IP with accuracy Hallucination rates, determinism, and safety.



## Scaling

Scaling is critical for intersection of AI with Traditional loads as well as both Training/Inferencing such Multimodal combined models to all end users



## Performance

Model, system and financial performance including latency, throughput and cost per inference TCO / ROI



## Governance

Monitor model accuracy, bias, drift and health in order to help mitigate risk



Red Hat

# OSFF 2025 NYC

Employing Neurosymbolic AI with the LLM at the center of the workflow




Source: OSFF NYC 2025



Red Hat

# Demo: Use Case

Mary is a PM at a leading asset manager who has a new client

Imagine Mary, a portfolio manager for a leading asset management firm. Her new client, a pension fund, wants a portfolio that meets their investment guidelines, meets their risk tolerance, avoids conflicts, and, most importantly, meets their investment targets. Now imagine this multiplied tens and hundreds of times.

The intuitive answer to this challenge is … AI to the rescue. For this to be true, it is necessary to combine the power of LLMs and neural networks, providing natural language, reasoning and prediction capabilities, with methods based on logical reasoning and knowledge, such as quantitative analysis and rules-based systems.

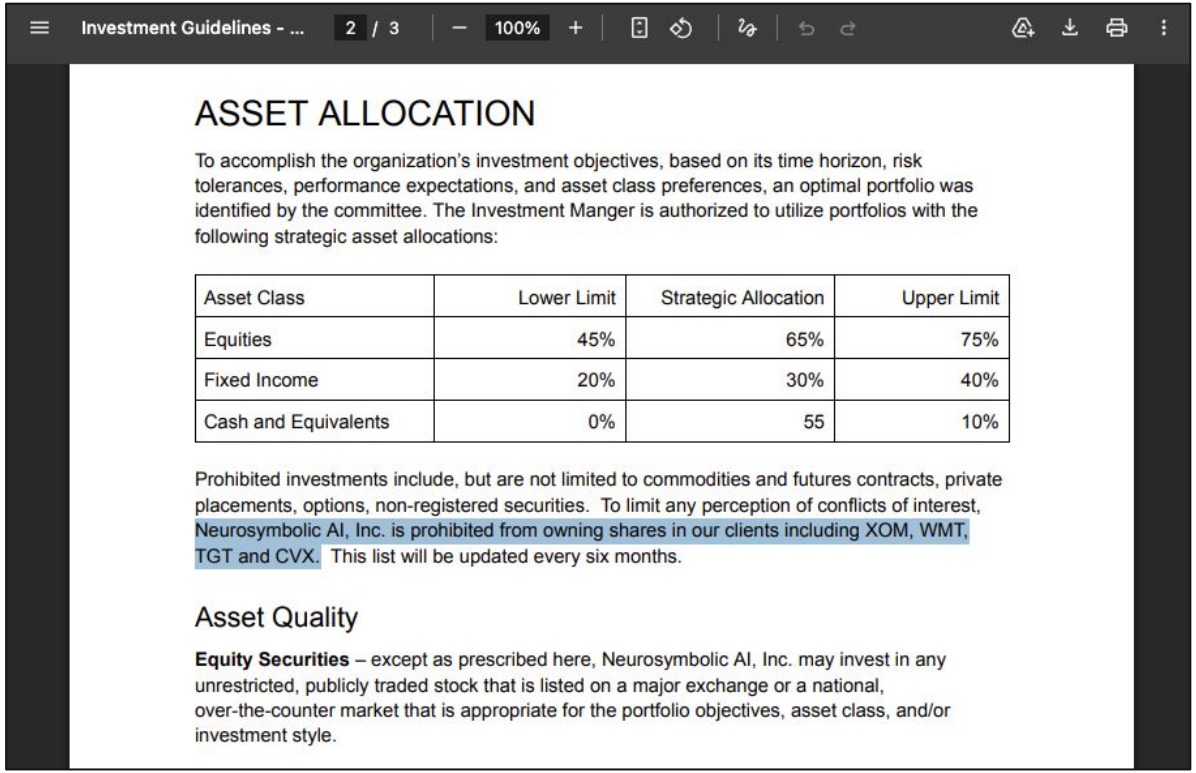
This demo will focus will be on a realistic example: an AI-supported portfolio manager that combines neural networks with symbolic AI into a single, coherent, multi-agent system.



Red Hat

# 80-90% of Enterprise Data is Unstructured Data

Investment guidelines determine what a client can / cannot invest in




## ASSET ALLOCATION

To accomplish the organization's investment objectives, based on its time horizon, risk tolerances, performance expectations, and asset class preferences, an optimal portfolio was identified by the committee. The Investment Manger is authorized to utilize portfolios with the following strategic asset allocations:

| Asset Class | Lower Limit | Strategic Allocation | Upper Limit |
|---|---|---|---|
| Equities | 45% | 65% | 75% |
| Fixed Income | 20% | 30% | 40% |
| Cash and Equivalents | 0% | 55 | 10% |

Prohibited investments include, but are not limited to commodities and futures contracts, private placements, options, non-registered securities. To limit any perception of conflicts of interest, Neurosymbolic AI, Inc. is prohibited from owning shares in our clients including XOM, WMT, TGT and CVX. This list will be updated every six months.

