2026 OPEN SOURCE IN FINANCE FORUM Toronto

Agents on a Leash:
Deterministic Agentic AI for Financial Services

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

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.

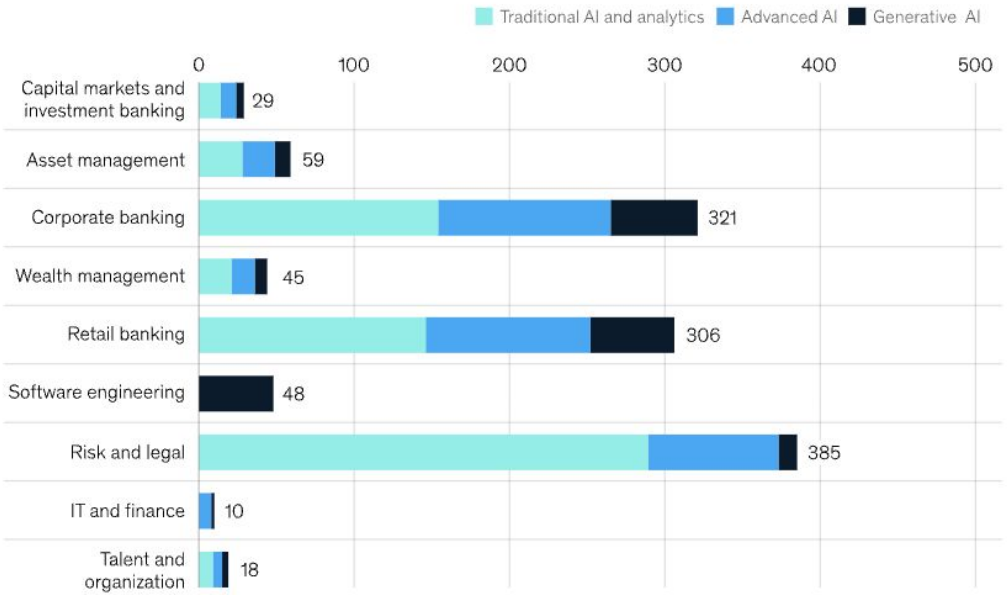
Aric served on the Governing Board of the Fintech Open Source Foundation (FINOS) from 2022-2024.

Red Hat

Gen AI Will Deliver Significant Value to Banks

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

Value created by AI at stake by segment and function,[1] $ billion



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

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.

OSFF 2025 NYC

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

Source: OSFF NYC 2025

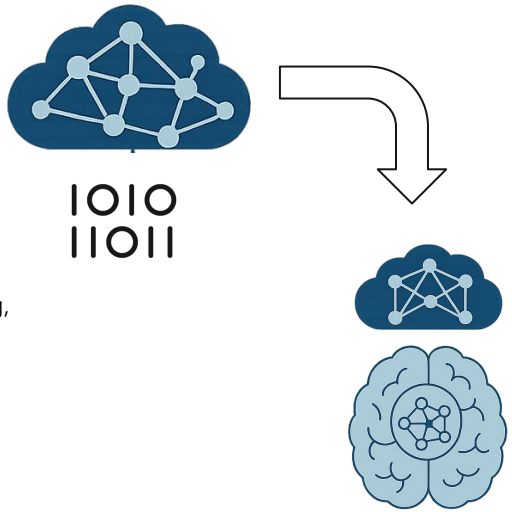
Red Hat

What is Neurosymbolic AI?

And why does it matter

Neural Networks (“System 1 Thinking”)

Pattern Recognition: Excels at extracting features from raw data (images, text, speech).
Learning from Data: Learns statistical correlations without explicit programming.
Scalability: Performance improves with more data and compute.
Limitations: Black-box behavior; weak at logic, reasoning, and generalization.