### Asset Quality

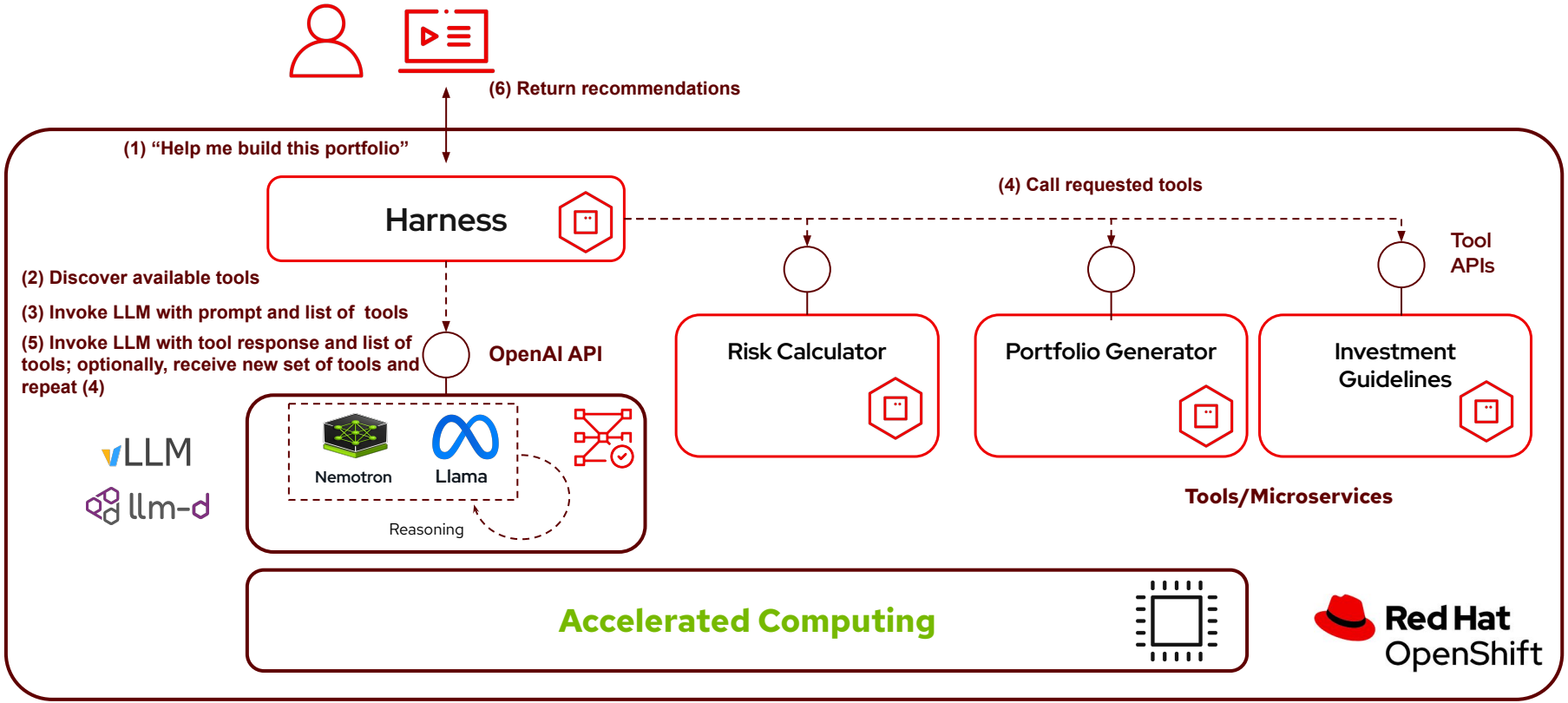
**Equity Securities** – except as prescribed here, Neurosymbolic AI, Inc. may invest in any unrestricted, publicly traded stock that is listed on a major exchange or a national, over-the-counter market that is appropriate for the portfolio objectives, asset class, and/or investment style.



Red Hat

# Demo: Architecture

Portfolio management with Neurosymbolic AI






Red Hat

# Demo: Use Case Trade Settlement

Steve is a Lead Trade analyst reviewing settlement exceptions

Meet Steve, a trade analyst responsible for investigating settlement exceptions—a process that is **often slow, fragmented, and highly manual**. He must **gather information** from internal and external systems, **reconcile** discrepancies, identify **recurring** issues, assess the **financial impact** of delayed settlement, and determine the right corrective action.

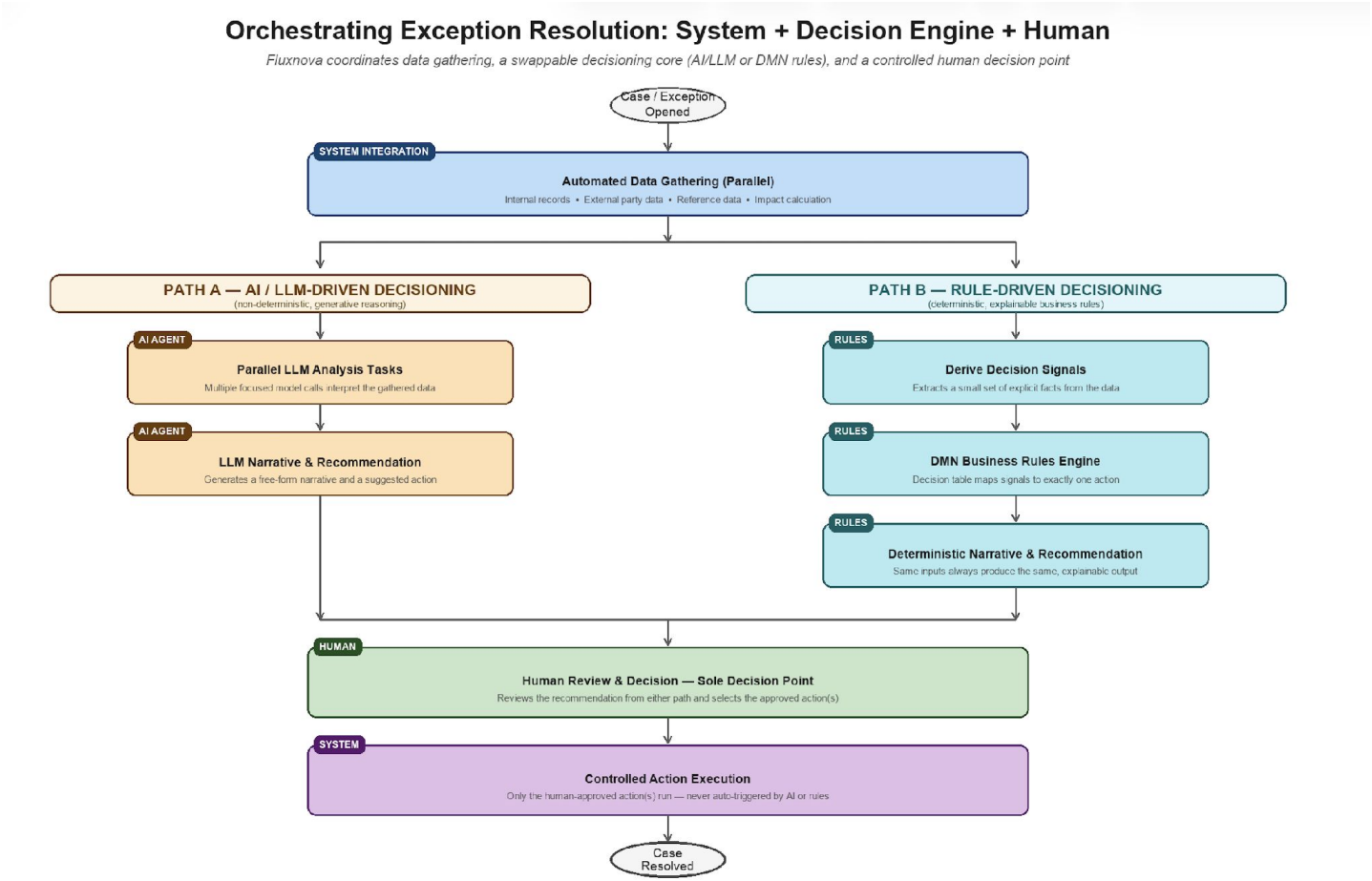
This demo shows how Fluxnova orchestrates systems, AI, and humans within a controlled, end-to-end process. It automates data gathering, uses LLMs to analyze evidence and generate resolution recommendations, and applies DMN rules to determine which cases qualify for straight-through processing (STP) and which require human review. Where approval is required, Steve reviews the findings and selects the corrective actions bringing it back to STP.

With this, Fluxnova helps shorten cycle time, reduce manual effort and settlement exposure, and bring greater consistency and traceability to exception handling along with **complete audit and history of every actions by system, human and AI.**