Neurosymbolic AI

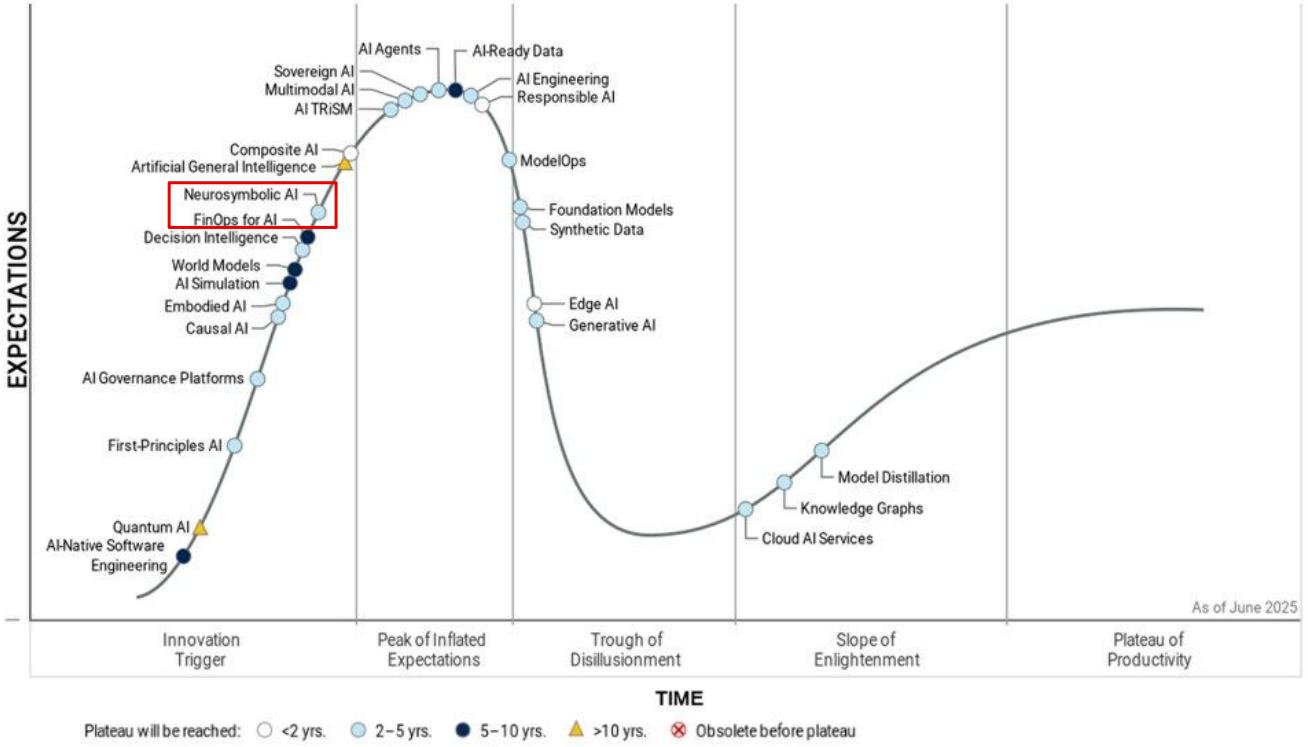
Best of Both Worlds: Combines learning from data with rule-based reasoning and analytics
Explainable and Robust: Adds logical consistency and transparency to neural outputs, eliminates hallucinations
Efficiency: Learns faster and generalizes better with less data
Applications: Vision + reasoning, healthcare, autonomous systems, AI assistants.

Symbolic AI (“System 2 Thinking”)

Rule-Based and Analytical Reasoning: Uses explicit logic, ontologies, analytics and knowledge graphs
Explainability: Provides transparent, traceable decision paths.
Strengths: Strong at structured problem-solving and compliance contexts.
Limitations: Brittle with noisy/unstructured data; struggles to scale.