Red Hat

# Design – Deterministic and Probabilistic

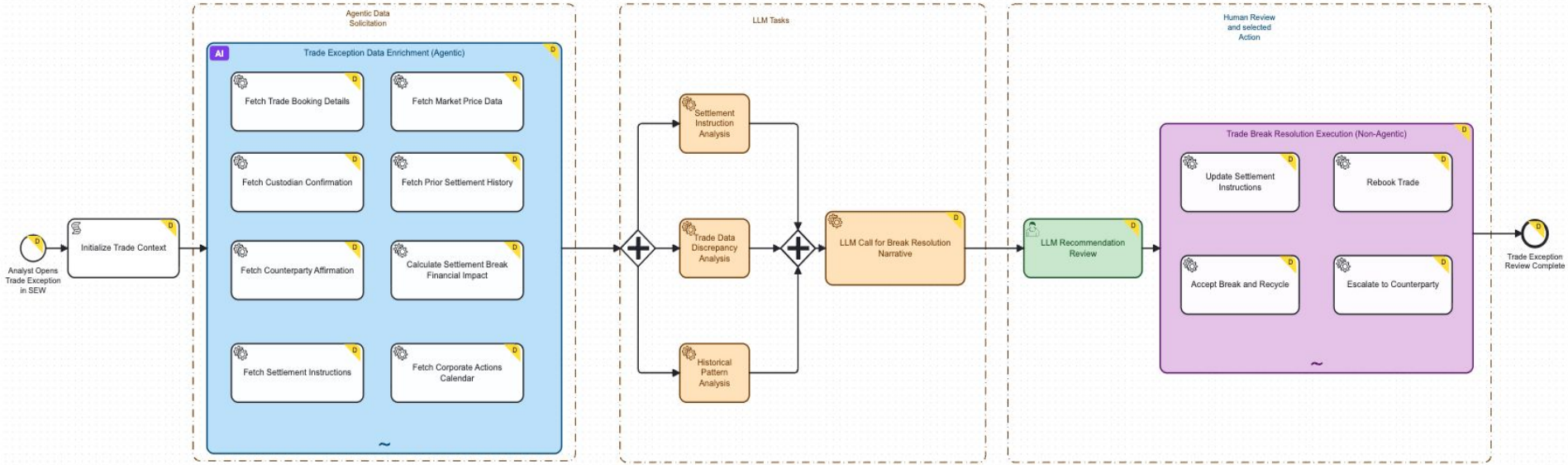





Red Hat

# Demo: Workflow with LLM & Fluxnova AI integrations

## Data Enrichment, LLM Summary, Human Review, auditability and transparency

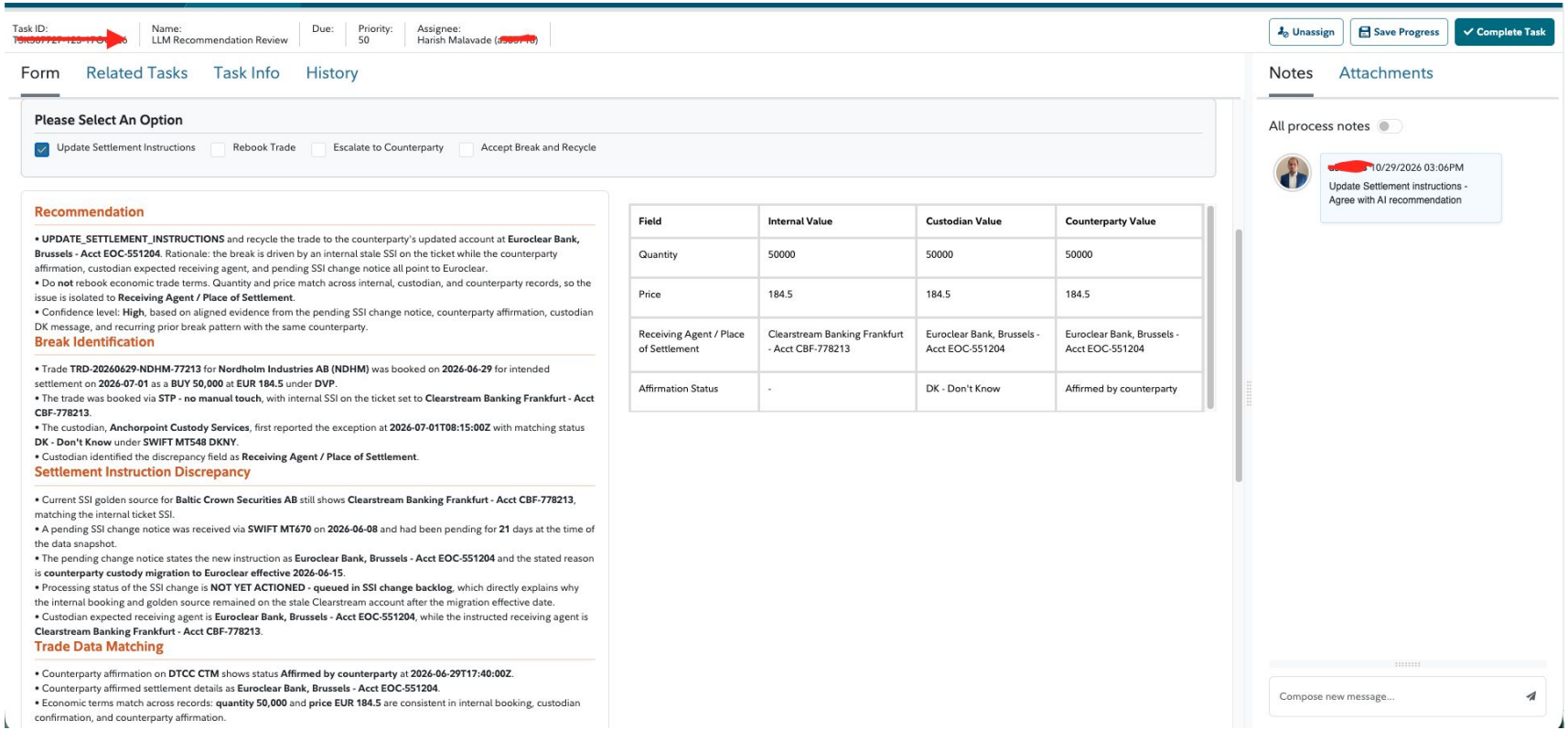





Red Hat

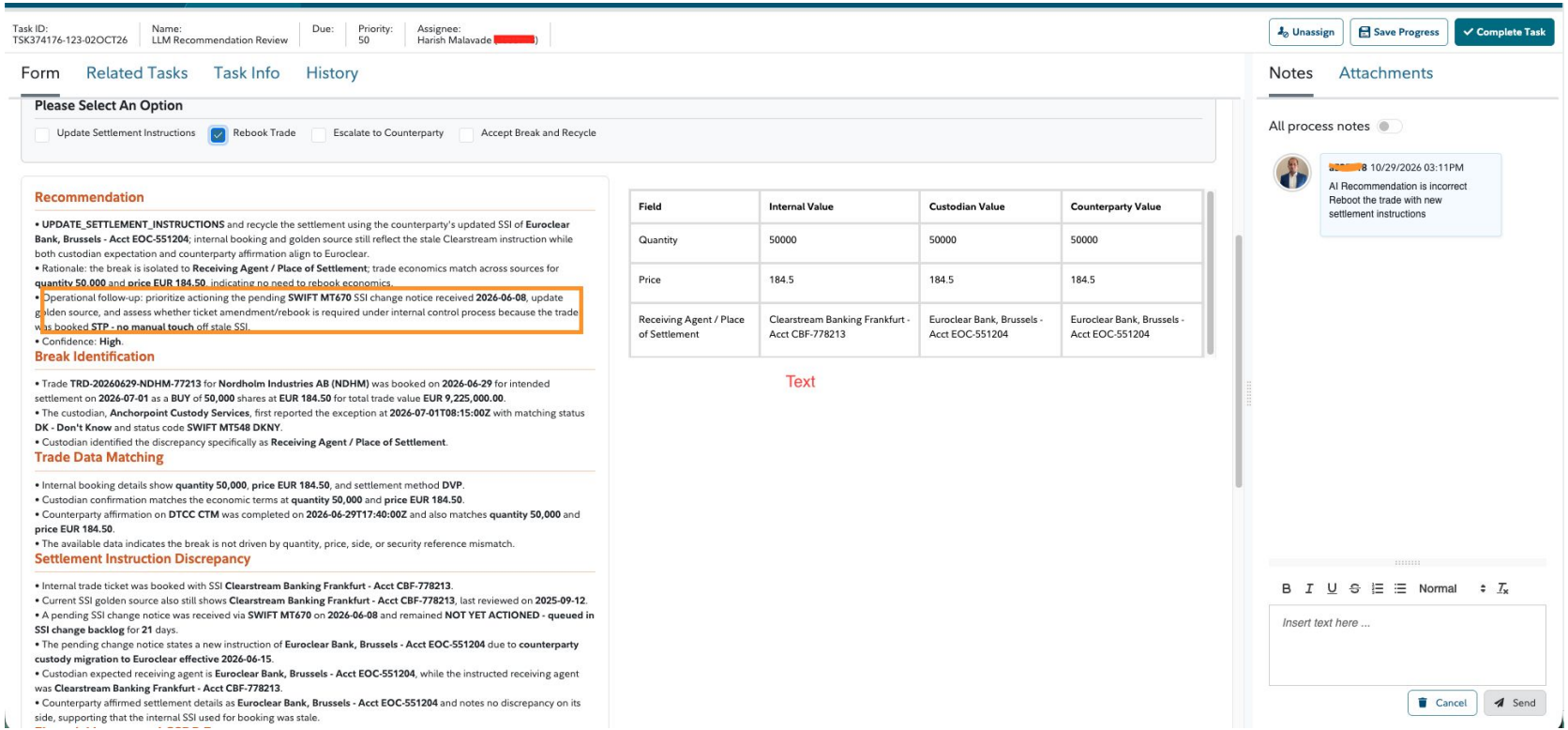
# Demo: LLM Output

## Human Reviews and approves the recommendation



# Demo: LLM Output probabilistic

## Human Reviews and approves the recommendation

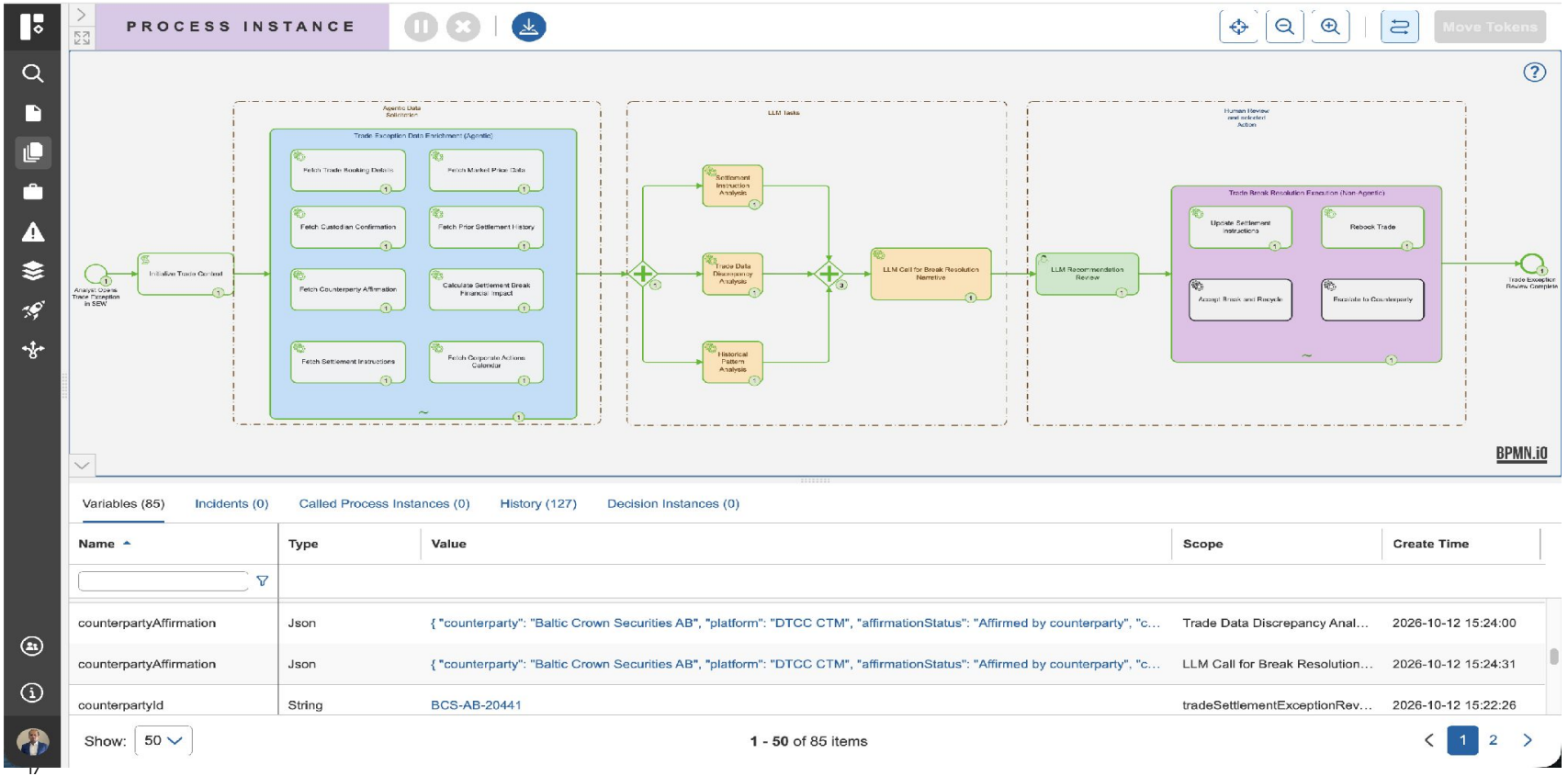





Red Hat

# Demo: Fluxnova Control center

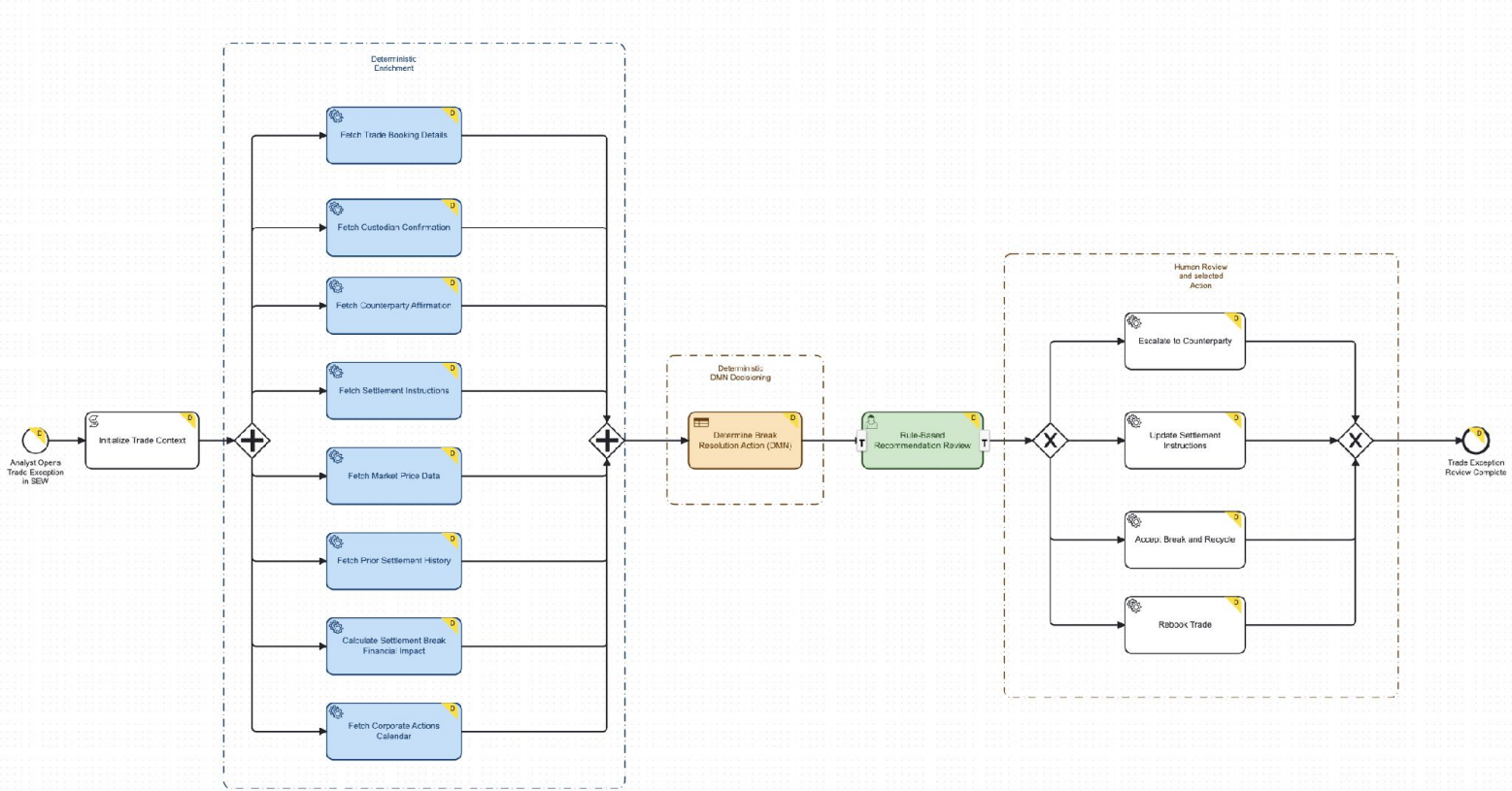
## Transparency, Visibility, Audit and History






Red Hat

# Process 101: Deterministic with DMN Rules






Red Hat

# Process 101: Deterministic with DMN Rules

Edit DRD | Open overview

**Trade Break Resolution Rules V4** — Hit policy: First

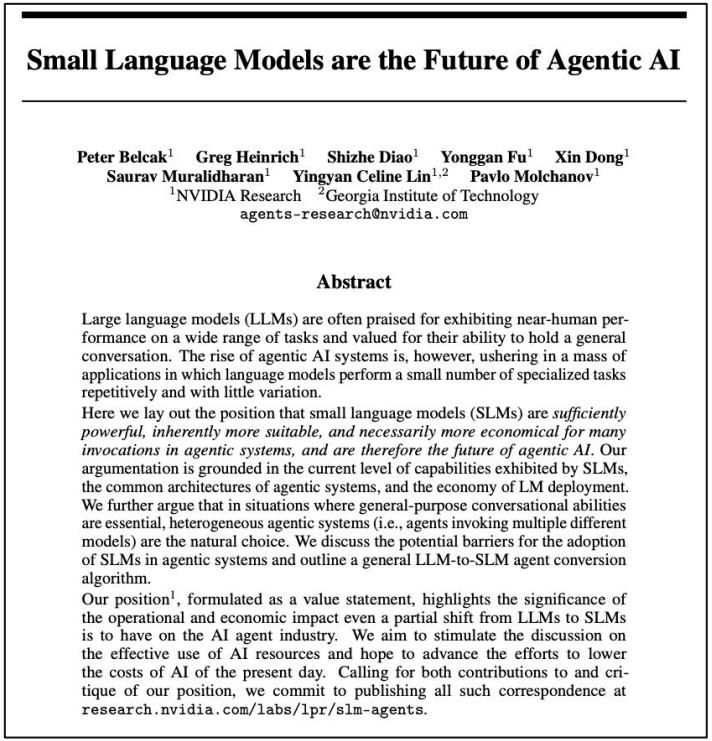