Neurosymbolic AI is an “Innovation Trigger”: Gartner

Plateau will be reached in 2–5 years



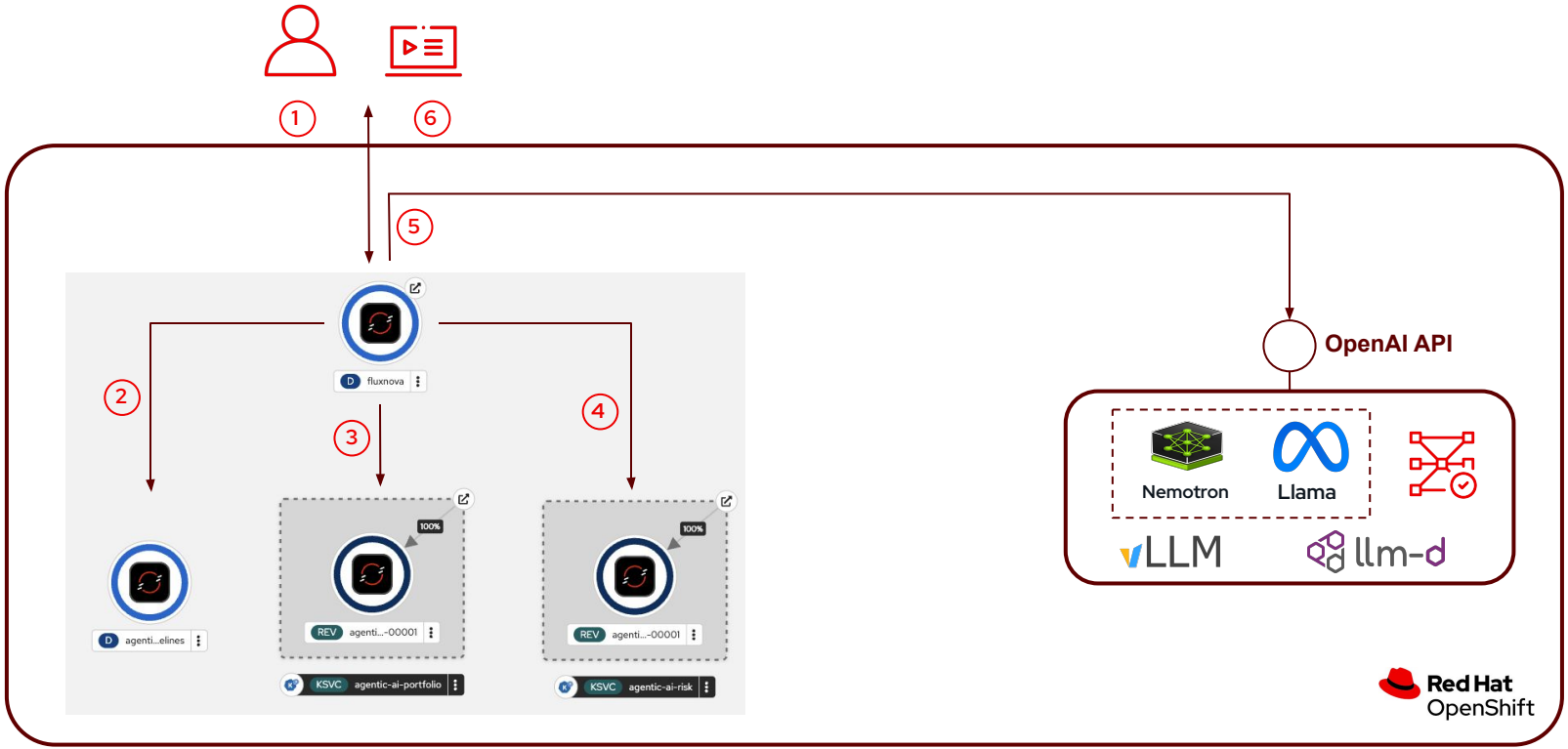
Source: Gartner Hype Cycle Identifies Top AI Innovations in 2025

Gartner

Red Hat

Demo: Architecture

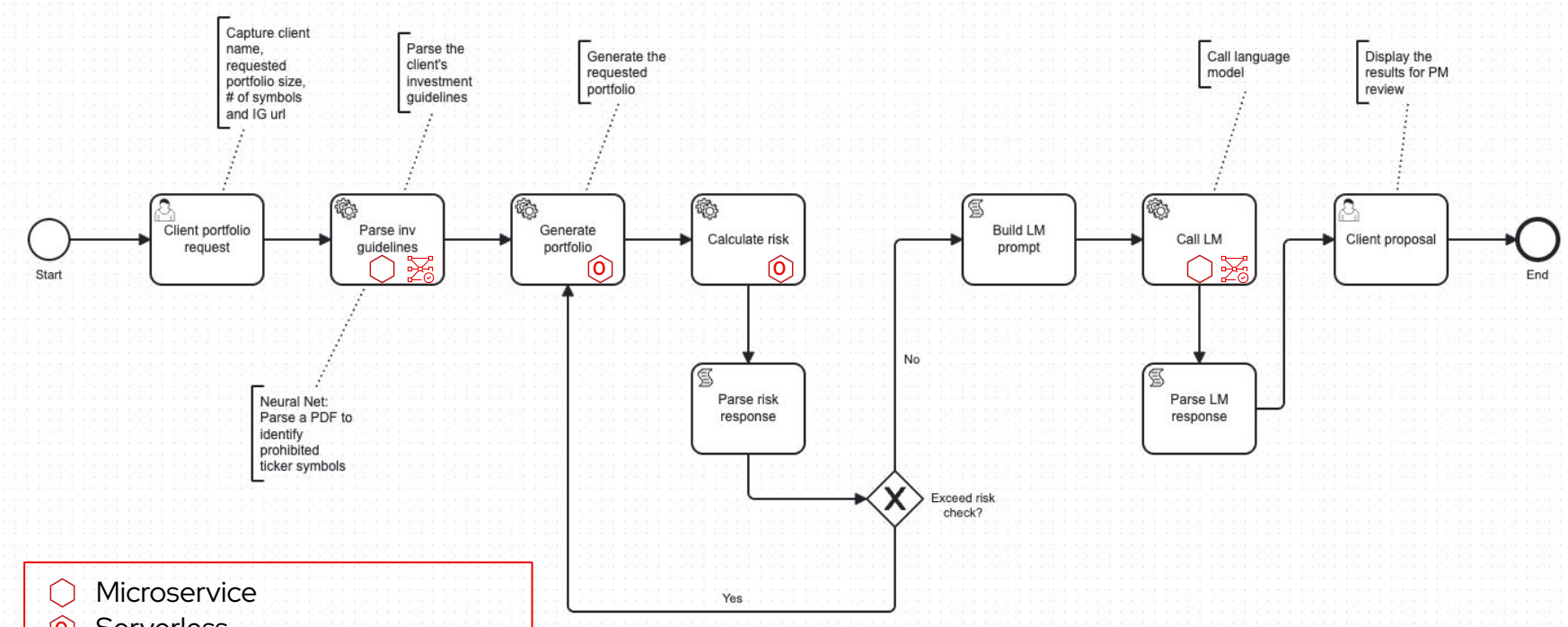
Portfolio management with Fluxnova and AI



Red Hat

Demo: Workflow

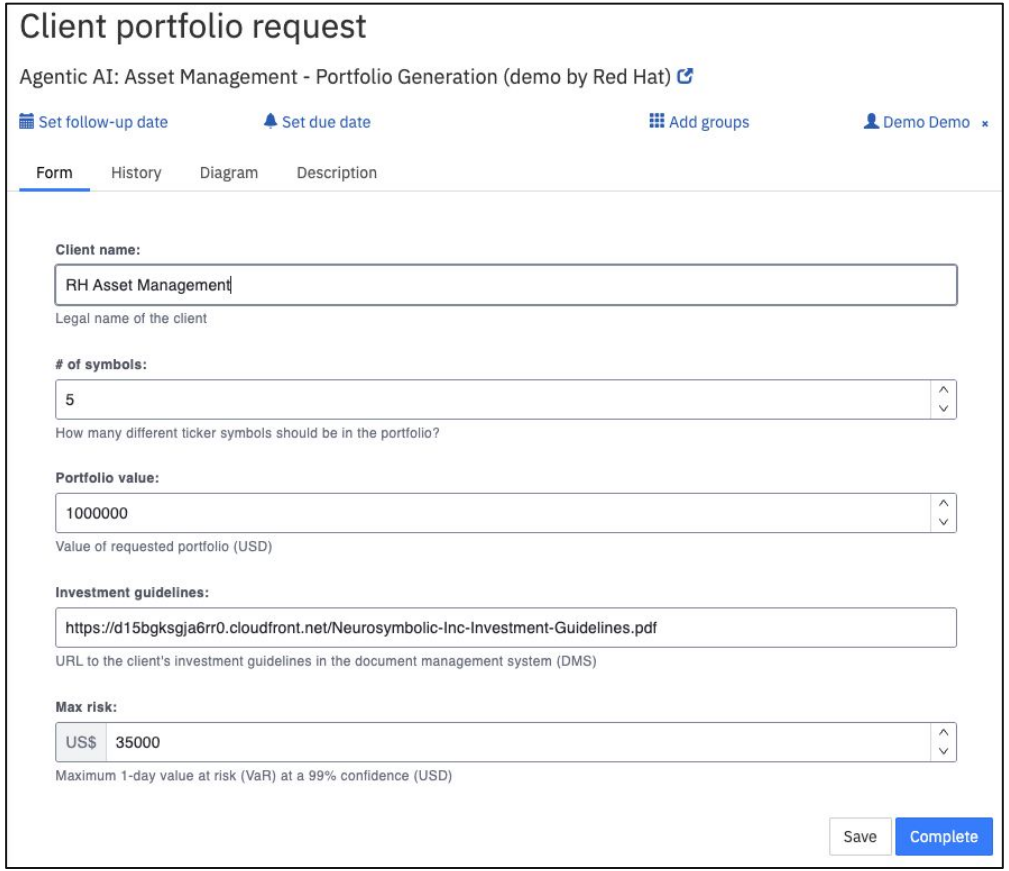
Determinism, auditability and explainability