| | When<br>SSI Change Pending<br>boolean | And<br>SSI Change Days Pending<br>long | And<br>Custodian Matching Status<br>string | And<br>Economic Mismatch<br>boolean | And<br>Recurring Breaks With Counterparty ...<br>long | And<br>Buy-In Risk Level<br>string | And<br>Days Past Intended Settlement<br>long | Then<br>Recommended Action<br>string | And<br>Risk Assessment<br>string |
|---|---|---|---|---|---|---|---|---|---|
| 1 | true | >=14 | "DK" | - | - | - | - | "UPDATE_SETTLEMENT_INSTRUCTIONS" | "HIGH" |
| 2 | true | - | "DK" | - | - | - | - | "UPDATE_SETTLEMENT_INSTRUCTIONS" | "MEDIUM" |
| 3 | false | - | not("MATCHED") | true | - | - | - | "REBOOK_TRADE" | "MEDIUM" |
| 4 | false | - | not("MATCHED") | false | >=2 | - | - | "ESCALATE_TO_COUNTERPARTY" | "HIGH" |
| 5 | - | - | - | - | - | "LOW" | <=2 | "ACCEPT_BREAK_AND_RECYCLE" | "LOW" |
| 6 | - | - | - | - | - | - | - | "ESCALATE_TO_COUNTERPARTY" | "MEDIUM" |
| + | - | - | - | - | - | - | - | | |