Red Hat

Demo: Workflow

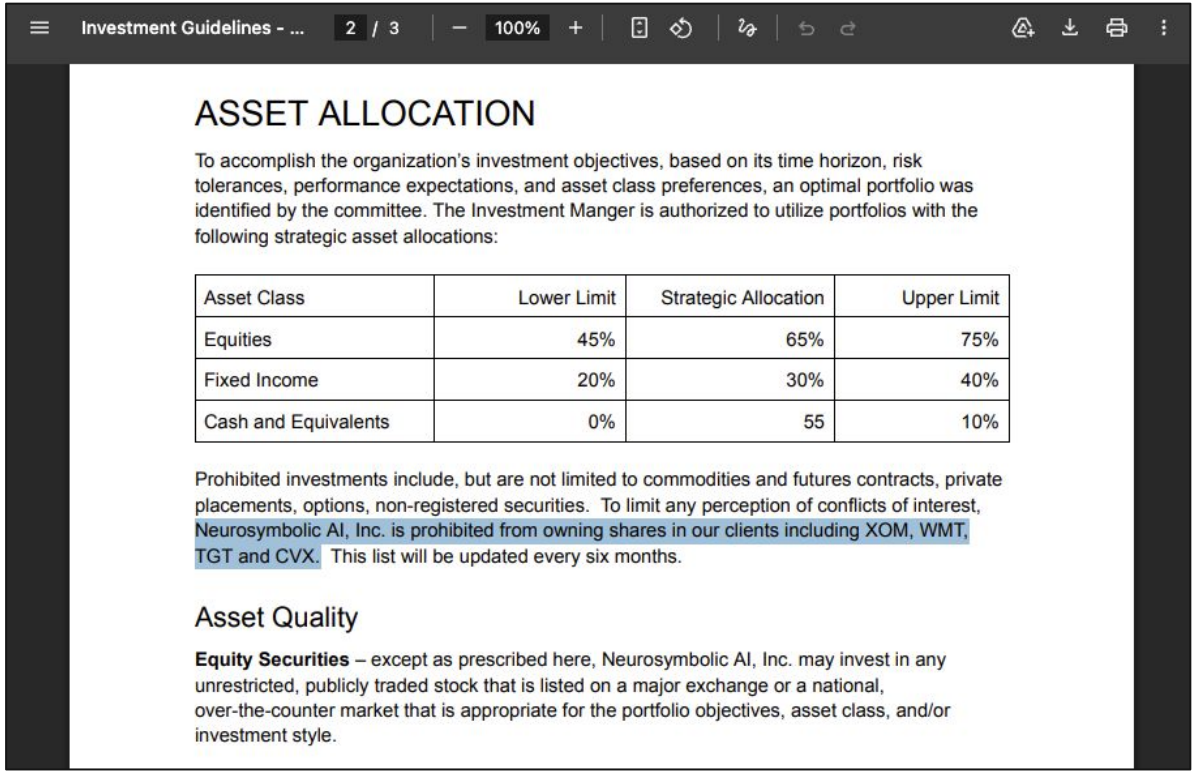
Portfolio management with Fluxnova and AI



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.

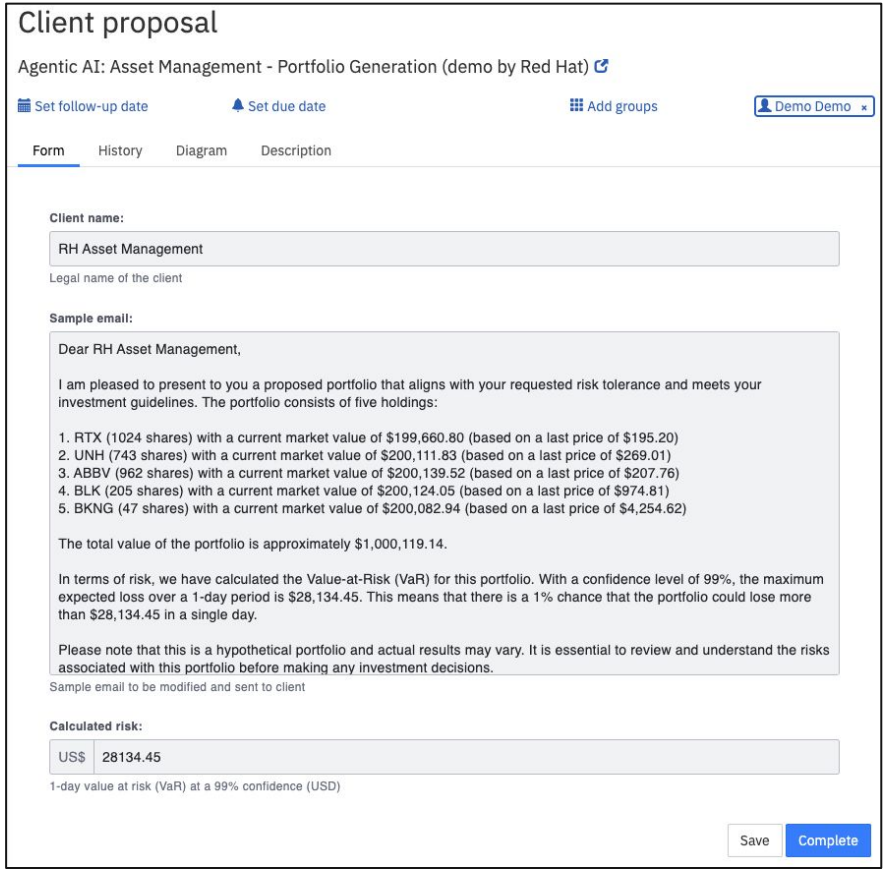
Asset Quality

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: Workflow

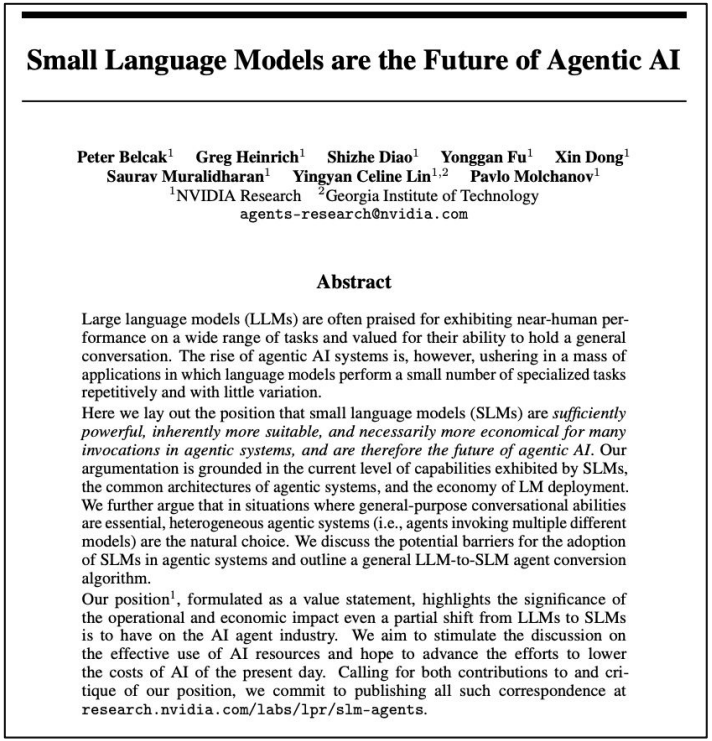
Portfolio management with Fluxnova and AI



Red Hat

Small language Models Outperform Language 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.