Red Hat

# Small language Models Outperform Large Language Models

NVIDIA / Georgia Tech: “Serving a 7b SLM is 10-30x cheaper than a 70-150bn LLM”



**Small Language Models are the Future of Agentic AI**


**Peter Belcak**[1] **Greg Heinrich**[1] **Shizhe Diao**[1] **Yonggan Fu**[1] **Xin Dong**[1]
**Saurav Muralidharan**[1] **Yingyan Celine Lin**[1,2] **Pavlo Molchanov**[1]
[1]NVIDIA Research [2]Georgia Institute of Technology
agents-research@nvidia.com



**Abstract**

Large language models (LLMs) are often praised for exhibiting near-human performance on a wide range of tasks and valued for their ability to hold a general conversation. The rise of agentic AI systems is, however, ushering in a mass of applications in which language models perform a small number of specialized tasks repetitively and with little variation.

Here we lay out the position that small language models (SLMs) are *sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems, and are therefore the future of agentic AI*. Our argumentation is grounded in the current level of capabilities exhibited by SLMs, the common architectures of agentic systems, and the economy of LM deployment. We further argue that in situations where general-purpose conversational abilities are essential, heterogeneous agentic systems (i.e., agents invoking multiple different models) are the natural choice. We discuss the potential barriers for the adoption of SLMs in agentic systems and outline a general LLM-to-SLM agent conversion algorithm.

Our position[1], formulated as a value statement, highlights the significance of the operational and economic impact even a partial shift from LLMs to SLMs is to have on the AI agent industry. We aim to stimulate the discussion on the effective use of AI resources and hope to advance the efforts to lower the costs of AI of the present day. Calling for both contributions to and critique of our position, we commit to publishing all such correspondence at research.nvidia.com/labs/lpr/slm-agents.


> small language models (SLMs) are *sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems, and are therefore the future of agentic AI.*

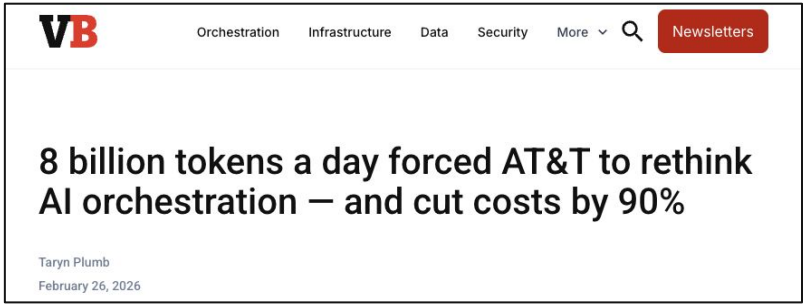
Source: arxiv



Red Hat

# Small language Models Outperform Large Language Models

## AT&T: Cut costs by 90% by utilizing SLMs




Source: Venture Beat

> I believe the future of agentic AI is many, many, many small language models (SLMs). We find small language models to be just as accurate, if not as accurate, as large language models on a given domain area."

– Andy Markus, Chief Data Officer, AT&T



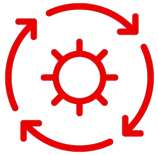
Red Hat

# How to Get Started?

Fork, learn, build, test, iterate







- 100% open source
- Full project instructions in GitHub
- Container images in Quay.io
- Runs in Podman or OpenShift (local, on-prem, cloud)
- Experiment: Always on containers or serverless
- Connect to a  language model via an OpenAI API
- Fluxnova
- Have Fun !

https://github.com/aric-rosenbaum/agents-on-a-leash



Red Hat

# Agents on a Leash:
## Deterministic Agentic AI for Financial Services



**Aric Rosenbaum**
Chief Technologist
Financial Services, Red Hat
aric@redhat.com



**Harish Malavade**
VP, Software Engineering
Fidelity Investments
Harish.Malavade@fmr.com

Red Hat