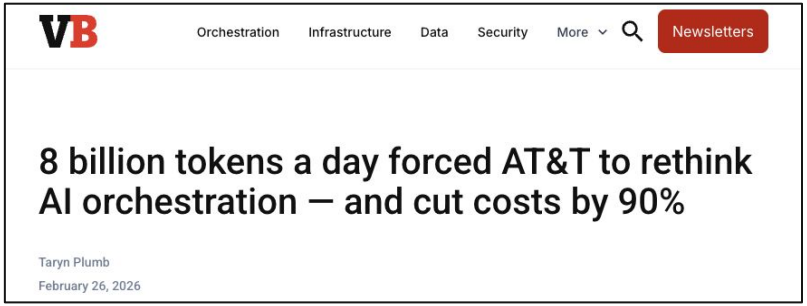
> small language models (SLMs) are *sufficiently powerful, inherently more suitable, and necessarily more economical for many invocations in agentic systems, and are therefore the future of agentic AI.*

Source: arxiv

Red Hat

Small language Models Outperform Language 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.”

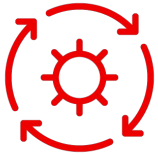
– Andy Markus, Chief Data Officer, AT&T

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

Agents on a Leash:
Deterministic Agentic AI for Financial Services

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

Red Hat

Appendix

Demo: Debug Mode

Portfolio management with Fluxnova and AI

Prohibited tickers (JSON):


```
{
  "matches": [
    {
      "score": 0.9486,
      "sentence": "is prohibited from owning shares in our clients including XOM, WMT, \n TGT and CVX.",
      "tickers": [
        "XOM",
        "WMT",
        "TGT",
        "CVX"
      ]
    }
  ],
  "meta": {
    "num_matches": 1,
    "num_sentences": 74,
    "threshold": 0.65
  },
  "prohibited_tickers": [
```

Demo: Debug Mode

Portfolio management with Fluxnova and AI

Suggested portfolio (JSON):


```
[
  {
    "last_price": 316.34,
    "quantity": 632,
    "symbol": "AXP"
  },
  {
    "last_price": 350.63,
    "quantity": 570,
    "symbol": "AVGO"
  },
  {
    "last_price": 97.21,
    "quantity": 2057,
    "symbol": "SBUX"
  },
  {
    "last_price": 140.76,
    "quantity": 1420,
```

Demo: Debug Mode

Portfolio management with Fluxnova and AI

Value at risk (JSON):


```
{
  "confidence": 0.99,
  "valueAtRisk": 34427.05,
  "valueAtRiskAsOf": 1775740274529
}
```

Demo: Debug Mode

Portfolio management with Fluxnova and AI

LM request (JSON):

{"model":"llama-3-3-70b-instruct-w8a8","messages":[{"role":"system","content":"You are a porfolio manager at an asset manager who has been asked to recommend a portfolio. You should compose a professional, human-reasable response in English that informs the client of the propsed portfolio that fits within their requested risk tolerance and meets their investment guidelines. Do not hallucinate. Only present what you know as fact. If you present false info, you will be fired and the firm will be sued."},{"role":"user","content":"{\"clientName\":\"RH\",\"portfolio\":[{\"last_price\":316.34,\"quantity\":632,\"symbol\":\"AXP\"},
{\"last_price\":350.63,\"quantity\":570,\"symbol\":\"AVGO\"},{\"last_price\":97.21,\"quantity\":2057,\"symbol\":\"SBUX\"},
{\"last_price\":140.76,\"quantity\":1420,\"symbol\":\"PLTR\"},
{\"last_price\":142.66,\"quantity\":1401,\"symbol\":\"EMR\"}],\"valueAtRisk\":
{\"maximumExpectedLoss\":34427.05,\"confidence\":0.99,\"period\":\"1-day\"}}"}],"temperature":0.7}

Red Hat

Demo: Debug Mode

Portfolio management with Fluxnova and AI

LM response (JSON):

{"id":"chatcmpl-bff35b51-34c2-41e4-9210-ae35425e8eb9","object":"chat.completion","created":1775740274,"model":"llama-3-3-70b-instruct-w8a8","choices":[{"index":0,"message":{"role":"assistant","reasoning_content":null,"content":"Dear RH,\n\nI am pleased to present to you a proposed portfolio that aligns with your requested risk tolerance and investment guidelines. The portfolio consists of the following five holdings:\n\n1. American Express (AXP) - 632 shares with a last price of $316.34\n2. Broadcom Inc. (AVGO) - 570 shares with a last price of $350.63\n3. Starbucks Corporation (SBUX) - 2057 shares with a last price of $97.21\n4. Palantir Technologies Inc. (PLTR) - 1420 shares with a last price of $140.76\n5. Emerson Electric Co. (EMR) - 1401 shares with a last price of $142.66\n\nIn terms of risk, I have calculated the Value-at-Risk (VaR) for this portfolio. With a confidence level of 99%, the maximum expected loss over a 1-day period is $34,427.05. This means that there is only a 1% chance that the portfolio will experience a loss greater than this amount over a single trading day.\n\nPlease note that this is a hypothetical portfolio and actual results may vary. It is essential to review and understand the risks associated with this portfolio before making any investment decisions. I recommend that you carefully consider your investment objectives, risk tolerance, and overall financial situation before proceeding.\n\nIf you have any questions or concerns, please do not hesitate to reach out to me. I am here to provide you with any additional information you may need.\n\nBest regards,\n[Your Name]\nPortfolio Manager","tool_calls":[]},"logprobs":null,"finish_reason":"stop","stop_reason":null}],"usage":{"prompt_tokens":239,"total_tokens":563,"completion_tokens":324,"prompt_tokens_details":null},"prompt_logprobs":null,"kv_transfer_params":null}

Red Hat

Demo: Train a Neural Net

The neural net will help identify prohibited ticker symbols in the investment guidelines

```python
def train_default_model() -> Pipeline:
    """
    A lightweight neural net (MLP) to classify whether a sentence/line indicates prohibition.
    Trains on synthetic examples that capture compliance phrasing.
    """
    positives = [
        # Direct "prohibited" / "do not own"
        "The following tickers are prohibited: AAPL, TSLA, MSFT.",
        "Do not
        "The po
```

```python
    X = positives + negatives
    y = [1] * len(positives) + [0] * len(negatives)

    model = Pipeline([
        ("tfidf", TfidfVectorizer(ngram_range=(1, 2), lowercase=True, max_features=8000)),
        ("mlp", MLPClassifier(hidden_layer_sizes=(64,),
                              activation="relu",
                              solver="adam",
                              random_state=42,
                              max_iter=400))
    ])
    model.fit(X, y)
    return model
```

Demo: Train a Neural Net

The neural net will help identify prohibited ticker symbols in the investment guidelines

```
240  def train_default_model() -> Pipeline:
241      """
242      A lightweight neural net (MLP) to classify whether a sentence/line indicates prohibition.
243      Trains on synthetic examples that capture compliance phrasing.
244      """
245      positives = [
246          # Direct "prohibited" / "do not own"
247          "The following tickers are prohibited: AAPL, TSLA, MSFT.",
248          "Do not own the following securities: AMZN and META.",
249          "The portfolio must not hold shares of GOOGL, BRK.B, JPM.",
```

Demo: Train a Neural Net

The neural net will help identify prohibited ticker symbols in the investment guidelines

```
373     X = positives + negatives
374     y = [1] * len(positives) + [0] * len(negatives)
375
376     model = Pipeline([
377         ("tfidf", TfidfVectorizer(ngram_range=(1, 2), lowercase=True, max_features=8000)),
378         ("mlp", MLPClassifier(hidden_layer_sizes=(64,),
379                               activation="relu",
380                               solver="adam",
381                               random_state=42,
382                               max_iter=400))
383     ])
384     model.fit(X, y)
385     return model
```